The Influence of Musical Training on Statistical Learning in Speech Segmentation: Combining Offline Methods With Online Neural Tracking (EEG)

Master's Thesis

Research Master Linguistics

30 EC

Mila Brandsen

6498299

Thesis Supervisors:

Iris van der Wulp, MA & Prof. Dr. Frank Wijnen

Second Reader:

Dr. Marijn Struiksma

14 August 2024

Acknowledgements

I would like to express my gratitude to my supervisor, Iris van der Wulp (MA), for providing me with the opportunity to do my internship and write this thesis within her PhD project. I have immensely enjoyed the past ten months, and I could not have wished for a better supervisor. You have taught me so much and were always ready to answer my questions, whether it was the practical "how do I pronounce this word from the Peabody" or the difficult "why do my results make no sense at all?!" I wish you all the best for the future.

Additionally, I would like to thank my other supervisors, Prof. Dr. Frank Wijnen and Dr. Marijn Struiksma, for their guidance and valuable feedback during the past ten months.

I also want to thank my fellow intern, Carmen Olsthoorn, for her (moral) support, insightful contributions, and advice on whether to keep or reject certain artefacts in the EEG data.

Finally, I want to thank my friends for taking the time to proofread my thesis and their help in catching the embarrassing typos I made.

Abstract

Background: While Statistical Learning (SL), which is crucial for speech segmentation, is well-researched, the exact underpinnings of individual differences in SL remain unclear. Research has found that Musical Training (MT) positively affects various linguistic abilities, but the influence of MT on SL in speech segmentation has yet to be studied.

Aim: The present study investigated the influence of MT on SL in speech segmentation, and examined whether MT is related to musical, specifically rhythmic, abilities.

Method: We used the data of 29 neurotypical Dutch monolingual adults who participated in the study by van der Wulp and colleagues (2023). With EEG, participants' neural entrainment to two artificial languages, namely a structured stream (consisting of trisyllabic non-words with a TP of 1.0 between syllables within words and 0.33 across words) and a random stream (with a TP of 0.09 between all syllables), was measured. Afterwards, they completed a rating task and several musicality tasks (the Gold-MSI, CA-BAT, and PROMS).

Results: In the rating task data, years of MT seemed to positively influence SL. In the EEG data, SL was positively affected by the CA-BAT (i.e., one of the musicality tasks that assesses rhythmic ability) instead of MT. The MT measures (i.e., years of MT and the Gold-MSI's MT subscale) positively correlated with each other and with the CA-BAT. The PROMS did not significantly correlate with any of the musicality measures.

Discussion and Conclusion: Musical ability involves aspects beyond just MT and seems positively related to SL, but it remains unclear which aspects of musicality affect SL. Future research is needed to establish the exact relationship between musicality and SL, preferably by exploring multiple aspects of musicality.

Keywords: *statistical learning, word segmentation, musical training, musicality, electroencephalography (EEG)*

Table of Contents

1. Introduction

Statistical Learning (SL) is the ability to detect regularities and statistical patterns in one's environment through passive exposure and is believed to be a fundamental mechanism for speech segmentation, enabling individuals to detect word boundaries from continuous auditory input (Batterink & Paller, 2017; Moreau et al., 2022). Previous research has shown that individual differences in SL performance are associated with variability in language acquisition and ability (e.g., Siegelman & Frost, 2015; Singh et al., 2012). The exact underpinnings of these individual differences, however, are still unknown.

Musical training (MT) could be one of the underlying factors. Previous research has already established a link between MT and linguistic abilities (e.g., Chobert et al., 2014; Flaugnacco et al., 2015). Notably, François and colleagues (2014) demonstrated that musicians are better and faster at segmenting a sung artificial language than nonmusicians. However, to date, no studies have explicitly connected MT with the SL of speech.

This study aims to broaden our knowledge of the factors underlying individual differences in SL by examining the influence of MT on the SL ability of speech in adults. We used data from Van der Wulp and colleagues (2023), which includes offline (a rating task and several musicality tasks) and online measured data (electroencephalography (EEG) recordings of a structured and random listening task). Additionally, we investigated to what extent musical, more specifically rhythmic, ability is correlated with MT experience.

This thesis starts with a theoretical framework discussing SL in speech segmentation and the influence of MT, followed by an introduction of the current study, including the research questions and hypotheses, in section 3. Section 4 details the methodology, and section 5 reports the results. Section 6 offers a discussion of these results, addresses limitations, and provides suggestions for future research. Finally, section 7 presents a conclusion. Additional materials are included in the appendices.

2. Theoretical Background

2.1 Statistical Learning in Speech Segmentation

Learning a new language entails the challenge of speech segmentation: detecting word boundaries in a continuous stream of speech. Considering that fluent natural speech has no clear pauses between words as opposed to written language, this seems like a difficult task (Saffran et al., 1996-a). However, in 1995, Jusczyk and Aslin found that even 8-month-old infants are able to discover word boundaries when learning a new language. Following this, an important study by Saffran, Newport and Aslin (1996-a) showed how infants manage to detect word boundaries solely based on the Statistical Learning (SL) of Transitional Probabilities (TPs), without any prosodic cues. SL is the process of using statistical properties of linguistic input to discover structure. In this case, learners can discover word boundaries by tracking TPs between neighbouring syllables: the probability that a syllable X is directly followed by a syllable Y, given the overall frequency of X (Saffran, 2003; Saffran et al., 1996-a). This notion can also be expressed as:

$$\text{TP of } Y|X \; = \frac{frequency\ of\ XY}{frequency\ of\ X}$$

In natural language, neighbouring syllables within words have higher TPs than neighbouring syllables between words, enabling learners to hypothesise where word boundaries are likely to occur. For example, in the sound sequence *pretty baby*, the TP from *pre* to *ty* is bigger than the one from *ty* to *ba* (Saffran, 2003; Saffran et al., 1996-a). Batterink and Paller (2017) further propose that SL consists of two components, namely word identification (i.e., identifying different word forms by segmenting the speech input based on TPs) and memory storage (i.e., storing these extracted representations in long-term memory).

In the experiment by Safran et al. (1996-a), infants were familiarised with two minutes of a continuous speech stream of four trisyllabic words from an artificial language, during which they had to identify word boundaries. The only cue for word boundaries were TPs

between syllable pairs; these were higher within words (1.0 in all cases) than between words (0.33 in all cases). The discovered word forms, then, must be stored in the infants' long-term memory for them to be recalled during the subsequent test phase. In the test phase, the Head-Turn Preference Procedure (HTPP) was used: the infants were presented with a trisyllabic word or non-word (i.e., a word that contained the same syllables as the artificial language but in an unfamiliar order) for as long as they looked at a blinking light. Following Jusczyk and Aslin (1995), results show a novelty effect: the infants had longer listening (and looking) times for non-words than words, indicating that they learned to segment words by detecting TPs, and recognised the difference between novel and familiar orderings. In other words, the authors infer that infants can perform SL.

Over the years, several studies have confirmed this observation (e.g., Choi et al., 2020; Hay et al., 2011; Singh et al., 2012). To further explore infants' SL abilities, Hay and colleagues (2011) examined whether infants can use the output of SL as the input for word-meaning association learning. Their experiments with 17-month-old infants combined a segmentation task, familiarising themselves with Italian speech, with a label-object association task. The results suggest a significant correlation between the statistics of the speech stream and infants' success in mapping labels to referents. The TPs internal to the labels did not affect the learning outcomes: infants were able to successfully map both low (forward TP of 0.33, backward TP of 1.0) and high (forward and backward TP of 1.0) TP words as labels for objects (Experiment 1). However, when no TP information was available (Experiment 2) or when the internal TP of labels was low in both forward and backward directions (Experiment 3), infants failed to connect labels to their referents. Across experiments, infants were able to learn new object labels if the TPs between syllables were high in at least one direction and if they had the opportunity to segment the sound sequences from continuous speech beforehand. Hence, word

learning in infants seems strongly influenced by prior experience with the distribution of sounds in natural languages.

However, infants are not the only ones who (are able to) do this: various studies have established that children and young adults can do the same (e.g., Batterink, 2017; Batterink & Paller, 2017; Choi et al., 2020; Misyak et al., 2010; Moreau et al., 2022; Pinto et al., 2022; Saffran et al., 1996-b). Choi and colleagues (2020) demonstrated that infants and young adults show a similar increase in neural entrainment (i.e., synchronisation of brainwaves, see section 2.1.2) to embedded words, and, thus, follow similar learning trajectories when tracking TPs. Moreover, children and adults have similar SL abilities: unlike other aspects of language learning, SL does not decline from childhood to early adulthood (Moreau et al., 2022). In older adults, however, cognitive ageing is likely to affect learning mechanisms, which may cause a decline in SL abilities (Schevenels et al., 2023). While the general ability appears to be preserved, they may resort to a different strategy or mechanism than young adults, presumably due to age-related declines in the relevant brain structures (Ong & Chan, 2019).

An example of a study with adult participants is the experiment by Batterink and Paller (2017), who combined behavioural measures (e.g., an explicit rating task and a Reaction Time (RT) task called the Target-Detection Task (TDT)) with electroencephalography (EEG). Participants were presented with both a structured (with a TP of 1.0 between syllables within words and 0.33 between syllables across words) and a random syllable stream (with a TP of 0.09 between all syllables; note that consecutive occurrences of the same syllable did not occur). The researchers found that neural entrainment to the words – or triplets in the random stream – was higher in the structured stream than in the random stream. This heightened entrainment increased progressively with the duration of exposure, suggesting an ongoing process of SL. Furthermore, it predicted performance on the TDT: participants who showed

higher entrainment also showed a larger RT effect, which indicates that SL learning facilitated processing and led to faster RTs.

2.1.1 Individual Differences

Although SL abilities are comparable across different age groups, many studies find that there are individual differences in learning trajectories or memory performance between participants (Batterink & Paller, 2017; Misyak et al., 2010; Siegelman & Frost, 2015). Frost and colleagues (2015) propose that these variations stem from two major sources: (1) variance in how individual elements in a sequence, such as syllables, are encoded (i.e., represented in memory), and (2) disparities in detecting TPs between syllables. Batterink and Paller (2017) posit that long-term memory interacts with and influences these word identification mechanisms, subsequently impacting SL performance: if people encounter challenges in effectively storing the extracted representations in their long-term memory, they are expected to show poorer performance on SL tasks during the test phase.

Individual differences in SL performance are associated with variability in language acquisition and ability (Siegelman, 2020; Singh et al., 2012). Variations in SL ability may play a role in differences observed in linguistic performance, like online processing and the development of vocabulary. The longitudinal study by Singh and colleagues (2012), for example, showed a strong degree of association between infant word segmentation abilities at 7 months and their productive vocabulary size at 24 months. A study by Misyak and colleagues (2010) found a correlation between individual differences in SL of nonadjacent dependencies and participants' online processing of long-distance dependencies. Finally, several studies indicate that individuals with Developmental Language Disorder (DLD) exhibit an SL deficit: those with DLD are less effective in auditory SL than their peers without DLD (see Lammertink

et al., 2017). The exact underpinnings of these individual differences in SL, however, remain unclear.

2.1.2 Various Measures in Statistical Learning Research

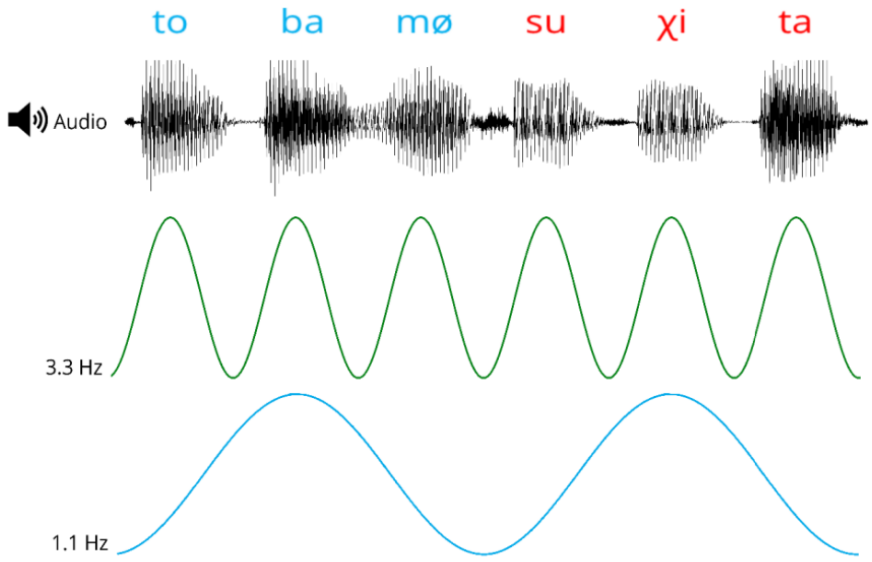
To study SL in speech segmentation, researchers typically expose participants to a continuous speech stream of repeating trisyllabic non-words. Their SL abilities, then, can be measured in several different ways, both offline with behavioural measures and online with EEG. Examples of the offline measures are the HTPP, the Two-Alternative Forced-Choice (2AFC) task, a rating task, and an RT task (e.g., Batterink & Paller, 2019). The HTPP is a frequently used measure in infant or child research which assesses listening behaviours towards auditory stimuli. After a blinking light attracts the child's attention, the auditory stimulus starts for as long as the child keeps looking at the light. In the case of SL, the exposure phase is followed by a test phase in which the HTPP is used to the duration of infants' listening times to different words and foils (Safran et al., 1996-a). Adult research often uses a 2AFC task in the test phase, which has participants discriminate between words and foils: two stimuli, a word and a foil, are presented and participants must decide which one is more familiar based on the stimuli presented during the exposure phase (Saffran et al., 1996-b; Batterink & Paller, 2019). Rating tasks are quite similar: participants are presented with a word or a foil and must rate their familiarity compared to the exposure phase (Batterink & Paller, 2017). Finally, in the TDT, participants are asked to repeatedly identify a target stimulus in a continuous stream. If SL has occurred, then participants show faster RTs to predictable targets (i.e., the second or third syllable of a trisyllabic word) compared to unpredictable ones (i.e., the first syllable of a trisyllabic word) (Batterink & Paller, 2017; Batterink & Paller, 2019).

Online measurements can be done with EEG, a functional brain research technique that can measure brain activity (see Batterink & Paller, 2017; Choi et al., 2020; Moreau et al., 2022;

Pinto et al., 2022). Contrary to the offline measures, this measure can assess SL while it is happening: it can record neural oscillations during the identification phase, i.e., when people start to identify word forms by segmenting the speech input based on TPs (Batterink & Paller, 2017). Neural oscillations, represented as successive "waves" in EEG output, reflect rhythmic brain activity driven by different neuron populations firing in specific frequency patterns. The frequency ranges of these oscillations, known as frequency bands, characterise their distinctive patterns. Neural oscillations can phase-lock to the rhythm of a stimulus (e.g., language), also known as neural entrainment or synchronisation (Peelle & Davis, 2012). In the identification phase of SL, phase-locking changes from only the syllable frequency (at a rate of 3.3 Hz) to the addition of a word frequency (at a rate of 1.1 Hz), as shown in Figure 1 (Batterink & Paller, 2017).

Figure 1

Syllable (3.3 Hz) and Word (1.1 Hz) Frequency Rates in a Stream of Trisyllabic Words.



Note. Reprinted from van der Wulp et al. (2023, Figure 2, p. 13).

To quantify neural entrainment at the syllabic and word frequency, the Inter-Trial Coherence (ITC), or: "phase-locking value", is calculated. ITC values range from 0 (non-phase-locked activity) to 1 (strictly phase-locked activity to a given frequency in the stimulus), based on the amount of phase-locking to the syllable and word frequencies. By dividing the ITC at the word frequency by the ITC at the syllable frequency, researchers can calculate a Word Learning Index (WLI), which offers insights into the progression of word learning during SL:

$$WLI = \frac{ITC\ of\ word\ frequency}{ITC\ of\ syllable\ frequency}$$

Higher WLI values indicate greater neural entrainment to the word frequency relative to the syllable frequency, which indicates the occurrence of SL (Batterink & Paller, 2017). The experiment by Batterink and Paller (2017) used both EEG and several offline measures (e.g., a rating task and TDT). While the offline (behavioural) tasks occur after participants have performed SL, EEG is able to measure these processes while they are happening. It captures only the identification phase of SL, while the rating task occurs after both the identification and the memory storage phase. Batterink and Paller (2017) show that EEG can offer valuable insights into the processes of speech segmentation and SL: variability in the WLI predicted performance on the RT task.

2.2 The Influence of Musical Training

EEG also plays an important role in rhythm and music research (e.g., Bouwer et al., 2016; Cantiani et al., 2022; François et al., 2014; Intartaglia et al., 2017; Schön & François, 2011). As mentioned before, neural oscillations have been shown to phase-lock to the rhythm of an auditory stimulus, for example to the rhythm of speech (Peelle & Davis 2012). Research has also shown that there is neural entrainment to the rhythm of music(al beats) (Nozaradan, 2014). To illustrate, in a recent study by Cantiani and colleagues (2022), 8-month-old infants were

exposed to two complex – non-speech (musical) and speech–rhythmic stimuli while their EEG signals were recorded. The researchers found that these infants were capable of entraining to the incoming auditory rhythms. A small control group of adults showed that their neural entrainment to the stimuli was very similar to the pattern observed in the infants.

EEG research with musicians reveals an effect of musical expertise, which increases the accuracy of phase-locking to music (Di Liberto et al., 2020). Musical expertise often results from years of Musical Training (MT), the latter of which, in turn, is shown to be associated with enhanced sensitivity to statistical cues in auditory SL tasks with pure tones (Mandikal Vasuki et al., 2017). MT does not only influence how people listen to music but could also affect their linguistic abilities. Research by Flaugnacco and colleagues (2015), for instance, revealed that MT was successful in enhancing reading and phonological skills in children with dyslexia. In their experiment, children with dyslexia (aged 8-11) were divided into two groups, a music group and a control group. The music group followed music classes and the other group completed painting classes for two hours a week for seven months. Before and after this intervention, their linguistic, musical, reading, and general cognitive abilities were administered. After the training, the music group outperformed the painting group in tasks about rhythmic abilities, phonological awareness and reading skills. These findings indicate that MT can positively influence reading and phonological skills in dyslexic children. In the same vein, children with hearing loss who wear cochlear implants or hearing aids could also benefit from MT: Hidalgo et al. (2017) conducted a study comparing temporal adaptation (i.e., mutual accommodation of temporal structure, such as speech rating and turn-timing) in speech interaction in children with normal hearing and their peers with hearing loss. Results indicate that, while children with normal hearing derive advantages from the temporal regularity of stress occurrences, children with hearing loss only develop sensitivity to this manipulation after rhythmic training. This training could assist them in organising the temporal flow of their

verbal interactions. Finally, a longitudinal study by Chobert et al. (2014) also provides evidence supporting the impact of MT intervention on linguistic abilities. They investigated whether active MT had an impact on preattentive processing, i.e., automatic and unconscious processing, of syllabic duration and Voice Onset Time (VOT). Children (aged 8-10) were randomly assigned to music or painting training. The researchers recorded the mismatch negativity (MMN) to syllables that differed in vowel frequency, vowel duration, and VOT. While both groups performed similarly before training, after 12 months, only the musically trained children showed enhancements of MMN amplitude to duration and VOT deviants. In other words, enhanced preattentive processing of syllabic duration and VOT was only found in the musically trained group.

Other studies on the effects of MT do not implement MT as an intervention but compare musicians, those who had MT, to nonmusicians, those who had not. An example is François et al. (2014), who had professional musicians and nonmusicians listen to an artificial language of sung pseudowords. After this exposure phase, a 2AFC task on pairs of pseudowords and melodies followed. Data from this task and ERPs reveal that musicians are better and faster at segmenting an artificial language than nonmusicians. As all musicians had over twelve years of MT and the nonmusicians not more than two, these results support the hypothesis that there is an MT effect that can be transferred from music to sound stream segmentation and perhaps even to SL. The findings of Intartaglia and colleagues (2017) are in line with these results. In their EEG experiment, English native nonmusicians and French non-native musicians and nonmusicians listened to an English syllable that does not exist in French. Whereas the natives had more robust subcortical representations of the syllable and an advantage in neural encoding of the features than non-native nonmusicians, the non-native musicians showed similar subcortical representations of the syllable compared to the natives. This suggests that MT may compensate for a lack of language experience.

2.2.1 Operationalisation

It is noteworthy that the operationalisation of the terms "musician" and "musically trained" varies among studies. While most studies use years of MT or music lessons as their main criterion, there are inconsistencies in the duration required for an individual to be classified as musically (un)trained or as (non)musicians. Generally, music psychology research follows the "six-year rule": a musician has at least six years of musical expertise, which is defined by years of MT, and a non-musician less than six (Zhang et al., 2018). However, not all research fields adhere to this rule. In some studies, nonmusicians are participants with no prior MT at all (e.g., Correia et al., 2023), while others classify nonmusicians as participants with no more than two years of MT (e.g., François et al., 2014). Another example is Cirelli et al. (2016), who divide infants into two groups based on whether their parents had five or more years of combined MT or less than five. These distinctions seem rather arbitrary. This variation in level of expertise is important to keep in mind when comparing studies: five years of MT represents a substantial amount compared to zero years.

Conversely, research by Correia and colleagues (2023) demonstrates that formal MT is not required to develop musical abilities and achieve musician-like performance on tests of musical and cognitive abilities. They state that the musicality of untrained participants and informal musical practice should also be considered in studies of musical expertise. To do so, there are various measures researchers could use, for example, the Goldsmiths Musical Sophistication Index (Gold-MSI). This self-report questionnaire measures musical expertise and experience, encompassing multiple subscales that assess different facets of musical sophistication (e.g., active engagement with music, perceptual abilities, and MT) (Müllensiefen et al., 2014). Another possible task is the Profile of Music Perception Skills (PROMS), which uses various subsets (e.g., melody, rhythm, and accent) to assess one's musical abilities

(Zentner & Strauss, 2017). Finally, researchers can use the Computerised Adaptive Beat Alignment Test (CA-BAT), in which participants listen to the same piece of music twice; once with beeps synchronised to the rhythm and once with beeps out of sync. Afterwards, they must indicate which of the two tracks had beeps in sync with the rhythm (Harrison & Müllensiefen, 2018).

3. Current Study

Previous studies have shown that Musical Training (MT) can positively influence Statistical Learning (SL) (e.g., François et al., 2014). To our knowledge, however, no study has made the explicit connection between MT and SL of speech segmentation before. Exploring how MT may influence SL processes involved in language learning could broaden our knowledge of the cognitive impact that MT has. Moreover, it could provide new insights into the underpinnings of individual differences in SL. Therefore, to investigate whether MT affects the SL of speech segmentation, we formulated the following research question:

RQ1: To what extent does musical training influence the statistical learning ability aiding speech segmentation in adults?

To study this, we used both online and offline measures. With electroencephalography (EEG), we recorded the neural oscillations of native Dutch participants who listened to two artificial languages with trisyllabic words based on the phonotactics of Dutch, namely a structured and a random stream. Afterwards, they completed a rating task.

We predicted MT to have a positive influence on SL in speech segmentation; overall, participants who are musically trained were expected to show a higher Word-Learning Index (WLI) than those who are less musically trained.

Additionally, research has shown that the definition of "musically trained" varies between studies, and that formal MT is not necessary to develop musical abilities and obtain musician-like performance on tests (Correia et al., 2023). Therefore, before answering RQ1, we examined the best way to define "musically trained" and looked at (possible) correlations between MT and musical ability. To this end, a second research question was formulated:

RQ2: To what extent is musical ability, measured with the CA-BAT and PROMS, related to (years of) musical training?

To measure MT, we first looked at the number of years of MT, which is most often used in other studies. Moreover, we also analysed the MT subscale of the Gold-MSI (Müllensiefen et al., 2014), as this self-report questionnaire is also frequently utilised in music research. Participants' musical ability was measured with more direct tests of musical ability, namely the CA-BAT (Zentner & Strauss, 2017) and PROMS (Harrison & Müllensiefen, 2018).

We expected that the CA-BAT and PROMS scores would positively correlate with (years of) MT and, therefore, that musical ability is related to MT.

4. Methodology

This study used the data from the in-principle accepted peer-reviewed study by van der Wulp and colleagues (2023), which has been approved by the Linguistics Chamber of the Faculty Ethics Assessment Committee of Humanities at Utrecht University (reference number: LK-22-174-02). Furthermore, the experiment was preregistered prior to testing (Brandsen et al., 2024).

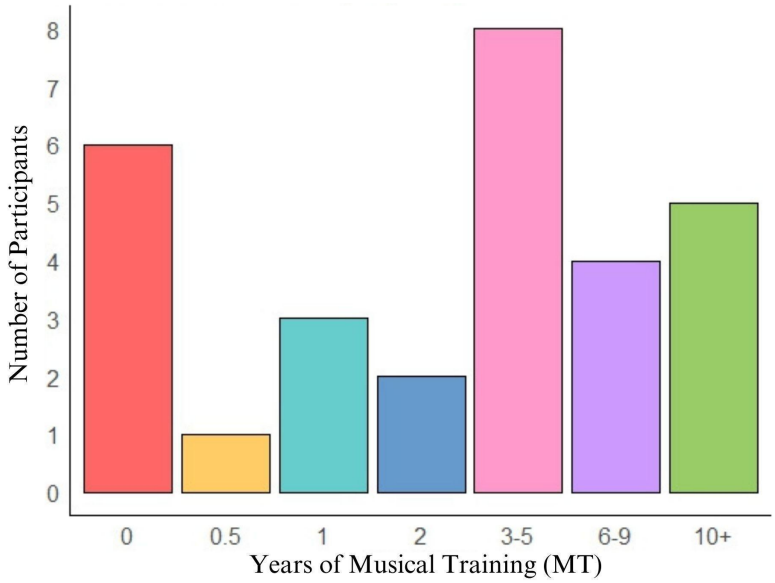
4.1 Participants

In total, 29 participants took part in the experiment. They were recruited using the participant database of the Institute for Language Sciences (ILS Labs) at Utrecht University, the researchers' own social circle and social networking sites, such as LinkedIn and Facebook. Participants were all Dutch monolinguals. Those with a history of hearing impairments or tinnitus, Autism Spectrum Disorder (ASD), DLD, Attention Deficit (Hyperactivity) Disorder (AD(H)D), dyslexia, and/or other neurological disorders were excluded from participation. Additionally, participants who are bald or have coily hair or dreadlocks were not invited to participate, as it is not possible to measure neural activity with electrode caps effectively. Participants were compensated for their time and efforts with a € 20,-- Yesty gift card sent to their e-mail address.

Of the 29 participants, four were male and 25 female. Their age ranged between 19 and 54 years old (M = 23.24, SD = 6.60). One participant had a notably high age of 54 years compared to the other participants whose age range was between 18 and 32 years old (M = 22.14, SD = 3.16). Regarding their level of education, two participants were currently or previously attending a university of applied sciences, and the other 27 participants were currently or previously attending university. Six participants had never followed Musical Training (MT), while the other 23 participants had at least 0.5 years of MT (see Figure 2 for the distribution of participants by the number of years of MT). Table A1 in Appendix A shows a full participant overview.

Figure 2

Distribution of Participants by Years of Musical Training (MT)



4.2 Stimuli

The stimuli include two artificial languages, namely a structured and a random stream, part of the experiment by van der Wulp et al. (2023):

> The stimuli consist of syllables which are combined into tri-syllabic non-words (from now on referred to as 'words') that adhere to Dutch phonotactics and have been piloted for their learnability [...]. 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 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 pausing 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 [...]. 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). (p. 13-14)

4.3 Study Design and Procedure

The data for this study is part of the project by van der Wulp et al. (2023), which makes the experimental design identical to their study. A visual representation of the experimental procedure is shown in Figure 3. This section will give a more detailed description of the tasks participants completed.

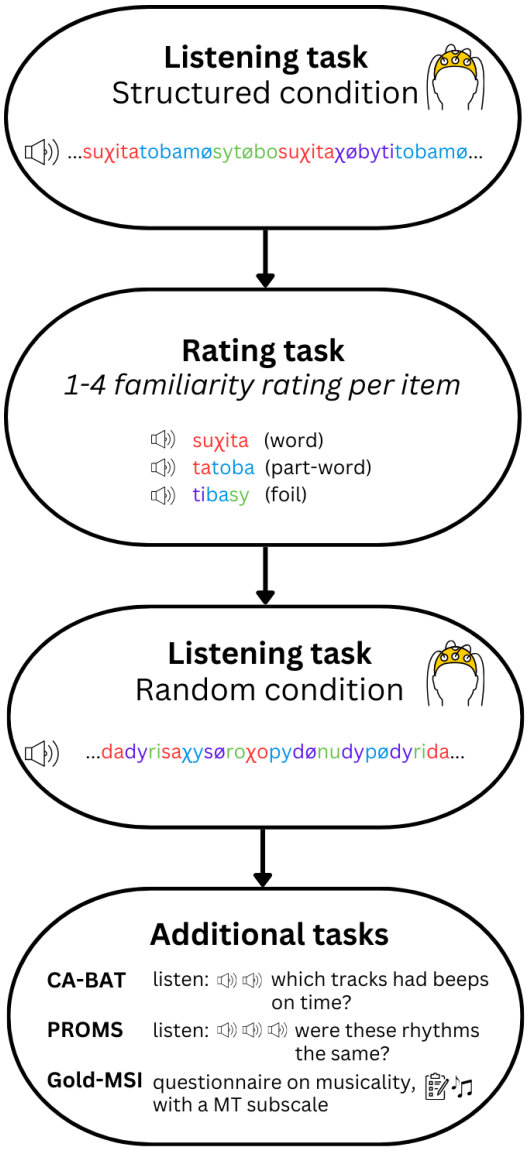
4.3.1 Listening Task

Participants started with listening to the structured stream, which took 13.5 minutes and was divided into three blocks of 4.5 minutes each. In between the three blocks, participants could take a break and press Enter when they wanted to continue with the experiment. These untimed breaks allowed participants to adjust the duration of their break to suit their individual needs. After the Rating Task, participants listened to the random stream that also consisted of three blocks of 4.5 minutes.

Participants were seated in a sound-attenuated booth. During both listening tasks, participants' neural oscillations were recorded with electroencephalography (EEG). Therefore, participants were instructed to listen as calmly as possible and minimise their blinking, to avoid too many unnecessary artefacts in the EEG data. They could use the break as an opportunity to move more freely.

Figure 3

Schematic Overview of the Experimental Procedure.



Note. Adapted from van der Wulp et al. (2023, Figure 2, p. 15).

4.3.2 Rating Task

The structured condition was immediately followed by the rating 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. 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).
>
> (Van der Wulp et al., 2023, p. 15)

Participants could take as long as they wanted to complete the task. Generally, this task took five minutes at most. Appendix B contains the full list of items used in this task.

4.3.3 Musicality Tasks

To measure musicality and rhythmic ability, three different tests were administered, namely the CA-BAT, the PROMS, and the Gold-MSI questionnaire.

> 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.
>
> (van der Wulp et al., 2023, p. 16)

All three measures exhibit good validity and reliability. The CA-BAT was contracted and validated in four empirical studies, which showed that the test achieves good reliability and validity in laboratory settings (Harrison & Müllensiefen, 2018). An independent evaluation determined that the PROMS displayed a good pattern of construct validity: it showed a remarkably good pattern of discriminant (i.e., whether it is *not* associated with a task measuring an unrelated concept, here: short-term or working memory), convergent (i.e., whether it correlates with another task measuring the same construct, here: other music ability tests) and criterion (i.e., whether it is related to 'real world' variables, here: years of MT, musicianship status, and musical talent) validity (Kunert et al., 2016). Finally, the Gold-MSI has good indices of internal consistency and test-retest reliability. Moreover, the Gold-MSI's subscales significantly correlate with auditory tests evaluating melodic memory (i.e., a test created by Müllensiefen and colleagues (2014) in which participants must indicate if two tunes have an identical pitch interval structure) and beat perception (i.e., Beat Alignment Test (BAT) by Iversen and Patel (2008)). This suggests that the instrument effectively measures individual performance levels, despite its self-reported nature (Müllensiefen et al., 2014).

The CA-BAT has a total of 25 questions, and the rhythm and accent subscales of the PROMS contain 18 questions (eight for rhythm; ten for accent) in total. Finally, the Gold-MSI includes a total of 39 questions, of which the following seven belong to the MT subscale:

- I have never been complimented for my talents as a musical performer.

- I would not consider myself a musician.
- I engaged in regular, daily practice of a musical instrument (including voice) for ___ years.
- At the peak of my interest, I practised ___ hours per day on my primary instrument.
- I have had formal training in music theory for __ years.
- I have had __ years of formal training on a musical instrument (including voice) during my lifetime.
- I can play ___ musical instruments.

4.4 EEG Hardware and Acquisition Settings

Electroencephalography (EEG) recordings were made with the BioSemi ActiveTwo system and the accompanying ActiView software. The EEG activity was recorded at a sampling rate of 512 Hz by using 64 BioSemi "pin-type" electrodes with a sintered Ag/AgCl-tip that were attached to a 64-channel electrode headcap from BioSemi using the 10-20 system. The Common Mode Sense (CMS) active electrode was used as the online reference channel, and the Driven Right Leg (DRL) passive electrode as the online ground channel. Additionally, six "flat-type" electrodes were attached to participants' left (EXG1) and right (EXG2) mastoid, the outer canthi of the left (EXG3) and right (EXG4) eye, and above (EXG5) and below (EXG6) the left eye. The two mastoid electrodes were used for offline re-referencing. The other four electrodes measured vertical and horizontal eye movements, such as blinking. To reduce impedance and improve conduction, Signa electrode gel was applied to the electrode holders in the headcap and the flat electrodes. The impedance of the electrodes was maintained under 20 mV. The local powerline frequency was 50 Hz.

4.5 Data

4.5.1 EEG Pre-Processing

Pre-processing steps occurred in the following order: importing, re-referencing, offline filtering, bad channel interpolation, artefact correction and rejection, and epoching. The EEG data will be processed through scripting in MATLAB (The Mathworks Inc., 2019), using EEGLAB (Delorme & Makeig, 2004) and the ERPLAB open-source toolbox (Lopez-Calderon & Luck, 2014).

First, the data was recorded with Actiview software and the *.bdf* files with all channels were imported into EEGLAB for MATLAB. The continuous data was filtered offline with a band-pass filter from 0.1 to 30 Hz and a notch filter of 50 Hz. Subsequently, bad channels identified during data collection or visual inspection of the data were interpolated in ERPLAB. A channel was deemed bad during the analysis if it was flagged as such during data collection due to high impedance (above 20 mV), or if it drifted or showed frequent noise upon visual inspection. Data sections containing large artefacts (e.g., large noise related to jaw muscle contractions), also identified through visual inspection, were manually rejected. Eye movement artefacts were not removed, because they are not time-locked to stimulus onsets and possess a broad power spectrum that does not interfere with narrow-band neural responses (Srinivasan & Petrovic, 2006). Finally, regarding epoching, the data was time-locked to the onsets of the trisyllabic words or every third syllable in the random condition, and segmented into non-overlapping epochs of 10.8 seconds, which corresponds to 12 words or 36 syllables per epoch. We used a Fast Fourier Transform (FFT) for each epoch across frequency bins of interest, namely between 0.6 to 5 Hz and with a bin width of 0.09 Hz (following Batterink & Choi, 2021; Benjamin et al., 2021; Moreau et al., 2022). Afterwards, we calculated the Inter-Trial Coherence (ITC) to quantify phase-locking to the word (1.1 Hz) and syllable (3.3 Hz)

frequencies. The ITC ranges from 0 to 1, with 0 indicating no phase-locking at all and 1 indicating perfect phase-locked neural activity to a certain frequency.

4.5.2 Analyses

Data was agreed to be excluded from the analysis if participants did not finish the experiment due to technical issues, if they wanted to retract or stop their participation during the experiment, or if they did not follow instructions during the experiment. Furthermore, EEG data from participants who did not show a clear ITC peak at the syllable frequency of 3.3 Hz must also be excluded, as this indexes basic auditory processing of syllables, in line with van der Wulp et al. (2023). We excluded EEG data from four participants because they did not show this ITC peak. Their behavioural data was kept and included in those analyses, because the absence of the ITC peak does not affect the validity or reliability of their behavioural responses on, for example, the musicality tasks. Moreover, the lack of the ITC peak does not necessarily indicate that participants were not listening to the stimuli; it could also be absent due to excessive noise in the data (e.g., caused by frequent movements or excessive blinking).

All analyses were performed in R (R Core Team, 2017). The car (Fox & Weisburg), lme4 (Bates et al., 2005), lmerTest (Kuznetsova et al., 2017) ordinal (Christensen, 2023), and tidyverse package (Wickham et al., 2019) were used. The statistical significance for each analysis was determined at an alpha level of $p < .05$ (Winter, 2019). Before examining the influence of MT (RQ1), possible correlations between the several musicality measures were checked (RQ2).

4.5.2.1 RQ2: Correlation Between Musicality Measurements

In this study, several behavioural measures aimed to measure the same construct. The PROMS and CA-BAT both measured rhythmic ability, the Gold-MSI measured musicality, and the

number of years of MT could also correlate with these abilities. This leaves us at risk of (multi)collinearity, i.e., the occurrence of significant correlations among two or more independent variables. This (multi)collinearity can make significant variables statistically insignificant when they are both included as predictors in a multiple regression model (Shrestha, 2020). To measure the degree of (multi)collinearity, we performed a Pearson's correlation between years of MT, the scores of the Gold-MSI's MT subscale, PROMS scores and CA-BAT scores. Furthermore, for the Linear Mixed Models (LMMs) (see section 4.5.2.2), we calculated the Variance Inflation Factor (VIF) for models with possible (multi)collinear variables. Correlation coefficients above .80, VIFs larger than 10, and an average VIF substantially larger than 1 indicate (multi)collinearity (Field et al., 2012; Shrestha, 2020). If this was the case, we either computed a composite variable that comprised the mean of the collinear variables or kept only one of the collinear variables.

4.5.2.2 RQ1: The Influence of Musical Training on Statistical Learning

Rating Task

First, we calculated binary accuracy scores from the familiarity ratings. For all stimuli, ratings of 1 or 2 were classified as 'not familiar' and ratings of 3 or 4 as 'familiar'. If participants gave words a familiar rating, they got an accuracy score of 1, because their judgment was correct (i.e., they correctly judged stimuli that appeared during the listening task as familiar). If they gave words an unfamiliar rating, they got an accuracy score of 0, as their answer was wrong (i.e., they incorrectly judged stimuli that appeared during the listening task as unfamiliar). For foils, it was the other way around. Each participant received a score representing the percentage of correct responses.

With the percentage of correct responses per participant (based on the binary accuracy scores), we performed a Shapiro-Wilk test to determine whether the data followed a normal

distribution, and a one-sided *t*-test to analyse whether performance was significantly above chance (> 50%). This indicated whether participants successfully segmented the words during the listening task and performed Statistical Learning (SL).

Afterwards, to test whether words were judged as more familiar than part-words and foils, Cumulative Linked Mixed Models (CLMMs) were built. The dependent variable was the familiarity score. First, a null model with only random intercepts for participants and item was created (model 0). Subsequently, a new model was created with word type (word, part-word or foil) as a predictor and random intercepts for participants and item (Model 1). A likelihood-ratio test showed if model 1 was a better fit than model 0, and, therefore, if SL took place on the behavioural level. Afterwards, we systematically added several predictors to our model to determine the influence of MT. Respectively, we added years of MT (Model 2)[1], CA-BAT scores (Model 3), and PROMS scores (Model 4) as predictors. Likelihood-ratio tests showed which model is the best fit for our data.

Our preregistered analysis plan also included the Gold-MSI's MT subscales scores as a predictor (see Brandsen et al., 2024). The results of RQ2, however, showed collinearity between years of MT and the Gold-MSI's MT subscale scores. Therefore, we decided to use only years of MT as a predictor and discard the Gold-MSI (see section 5.1 for the analysis).

EEG

To analyse the occurrence of SL in the EEG data, we computed Word-Learning Index (WLI) values for both conditions as a mean for each participant over the entire exposure period:

$$WLI = \frac{ITC\ of\ word\ frequency}{ITC\ of\ syllable\ frequency}$$

[1] Participants' number of years of MT was assessed by a Likert-scale question in the Gold-MSI questionnaire, which is why this variable was treated as a factor. The categories, however, were not equally distanced.

LMMs were built with WLI as the dependent variable. First, we created a null model with only random intercepts for participants (model 0). To discover whether participants performed SL, a new model with language condition (structured or random) as a predictor and random intercepts for participants was created (model 1). To establish whether language condition significantly improved the model, a likelihood-ratio test between models 0 and 1 was conducted ($p < .05$; Winter, 2019). The model that showed the best fit to the data was selected for further analyses. Then, to determine whether MT could account for individual differences in WLI, we systematically added several predictors to our model and compared models with likelihood-ratio tests to determine the best-fitting one. First, we added years of MT as a predictor (model 2). Second, CA-BAT scores were included in the model (model 3), and, finally, we included PROMS scores (model 4). Again, because the results of RQ2 showed collinearity between years of MT and the Gold-MSI's MT subscale scores, only years of MT was included as a predictor.

5. Results

5.1 RQ2: Correlations Between Musicality Measures

A Pearson correlation coefficient was computed to assess the degree of (multi)collinearity between years of Musical Training (MT), the Gold-MSI's MT subscale, the CA-BAT scores, and the PROMS scores. There was a very strong positive correlation between years of MT and the Gold-MSI's MT subscale ($r = .86$, $p < .001$). Additionally, there was a weak to moderate positive correlation between the CA-BAT and both years of MT ($r = .39$, $p = .04$) and the Gold-MSI's MT subscale ($r = .47$, $p = .01$). Table 1 shows an overview of all correlation coefficients and Appendix D contains the scatter plots for each correlation.

Table 1

Correlations Between the Gold-MSI's MT Subscale, Years of MT, CA-BAT Scores, and PROMS Scores.

Variable		Gold-MSI_MT	Years of MT	CA-BAT
Years of MT	Pearson's *r*	.86***	–	
	p-value	<.001	–	
CA-BAT	Pearson's *r*	.47*	.39*	–
	p-value	.01	.04	.04
PROMS	Pearson's *r*	.36	.34	.34
	p-value	.05	.07	.07

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

Because the coefficient of the correlation between years of MT and the Gold-MSI's MT subscale is above .80, there is a risk of collinearity. To resolve this issue, we opted to use only one of the variables, namely years of MT. That variable more directly aligns with the focus of this thesis than a variable of the Gold-MSI's MT subscale scores, because the MT subscale also contains questions that are more indirectly related to MT (e.g., "I have never been complimented for my talents as a musical performer"). Computing a composite variable would not be beneficial here, because years of MT is a subquestion of the Gold-MSI. The remaining significant correlations are relatively weak to moderate, indicating that the risk of collinearity is minimal. Finally, the VIF of our best-fitting LMM was below 10 and did not indicate problematic collinearity (see section 5.2.1.2).

5.2 RQ1: The Influence of Musical Training on Statistical Learning

To maintain focus on the most relevant findings, only the best-fitting models are reported in this section. Appendix E shows an overview of the results of the likelihood ratio tests for all

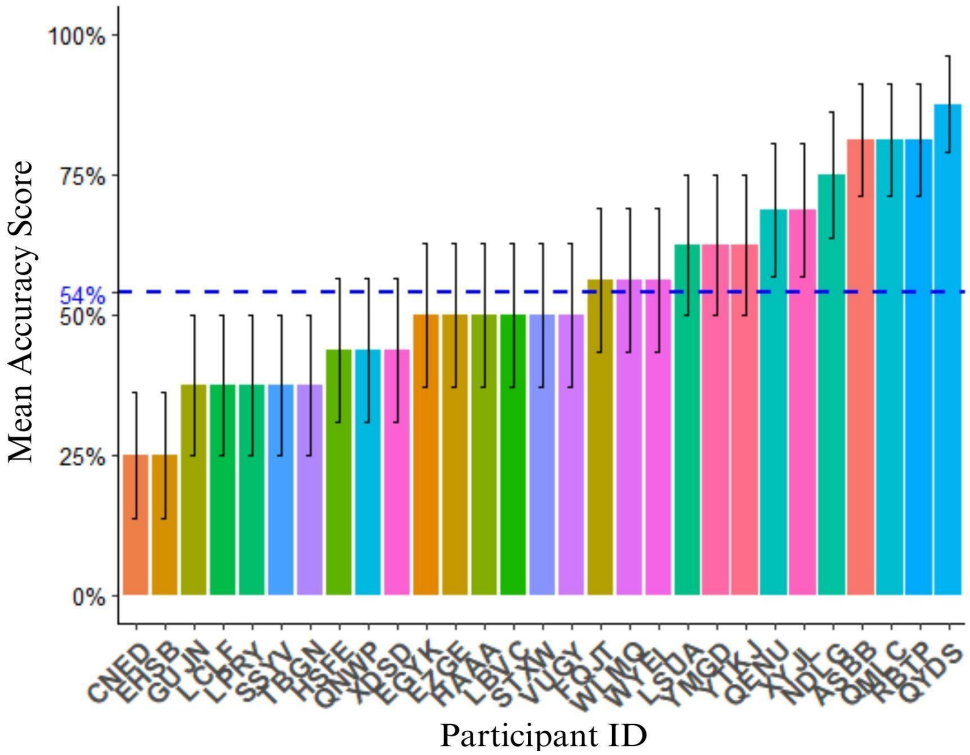
models and the emmeans contrasts. Based on the results of the planned analyses, several exploratory analyses were conducted.

5.2.1 Rating Task

The Shapiro-Wilk test indicated no significant deviation from normality for the binary accuracy score data ($W = 0.96$, $p = .25$), implying a normal distribution (see Appendix C for the Q-Q plots of the rating task data). A one-sided *t*-test revealed that participants did not significantly score above chance ($M = 54\%$, $SD = 16.8\%$) on their rating accuracy ($t(28) = 1.31$, $p = .20$). The mean scores per participant are shown in Figure 5. The mean familiarity ratings per participant, on which these accuracy scores are based, can be seen in Figure F1 in Appendix F.

Figure 5

Mean Accuracy Scores per Participant and the Overall Mean Accuracy Across All Participants (in %)



The best-fitting model was Model 1[2] with familiarity rating as a dependent variable and word type as a predictor (see Table 2 for the model coefficients). It was significantly better than the null model[3] ($\chi^2(2) = 12.37$, $p = .002$). There was a main effect of word type: pairwise comparisons using emmeans indicated that words were rated significantly higher than part-words ($\beta = 0.76$, $p < .001$) and foils ($\beta = 0.95$, $p < .001$) (see Table E3 in Appendix E). The difference in rating between part-words and foils was, however, not significant ($\beta = -0.20$, $p = 0.68$).

[2] model1 = clmm(resp ~ wordtype + (1|ppID) + (1|item), data=data_rating, REML=FALSE)
[3] model0 = clmm(resp ~ (1|ppID) + (1|item), data=data_rating, REML=FALSE)

Table 2

Model Coefficients of Model 1 of the Rating Task Data with Familiarity Rating as the Dependent Variable.

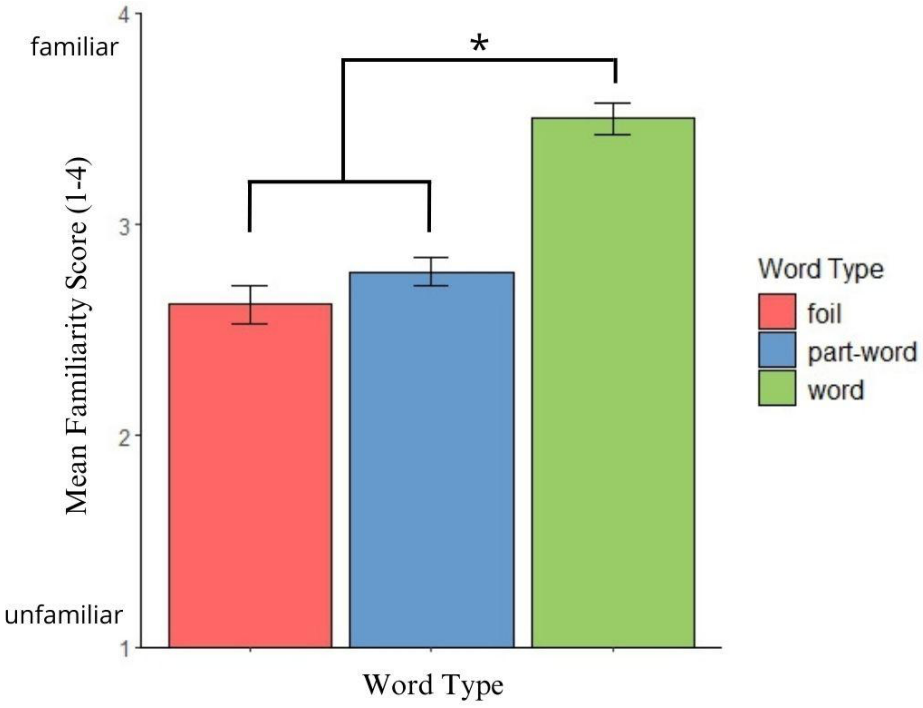
Fixed Effects	*β*	*SE*	*z-value*	*p-value*
word type: partword	0.37	0.45	0.83	.41
word type: word	2.06	0.53	3.85	< .001***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. The foil word type is the reference category, which is implicitly set to 0 in this model.

Figure 6 shows the difference in mean familiarity ratings between the different word types. It can be seen that words were rated significantly more familiar than part-words and foils.

Figure 6

Mean Familiarity Score (1-4) per Word Type



5.2.1.1 Exploratory Analyses and Results

Exploratorily, we looked further into the accuracy scores by creating Generalised Linear Mixed Models (GLMMs). The dependent variable was accuracy score. The null model included only random intercepts for participants and item (Model 0). Similar to the CLMMs with familiarity score, we systematically added several predictors, namely word type (Model 1), years of MT (Model 2), CA-BAT scores (Model 3), and PROMS scores (Model 4). Likelihood-ratio tests determined the best-fitting model.

Model 2[4] was chosen as the final, best-fitting model, because it provided a significantly better fit than Model 1 ($\chi^2(6) = 13.25$, $p = .04$). Table 2 shows the model coefficients. Pairwise comparisons with emmeans revealed that words were rated significantly more accurate than part-words ($\beta = 2.61$, $p < .001$) and foils ($\beta = 2.41$, $p < .001$) (Figure F2 in Appendix F shows the difference in accuracy scores across word type and Table E5 in Appendix E shows the emmeans contrasts). Additionally, there was a significant effect of 1 ($\beta = 1.38$, $p = .01$) and 10+ year(s) of MT ($\beta = 1.01$, $p = .01$) compared to 0 years of MT. However, these effects disappeared after performing emmeans and contrasts with a Tukey HSD correction of word type and MT.

[4] model2ac = glmer(famrating_accur ~ wordtype + Years_MT + (1|ppID) + (1|item), data=data_rating, family = binomial)

Table 2

Model Coefficients of Model 3 of the Rating Task Data with Accuracy Score as the Dependent Variable.

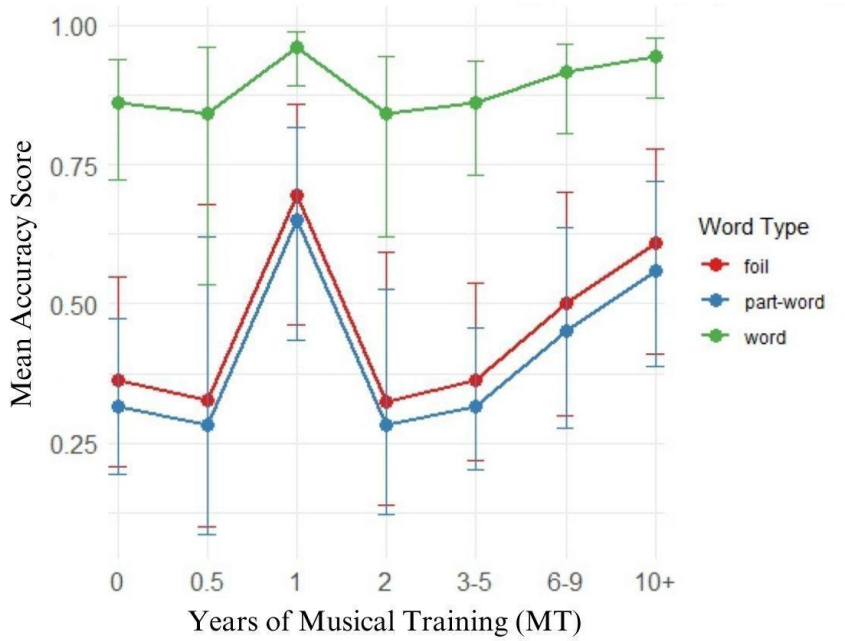
Fixed Effects	*β*	*SE*	*df*	*p-value*
(Intercept)	-0.57	0.39	-1.45	.15
word type: part-word	-0.20	0.35	-0.57	0.57
word type: word	2.40	0.47	5.11	< .001***
Years_MT2	-0.16	0.76	-0.22	.83
Years_MT3	1.38	0.51	2.72	.01*
Years_MT4	-0.16	0.57	-0.29	.77
Years_MT5	0.00	0.38	0.00	1.00
Years_MT6	0.57	0.45	1.27	0.21
Years_MT7	1.01	0.43	2.37	.02*

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Years_MT1 = 0 years; 2 = 0.5 years; 3 = 1 year; 4 = 2 years; 5 = 3-5 years; 6 = 6-9 years; 7 = 10 + years. The intercept represents the non-word word type.

Figure 7 shows the effect of MT on the mean accuracy scores per word type. It can be seen that in general, words were rated significantly more accurate than part-words and foils, regardless of years of MT. Furthermore, the figure reveals a notable peak in accuracy score at one year of MT for all word types. After two years of MT, there is an upward trend and the accuracy scores seem to progressively increase with more years of MT.

Figure 7

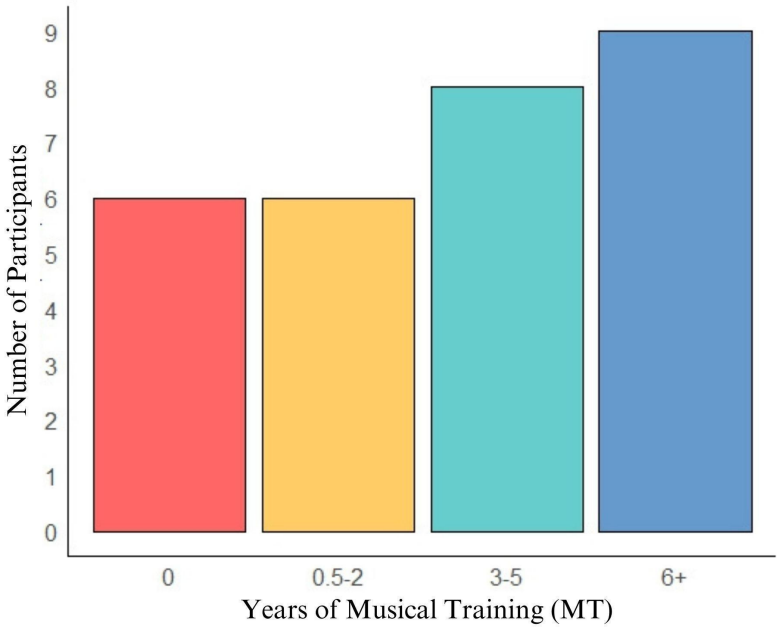
The Effect of Years of Musical Training (MT) on the Mean Accuracy Scores



Inspection of Figure 3 with the distribution of participants by years of MT (see section 4.1) reveals an uneven distribution across the categories, with relatively few participants in most categories (e.g., only three participants with 1 year of MT compared to eight with 3-5 years). This sparse distribution may lead to misleading results, such as the peak at 1 year of MT in Figure 7. This peak could result from the limited number of participants in the 1 year category: a small sample size may distort the apparent impact of MT on the accuracy ratings. The peak could reflect an artefact of the distribution rather than a true effect. To address this potential issue, we opted for a second exploratory analysis. New categories of years of MT (0 years, 0.5 to 2 years, 3 to 5 years and 6+ years) were created to achieve a more equal distribution of participants and a larger number of participants per category. Figure 8 shows the distribution of participants across these new categories.

Figure 8

Distribution of Participants Across New Categories of Years of Musical Training (MT)



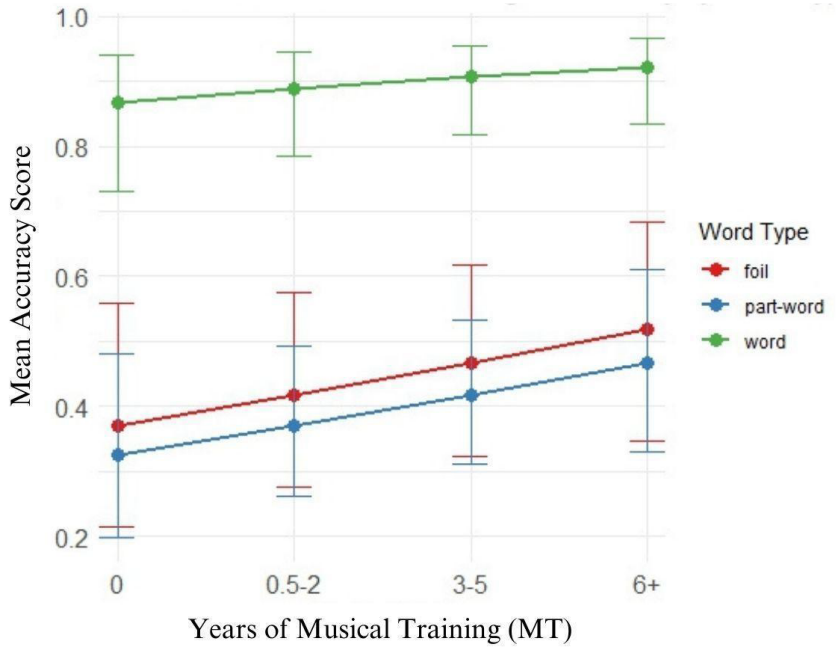
With these new categories, we redid the analyses on familiarity ratings and accuracy scores. Regarding the familiarity rating, this did not significantly change the results: Model 1 with only word type remained a better fit compared to the new Model 2[5] with the new categories of years of MT ($\chi^2(3) = 4.15$, $p = .24$). With respect to the accuracy scores, the new categories did change the results. In the new Model 2[6], 6+ years of MT has a significant effect on accuracy scores compared to 0 years of MT ($\beta = 0.83$, $p = .049$), but this effect disappeared after conducting emmeans and contrasts. Figure 9 shows the effect of MT on the accuracy scores using the new MT categories. It shows that the peak observed at 1 year in Figure 7 has dissipated in Figure 9, leading to a consistently rising trend. However, this trend did not reach significance. While Model 2 with word type and the default categories of MT had proven to be the best fit, the new Model 2 was not: Model 1 with only word type fits the data better ($\chi^2(3) = 6.00$, $p = .11$) and is, therefore, the final model.

[5] model2_cat = clmm(resp ~ wordtype + cat_MT + (1|ppID) + (1|item), data=data_rating, REML=FALSE)

[6] model2ac_cat = glmer(famrating_accr ~ wordtype + cat_MT (1|ppID) + (1|item), data=data_rating, family=binomial)

Figure 9

The Effect of Years of Musical Training (MT) on the Mean Accuracy Scores (New Categories)



Additionally, we looked at the accuracy percentages (i.e., the percentage of correct ratings). An LMM with accuracy percentages as the dependent variable was built. Again, like in the previous analyses, we added multiple predictors to the best-fitting model in the following order: years of MT (Model 1), CA-BAT scores (Model 2), and PROMS scores (Model 3). We did not add word type because the accuracy percentages are aggregated per participant, resulting in a single value per participant. Consequently, there is no within-participant variability in accuracy percentages related to word type. None of the predictors exhibited an improvement over the null model[7]: likelihood-ratio tests showed that neither Model 1[8] ($\chi^2(1) = 0$, $p = 1$) and Model 2[9] ($\chi^2(1) = 0$, $p = 1$), nor Model 3[10] ($\chi^2(1) = 0$, $p = 1$) were a significantly better fit.

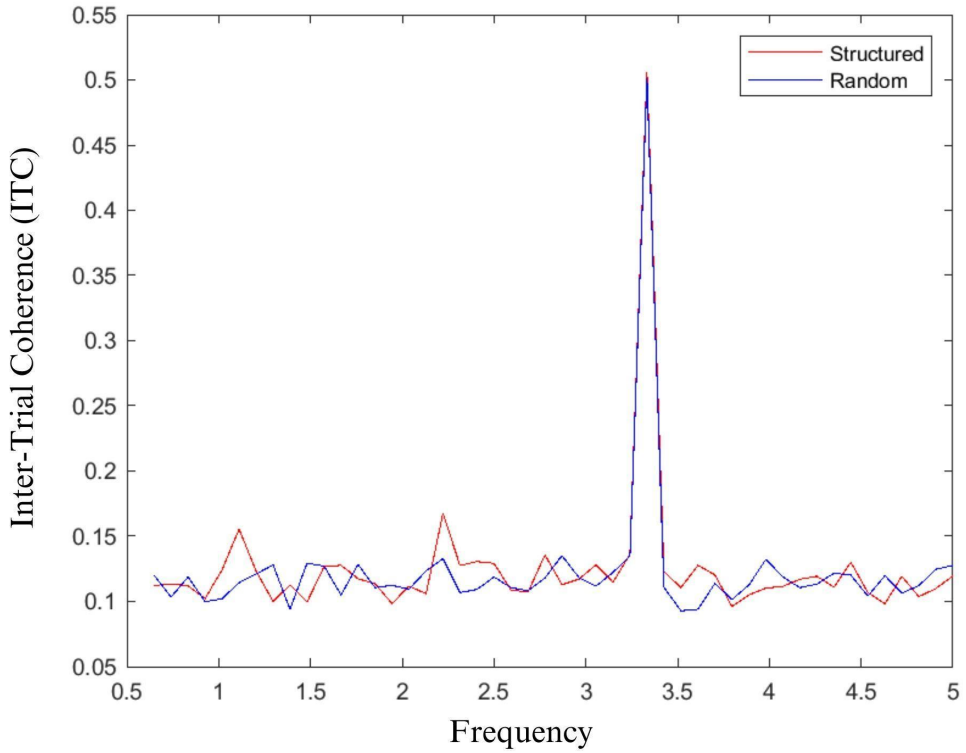
[7] model0ac_pct = lmer(accur_percentage ~ (1|ppID) + (1|item), data=data_rating, REML=FALSE)
[8] model2ac_pct = lmer(accur_percentage ~ Years_MT + (1|ppID) + (1|item), data=data_rating, REML=FALSE)
[9] model3ac_pct = lmer(accur_percentage ~ CA.BAT + (1|ppID) + (1|item), data=data_rating, REML=FALSE)
[10] model4ac_pct = lmer(accur_percentage ~ PROMS_average + (1|ppID) + (1|item), data=data_rating, REML=FALSE)

5.2.2 EEG Data

Visual inspection of the electroencephalography (EEG) data showed a clear Inter-Trial Coherence (ITC) peak at the syllable frequency of 3.3 Hz in both the random and structured condition (see Figure 4). In the structured condition, there is also a peak at the word frequency of 1.1 Hz, suggesting neural entrainment to the trisyllabic words. Appendix G contains an additional figure of the frequency peaks and topographical plots showing the mean ITC distribution across the scalp for both conditions and frequencies (word and syllable).

Figure 4

Frequency Peaks for the Structured (Red) and Random Condition (Blue) (0-5 Hz).



Note. In both conditions, there is a peak at 3.3 Hz (syllable frequency). The structured condition also includes a peak at 1.1 Hz (word frequency).

Model 3[11] with Word-Learning Index (WLI) as the dependent variable and condition and CA-BAT as predictors was the best-fitting model with a significant main effect of CA-BAT (p = .04). In line with the observations from our visual inspection, there was also a main effect of condition (p = .01). The model coefficients can be seen in Table 3.

Table 3

Model Coefficients of Model 3 of the EEG Data with WLI as the Dependent Variable

Fixed Effects	*β*	*SE*	*df*	*t-value*	*p-value*
(Intercept)	0.24	0.02	25.93	11.05	< .001***
condition: structured	0.09	0.01	775.00	15.44	< .001***
CA.BAT	0.04	0.02	25.00	2.14	.04*

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. The intercept represents the random condition.

5.2.2.1 Exploratory Analyses and Results

Similar to the rating task analysis in section 5.2.1.1, we exploratorily analysed the EEG data with new categories of MT (0 years; 0.5-2 years; 3-5 years; 6+ years), shown in Figure 8. This, however, did not alter the results: similar to the previous Model 2 with the default categories, the new Model 2[12] with the new categories of years of MT as a predictor did not prove to be a better fit than Model 1 with only condition as a predictor ($\chi^2(1) = 0.16$, p = .69). Thus, Model 3 with condition and CA-BAT as predictors remains the best fitting model.

Additionally, we explored the influence of MT on syllable ITC. Research shows that MT positively influences auditory processing (see Neves et al. (2022) for a systematic review and meta-analysis). This can be examined by analysing the syllable ITC, because larger neural entrainment to the syllable frequency indicates enhanced sensory processing of individual

[11] EEGmodel3 = lmer(WLI_ROIchans ~ condition + CA.BAT + (1|ppID), data=data_EEG, REML=FALSE)
[12] EEGmodel2_cat = lmer(WLI_ROIchans ~ condition + cat_MT + (1|ppID), data=data_EEG, REML=FALSE)

syllables (Batterink & Paller, 2017). To this end, a null model[13] with syllable ITC as the dependent variable and random intercepts for participants was created. Systematically, we added years of MT, CA-BAT, and PROMS scores as predictors. Model 4[14] with PROMS scores proved to be the best-fitting model for the data. It was significantly better than the null mo2del ($\chi^2(1) = 4.35$, $p = .04$) with a main effect of PROMS scores ($p = .04$).

Finally, given the discrepancy in outcomes between the EEG data and the rating task data, we sought to determine the extent to which the listening tasks with EEG and the rating task measure the same construct and should yield comparable results. To this end, we computed correlations between the WLI and accuracy scores. The results show a non-significant, albeit slightly negative correlation between the two scores ($r = -.03$, $p = .06$) (see Figure G3 in Appendix G for the correlation plot).

6. Discussion

This study aimed to establish the influence of musical training (MT) on the statistical learning (SL) of speech segmentation in adults by using both online and offline methods, namely electroencephalography (EEG) and a rating task. We hypothesised MT to have a positive influence on SL, with musically trained participants showing a higher Word-Learning Index (WLI) than untrained participants. Prior to that, we examined the relation between the different MT and musicality measurements used in the experiment, namely the number of years of MT, the Gold-MSI's MT subscale, the CA-BAT and the PROMS. We expected a positive relationship between all measures.

[13] Syllmodel0 = lmer(SyllITC_ROIchans ~ (1|ppID), data=data_EEG, REML=FALSE)
[14] Syllmodel3 = lmer(SyllITC_ROIchans ~ PROMS_average + (1|ppID), data=data_EEG, REML=FALSE)

6.1 RQ2: Correlations Between Musicality Measures

RQ2: To what extent is musical ability, measured with the CA-BAT and PROMS, related to (years of) musical training?

The results showed a very strong positive correlation between years of MT and the Gold-MSI's MT subscale, confirming that participants with more years of MT tend to score higher on the Gold-MSI's MT subscale. While this result is not entirely unexpected, given that years of MT is a sub question of the subscale, it nonetheless reinforces the validity of the MT subscale by demonstrating that it accurately reflects an individual's MT experience. In line with previous research by Müllensiefen and colleagues (2014), the MT subscale of the Gold-MSI is confirmed as a valid self-report measure. Unfortunately, we do not have data to confirm the validity of the entire Gold-MSI questionnaire or discuss the relationship between the MT subscale and the other subscales, because we only analysed the MT subscale scores.

The significant, positive correlation between the CA-BAT and both years of MT and the Gold-MSI's MT subscale is weak to moderate. This means that, while there is a positive relationship between the number of years of MT and performance on the CA-BAT, the correlation is not very strong. MT contributes to, but does not fully explain, performance on the CA-BAT, and other factors, such as innate rhythmic abilities, may also affect CA-BAT scores. Conversely, high CA-BAT scores do not automatically indicate extensive MT, as other variables may also influence these scores. Overall, this suggests that musical ability is not necessarily achieved solely through MT, corroborating the findings of Correia and colleagues (2023) that musical ability is multifaceted, and that MT is not necessary for developing musician-like abilities. However, as discussed in section 2.2.1, previous studies that investigated the influence of musicality traditionally rely on years of MT as the primary criterion for defining musicians. Initially, the positive correlation between the CA-BAT and

MT may seem to affirm this decision, but it becomes less convincing when taking into account that the correlation is moderate at best. Instead, the correlation underscores the need for more nuanced and comprehensive measures of musicality. We believe that participants' musicality should be assessed using more direct measures than just the number of years of MT, and, ideally, by employing a range of different assessments.

The CA-BAT and PROMS did not significantly correlate: a high score on the CA-BAT does not necessarily equal a high score on the PROMS, or vice versa, implying that both tasks measure different constructs. This result seems remarkable, given that both tasks aim to assess musical, and more specifically rhythmic, abilities. However, upon closer examination, this is not as striking as it initially appeared to be. For the PROMS, in which participants must remember a musical piece to correctly compare the third one to the previous ones, musicality is not the only aspect being assessed; memorisation also plays a significant role. In contrast, the CA-BAT requires participants to decide which of two rhythms has beeps in synch, a task where memorisation is not essential. If participants do not (fully) remember the first rhythm, they can still take an educated guess based on whether the beeps in the second rhythm were in or out of sync. For the PROMS, this strategy is impossible because making a comparison requires having a reference (the first musical piece) in mind. This distinction highlights the importance of considering the specific cognitive processes involved in different musical tasks when interpreting correlations between them.

Additionally, the PROMS did not correlate with neither years of MT, nor the Gold-MSI's MT subscale. This, again, is noteworthy, considering that an independent evaluation of the task showed that it was indeed related to years of MT (Kunert et al., 2016). A possible explanation is that the PROMS taps into aspects of musicality that are less dependent on MT and more influenced by innate perceptual abilities or cognitive skills, such as working memory, processing speed and attention, which are not directly enhanced through MT. Individual

differences in these cognitive abilities could play a significant role in PROMS performance, independently of one's MT background. However, there is still a debate about whether MT enhances cognitive abilities or whether individuals who follow MT already possess higher cognitive abilities from the start (Escobar et al., 2020).

Another potential explanation is that the correlations could not reach significance due to the small sample size of this study. The near-significant *p*-values (PROMS - Gold-MSI's MT subscale: $p = .05$; PROMS - years of MT: $p = .07$; PROMS - CA-BAT: $p = .07$) suggest that these correlations might reach significance with more statistical power. Another possibility is that our selection of only two PROMS subsets, namely rhythm and accent, has influenced the results. The PROMS is designed to assess a wide range of musical abilities (Zentner & Strauss, 2017), and by limiting our focus to just two subsets, we may have missed capturing the broader spectrum of musicality. Perhaps including the other subsets would alter the results.

Summarising, the results indicate a weak to moderate relationship between musical ability, as measured by the CA-BAT, and (years of) MT. However, musical ability measured with the PROMS is not related to MT. This implies that musical ability is influenced by a broader range of factors beyond MT alone.

These findings contribute to the ongoing debate about the nature of musical ability and the influence of MT. It challenges the traditional reliance on years of MT as the primary indicator of musical ability (as employed by, for example, François and colleagues (2014) and Zhang and colleagues (2018)), which may oversimplify the complexity of musicality. The idea that musicality encompasses a range of factors beyond the number of years of MT suggests a need for a more nuanced approach that integrates several different measures.

6.2 RQ1: The Influence of Musical Training on Statistical Learning

RQ1: To what extent does musical training influence the statistical learning ability aiding speech segmentation in adults?

As expected and in line with previous research (e.g., Batterink & Paller, 2017), evidence for SL was observed in both the rating task and the EEG data. In the rating task, word type had a significant effect on the familiarity ratings and accuracy scores. Participants rated words significantly higher and more accurately than part-words and foils, which indicates that they learned the words from the structured stream and performed SL. The difference between part-words and foils was insignificant, meaning that participants assigned comparable ratings to both word types. They did not perceive part-words to be significantly more challenging to judge correctly. This finding contrasts with the results reported by Batterink and Paller (2017), who demonstrated that participants rated words as the most familiar, followed by part-words, with foils rated as the least familiar. The absence of a (significant) difference between part-words and foils could be attributed to a response bias, particularly a "yes-bias" (Fritzley & Lee, 2003), where participants rely on a familiarity heuristic, rating sequences with familiar syllables as familiar even if they do not fully recognise the sequence. Furthermore, a conservatism bias may prompt participants to avoid extreme ratings (1 or 4) and give both part-words and foils a similar familiarity rating of fairly (un)familiar (2 or 3) (Deason et al., 2017). Moreover, the items included twelve part-words and foils, but only four words. This imbalance (i.e., rating only four items as familiar and twelve items as unfamiliar) could feel counterintuitive. In combination with a potential yes-bias, which could lead participants to rate unfamiliar sequences as familiar, this may result in higher ratings for part-words and foils.

The EEG data also showed that SL occurred: the ITC peak at 1.1 Hz (word frequency) in the structured condition suggests neural entrainment to the trisyllabic words. Furthermore,

there was a significant effect of condition: the WLI was significantly higher in the structured condition compared to the random condition, supporting the occurrence of SL in the structured condition.

To establish the influence of MT, we first looked at the rating task data. Concentrating on the familiarity ratings first, neither MT nor the CA-BAT or PROMS had an effect on the ratings. The new categories of years of MT did not change the results; only word type affected the familiarity ratings. On the accuracy scores, however, MT did have a significant effect: participants with MT experience gave more accurate ratings than those without MT experience, indicating that MT benefits SL. The effects of 1 and 10+ years compared to 0 years of MT imply that participants with only one year of MT showed a significant improvement in their accuracy compared to their untrained counterparts. Additionally, 10+ years also makes a significant difference compared to zero years. These results, however, are based on an uneven distribution and low number of participants per category of years of MT. With the new categories of years of MT, the main effect of MT disappeared. The significant effect of one year of MT was found to be a misleading artefact. Nonetheless, there is a clear trend showing that accuracy scores tend to increase with more years of MT (see Figure 9), with 6+ years of MT reaching significance. Therefore, we believe that the sample size of this experiment might have prevented these results from reaching significance and that more statistical power might have yielded a significant effect of MT on accuracy scores. That would be consistent with previous research by, for example, François and colleagues (2014) who showed an MT effect in an experiment with an artificial language of sung pseudowords. Finally, MT did not affect the accuracy percentages.

For the EEG data, the results were rather different. Contrary to our expectations, based on the results from the rating task and previous studies such as François et al. (2014) and Mandikal Vasuki and colleagues (2017), we did not find an MT effect on participants' WLI.

Surprisingly, instead of MT, CA-BAT scores had a significant effect on the WLI. Participants with higher CA-BAT scores exhibited a higher WLI, indicating greater SL for these participants during the exposure phase. The new categories of MT did not alter these results. A possible explanation can be found in previous research only briefly discussed in section 2.2. Such studies found a correlation between rhythmic (perception) ability (i.e., the ability to accurately extract and synchronise to an auditory pulse) and language outcomes, and revealed that people can process both musical and linguistic rhythm through neural entrainment (e.g., Cirelli, 2016; Di Liberto et al., 2020; Nozaradan, 2014). Additionally, Peelle and Davis (2012) suggest that more precise synchronisation of neural oscillations to auditory input leads to better processing. Following this, van der Wulp and colleagues (2023) hypothesise that efficient phase-locking could be supported by rhythmic abilities and that, in turn, rhythmic abilities support SL (see Van der Wulp et al. (2023) for details).

While both the rating task and EEG data indicate the occurrence of SL, the discussed results present a clear contrast. Moreover, we found a non-significant but slightly negative correlation between both, meaning that – when significant – as participants' WLI increased, their rating accuracy slightly decreased, and vice versa. This correlation did not reach significance, but there is still a difference in results between these measures. This discrepancy may arise from the distinct ways each method captures aspects of SL: the rating task is an explicit measure that relies on participants' subjective judgements and conscious recall of learned words, while EEG is an implicit measure that can capture SL while it is happening and dissociate the word identification component from the explicit memorisation component (Batterink & Paller, 2017). Years of MT is an explicit measure of musicality, while the CA-BAT assesses rhythmic abilities more implicitly. Previous studies have found that MT is associated with an advantage in working memory, which could explain the positive influence of MT on the rating task data. The influence of participants' rhythmic abilities on the implicit

EEG could be explained with van der Wulp et al.'s (2023) hypothesis, mentioned in the previous paragraph. There is, however, no consensus yet about whether MT enhances cognitive abilities or if those with MT already have higher cognitive abilities to begin with (Escobar et al., 2020). Moreover, this would contradict our explanation for the lack of correlation between years of MT and PROMS (see section 6.1).

Finally, we expected MT to influence syllable Inter-Trial Coherence (ITC), because previous studies like Neves and colleagues (2022) demonstrated that MT enhances auditory processing, indicated by larger syllable ITC values. Interestingly, the PROMS, rather than MT, had a significant impact on the syllable ITC. Participants with high PROMS scores showed higher phase-locking to the syllables and, thus, showed better auditory processing than their counterparts with lower PROMS scores. Based on this, we can conclude that participants with MT experience or high CA-BAT scores did not have significantly better auditory processing skills than other participants. Since the PROMS did not influence the WLI and auditory processing was not influenced by CA-BAT scores, we can infer that differences in WLI are not due to auditory processing but are genuinely driven by SL. If auditory processing was responsible, the PROMS would significantly affect the WLI and participants with higher PROMS scores would have shown higher WLIs, which was not the case. This suggests that a higher WLI reflects actual learning rather than superior auditory perception skills, supporting the validity of the WLI and the findings of Batterink and Paller (2017) who introduced the WLI as a measure for SL.

Summarising, the rating task and EEG data yielded mixed results: according to the rating task data, participants with MT experience had better accuracy scores and, therefore, showed better SL than those without. The EEG data, however, demonstrated that only CA-BAT scores positively influenced SL. Considering this, musicality appears to be positively related to SL, whether in the form of MT experience or rhythmic abilities. While future research

is needed to confirm these observations, participants with enhanced musicality – either in the form of MT experience or superior rhythmic abilities – seem to demonstrate enhanced SL of speech segmentation. Perhaps the specific aspect of musicality that affects SL depends on the type of measure (here: an explicit rating task or implicit EEG).

These results deepen our knowledge of the factors underlying individual differences in SL for word segmentation and enhance our understanding of the influence that MT and musical and rhythmic abilities have on, in this case, SL. Similar to the results of RQ2, the findings corroborate the idea that musicality consists of several aspects that need to be considered when studying the influence of musicality on phenomena. Finally, the different results between the rating task and the EEG data underline the importance of using multiple measures to assess SL to provide a more comprehensive understanding of different aspects of SL.

6.3 Limitations

This study had three important limitations that should be considered, some of which have already been mentioned in sections 6.1 and 6.2. The first limitation is our small and homogeneous sample of participants consisting of only 29 participants for the rating task and 25 for the EEG listening task (after exclusion), most of whom were female and (had) attended university. The small sample size reduces statistical power and could limit the ability to detect significant effects, as suggested for the rating task data with the new categories of MT and the PROMS correlations. Moreover, the results could present a distorted view due to the homogeneity of our sample. Therefore, future research should consider using a larger and more heterogeneous sample of participants, for example by ensuring a broader range of educational backgrounds.

The second limitation concerns the years of MT variable, which we based on a question from the Gold-MSI. This Likert-scale question had fixed categories, namely 0, 0.5, 1, 2, 3-5,

6-9 and 10+ years of MT. These unequally distanced categories prohibited us from differentiating between, for instance, three and five years of MT. Moreover, we did not include a question allowing participants to elaborate on their answers, leaving us unaware of the frequency and intensity of their MT sessions. For instance, participants with two years of MT might have engaged in MT sessions twice a week or only once a month, and this variability could have influenced the results. Future research on this topic should keep this in mind.

Finally, we only administered two subsets of the PROMS and focused exclusively on the MT subscale of the Gold-MSI in our analysis. Including the entire PROMS might have offered a more comprehensive assessment of musical and rhythmic abilities, potentially altering the results. This was not possible, however, because the data we used (from Van der Wulp et al., 2023) only contained the rhythm and accent subsets. The Gold-MSI was administered entirely, but we decided to focus on the MT subscale, discarding the other subscales and the mean score. In hindsight, it would have been interesting to include these as well to enrich the analysis and potentially offer a more nuanced understanding of the influence of musicality. Due to time constraints, however, we only analysed the Gold-MSI's MT subscale.

6.4 Future Research

Based on our findings, we can give several recommendations for future research. First, given our mixed results about the influence of MT and rhythmic abilities, more research is needed to disentangle the specific aspects of musicality that influence SL and how they contribute to it. It could, furthermore, be interesting to investigate how aspects of musicality interact with different SL measures (e.g., offline vs. online).

Another idea for future research is to use MT as an intervention, rather than merely examining pre-existing differences in MT experience as we did. This allows for a more

controlled and systematic investigation of how (specific aspects of) MT influence SL, as researchers can standardise the type, duration, and intensity of the training. Additionally, implementing pre- and post-tests (with different artificial languages) can track individual progress, providing valuable insights into individual differences. Finally, the temporal dynamics of SL improvement (i.e., how quickly and effectively MT can enhance SL abilities) could be explored.

Finally, future studies could also expand on this study by applying EEG to both the listening and musical tasks (here: the PROMS and the CA-BAT). This approach would allow researchers to measure participants' brain activity and their neural entrainment to the rhythm of the music(al beats) during these tasks. For the PROMS, this could, for example, provide insights into how well the brain's rhythmic processing mechanisms align with the presented rhythms and how this synchronisation might differ for familiar versus novel rhythms.

7. Conclusion

This study aimed to establish whether Musical Training (MT) influences the Statistical Learning (SL) of speech segmentation by analysing both offline measures (rating task, PROMS, CA-BAT and Gold-MSI data) and online neural tracking (listening tasks with electroencephalography (EEG)). In line with our expectations, we found evidence for the occurrence of SL in both the rating task and the EEG data. The factors that significantly influenced SL, however, differed between the two measurements. On the rating accuracy scores, years of MT initially had a positive effect, with accuracy improving as the number of years of MT increased. After reorganising the categories of MT, however, this significant effect disappeared, although a consistent upward trend persisted. We anticipate that this trend may become significant with a larger sample size. Conversely, the EEG data indicated that rhythmic abilities, measured with the CA-BAT, had a significant impact on SL, while MT itself did not

show a significant effect. This discrepancy underscores the importance of interpreting SL through various measures: different methods can yield different results. Summarising, musicality seems to be associated with SL, but it remains unclear which specific dimension of musicality (whether it be MT, rhythmic ability, or another aspect) most strongly influences SL.

Additionally, we established a positive relationship between the Gold-MSI's MT subscale and (years of) MT, and both are positively related to the CA-BAT. Performance on the PROMS, however, is not related to the CA-BAT, nor to the Gold-MSI's MT and (years of) MT. This indicates that musical ability is shaped by a broader set of factors beyond just MT, which highlights the need for studies to take a more nuanced approach that incorporates various measures of musical ability.

We recommend future studies to further investigate the precise nature of the relationship between musicality and SL in speech segmentation with a larger sample size.

References

Bates, D., Mächler, 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., & 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

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

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

Brandsen, M. N., van der Wulp, I. M., Struiksma, M. E., & Wijnen, F. N. K. (2023). The influence of musical training on statistical learning in speech segmentation [Preregistration]. *OSF Repository.*

Cantiani, C., Dondena, C., Molteni, M., Riva, V., & Piazza, C. (2022). Synchronizing with the rhythm: Infant neural entrainment to complex musical and speech stimuli. *Frontiers in Psychology, 13*, 944670. https://doi.org/10.3389/fpsyg.2022.944670

Chobert, J., François, C., Velay, J. L., & Besson, M. (2014). Twelve months of active musical training in 8-to-10-year-old children enhances the preattentive processing of syllabic duration and voice onset time. *Cerebral Cortex, 24*(4), 956-967. https://doi.org/10.1093/cercor/bhs377

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. (2023). ordinal: Regression Models for ordinal data. R package version 2023.12-4.1. https://CRAN.R-project.org/package=ordinal.

Cirelli, L. K., Spinelli, C., Nozaradan, S., & Trainor, L. J. (2016). Measuring neural entrainment to beat and meter in infants: Effects of music background. *Frontiers in Neuroscience, 10*, 229. https://doi.org/10.3389/fnins.2016.00229

Correia, A. I., Vincenzi, M., Vanzella, P., Pinheiro, A. P., Schellenberg, E. G., & Lima, C. F. (2023). Individual differences in musical ability among adults with no music training. *Quarterly Journal of Experimental Psychology, 76*(7), 1585-1598. https://doi.org/10.1177/17470218221128557

Deason, R. B., Tat, M. J., Flannery, S., Mithal, P. S., Hussey, E. P., Crehan, E. T., Ally, B. A., & Budson, A. E. (2017). Response bias and response monitoring: Evidence from healthy

older adults and patients with mild Alzheimer's disease. *Brain Cognition, 119*, 17-24. https://doi.org/10.1016/j.bandc.2017.09.002

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods, 134,* 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Di Liberto, G. M., Pelofi, C., Shamma, S., & de Cheveigné, A. (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

Escobar, J., Mussoi, B. S., & Silberer, A. B. (2020). The effect of musical training and working memory in adverse listening situations. *Ear and Hearing, 41*(2), 278-288. https://doi.org/10.1097/AUD.0000000000000754

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* SAGE Publications.

Flaugnacco, E., Lopez, L., Terribili, C., Montico, M., Zoia, S., & Schön, D. (2015). Music training increases phonological awareness and reading skills in developmental dyslexia: A randomized control trial. *PLOS ONE, 10*(9), e0138715. https://doi.org/10.1371/journal.pone.013871

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). SAGE Publications. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

François, C., Jaillet, F., Takerkart, S., & Schön, D. (2014). Faster sound stream segmentation in musicians than in nonmusicians. *PLOS ONE, 9*(7), e101340. https://doi.org/10.1371/journal.pone.0101340

Fritzley, V. H., & Lee, K. (2003). Do young children always say yes to yes-no question? A metadevelopmental study of the affirmation bias. *Child Development, 74*(5), 1297–1313. https://doi.org/10.1111/1467-8624.00608

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

Harrison, P. M. C., & Müllensiefen, D. (2018). 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

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology, 63*(2), 93-106. https://doi.org/10.1016/j.cogpsych.2011.06.002

Hidalgo, C., Falk, S., & Schön, D. (2017). Speak on time! Effects of a musical rhythmic training on children with hearing loss. *Hearing Research, 351*, 11-18. https://doi.org/10.1016/j.heares.2017.05.006

Intartaglia, B., White-Schwoch, T., Kraus, N., & Schön, D. (2017). Music training enhances the automatic neural processing of foreign speech sounds. *Scientific Reports, 7*(1), 12631. https://doi.org/10.1038/s41598-017-12575-1

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology, 29*(1), 1-23. https://doi.org/10.1177/17470218221128557

Kunert, R., Willems, R. M., & Hagoort, P. (2016). An independent psychometric evaluation of the PROMS measure of music perception skills. *PLOS ONE, 11*(7), e0159103. https://doi.org/10.1371/journal.pone.0159103

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

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

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8*, 1–14. https://doi.org/10.3389/fnhum.2014.00213

Mandikal Vasuki, P. R., Sharma, M., Ibrahim, R. K., & Arciuli, J. (2017). Musicians' online performance during auditory and visual statistical learning tasks. *Frontiers in Human Neuroscience, 11*, 114. https://doi.org/10.3389/fnhum.2017.00114

Misyak, J., 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*, 1–10. 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

Neves, L., Correia, A. I., Castro, S. L., Martins, D., & Lima, C. F. (2022). Does music training enhance auditory and linguistics processing? A systematic review of meta-analysis of behavioral and brain evidence. *Neuroscience & Biobehavioral Reviews, 140*, 104777. https://doi.org/10.1016/j.neubiorev.2022.104777

Nozaradan, S. (2014). Exploring how musical rhythm entrains brain activity with electroencephalogram frequency-tagging. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1658), 20130393. https://doi.org/10.1098/rstb.2013.0393

Ong, J. H., & Chan, A. H. D. (2019). Statistical learning among young and older adults: Similar yet different? *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019)*, 2109-2113.

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

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, 3*(2), 214-234. https://doi.org/10.1162/nol_a_00061

R Core Team. (2017). R: A language and environment for statistical computing. [Computer software]. http://www.r-project.org/

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-a). 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-b). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606-621. https://doi.org/10.1006/jmla.1996.0032

Schevenels, K., Altvater-Mackensen, N., Zink, I., De Smedt, B., & Vandermosten, M. (2023). Aging effects and feasibility of statistical learning tasks across modalities. *A Journal on Normal and Dysfunctional Development, 30*(2), 201-230. https://doi.org/10.1080/13825585.2021.2007213

Schön, D., & François, C. (2011). Musical expertise and statistical learning of musical and linguistic structures. *Frontiers in Psychology, 2,* 167. https://doi.org/10.3389/fpsyg.2011.00167

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics, 8*(2), 39-42. https://doi.org/10.12691/ajams-8-2-1.

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

Singh, L., Steven Reznick, J., & 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

Srinivasan, R., & Petrovic, S. (2006). MEG phase follows conscious perception during binocular rivalry induced by visual stream segregation. *Cerebral Cortex, 16,* 597–608. https://doi.org/10.1093/cercor/bhj016

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

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

van der Wulp, I. M., Struiksma, M. E., Batterink, L. J., & Wijnen, F. N. K. (2023). Investigating individual differences in linguistic statistical learning and their relation to rhythmic and cognitive abilities: A speech segmentation experiment with online neural tracking. In principle acceptance of Version 4 by Peer Community in Registered Reports. https://doi.org/10.17605/OSF.IO/2Y6SX

Wickham, H,, Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., 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. (2019). *Statistics for linguists: An introduction using R*. Routledge.

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, J. D., Susino, M., McPherson, G. E., & Schubert, E. (2018). The definition of a musician in music psychology: A literature review and the six-year rule. *Psychology of Music, 48*(3), 389-409. https://doi.org/10.1177/0305735618804038

Appendix A: Participant Overview

Table A1

Overview of Participant Information per Participant

Participant	Gender	Age (in years)	Education Level	Years of MT
VUGY	F	22	University	3-5
TBGN	F	24	University	10+
WYEL	F	24	University	3-5
LBVC	F	22	University	6-9
QENU	F	21	University	10+
HSFE	M	23	University	3-5
WLMQ	F	23	University	0
LPRY	F	20	University	2
EHSB	F	32	University	3-5
STXW	F	22	University	6-9
EGYK	F	24	University	6-9
SSYV	F	23	University	3-5
QYDS	F	19	University	1
LCLF	F	19	University	0
XYJL	F	21	University	0
HAAA	M	23	University	0
FQJT	F	19	University	1
RBTP	M	22	University	10+
YMGD	F	20	University	3-5
QNWP	F	20	University	0
ASBB	F	20	University	10+
CNED	F	22	University	0
YTKJ	F	32	University of Applied Sciences	3-5
LSUA	M	20	University	10+

QMLC	F	21	University	6-9
GUJN	F	23	University	3-5
NDLG	F	20	University	1
XDSD	F	19	University	0.5
EZGE	F	54	University of Applied Sciences	2

Appendix B: Rating Task Items

Table B1

Items for the Rating Task per Category

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

Note. Reprinted from van der Wulp et al. (2023, Table C2, p. 56).

Appendix C: Q-Q Plots

Figure C1

Q-Q Plot of the Familiarity Scores (1-4) from the Rating Task

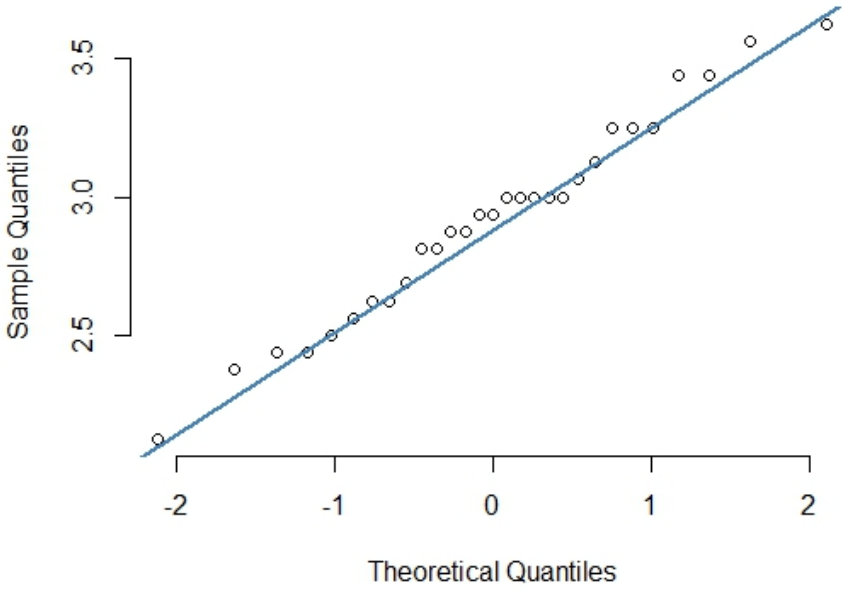
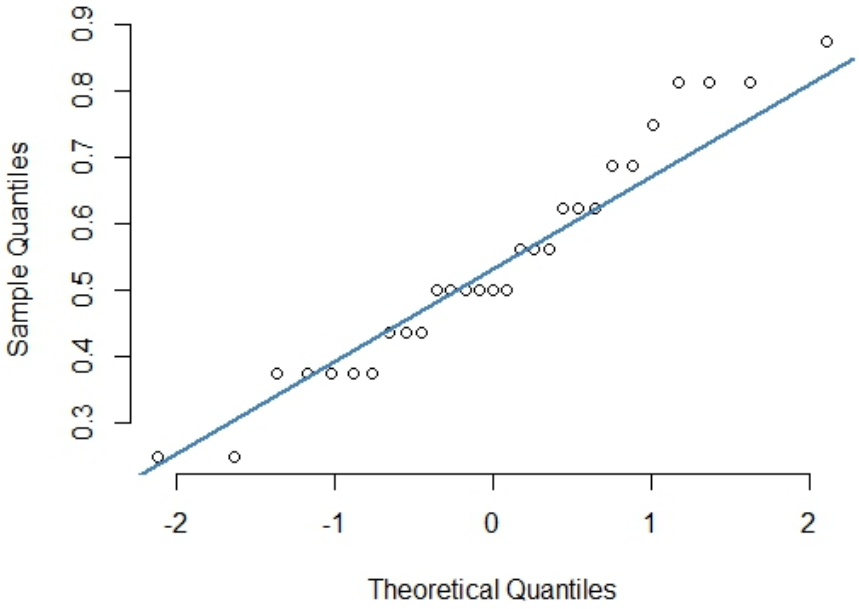


Figure C2

Q-Q Plot of the Accuracy Scores (0-1) from the Rating Task



Appendix D: Correlations Musicality Measures

Figure D1

Scatter Plot of the Correlation Between Gold-MSI and Years of Musical Training (MT)

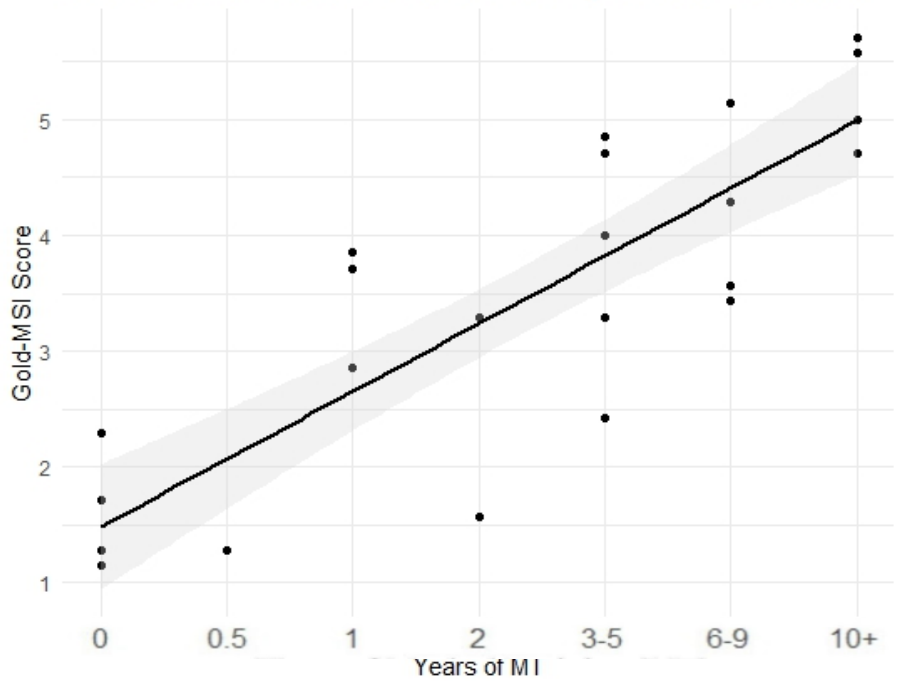


Figure D2

Scatter Plot of the Correlation Between Gold-MSI and CA-BAT

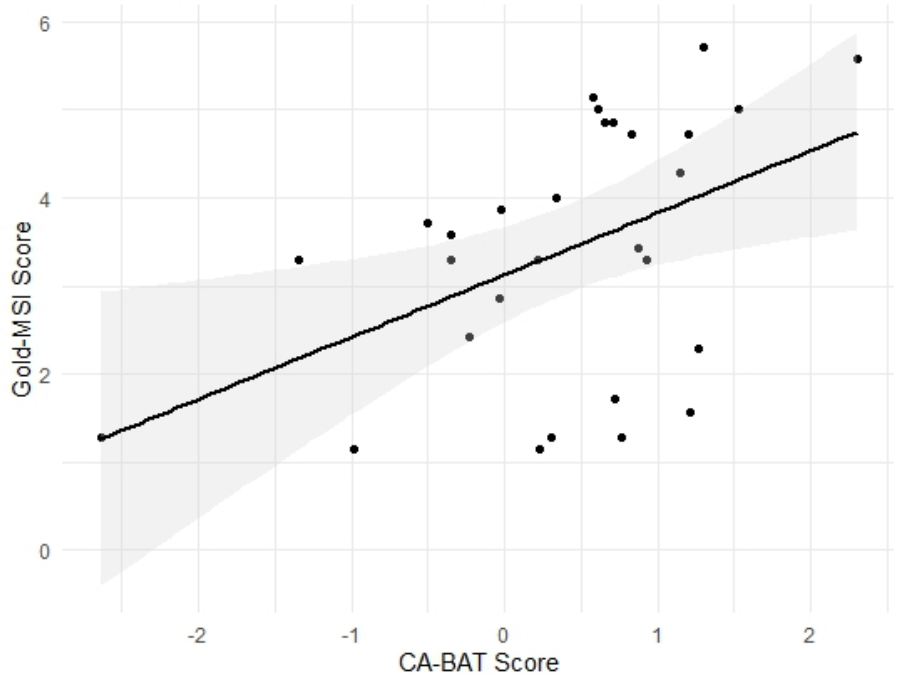


Figure D3

Scatter Plot of the Correlation Between Gold-MSI and PROMS

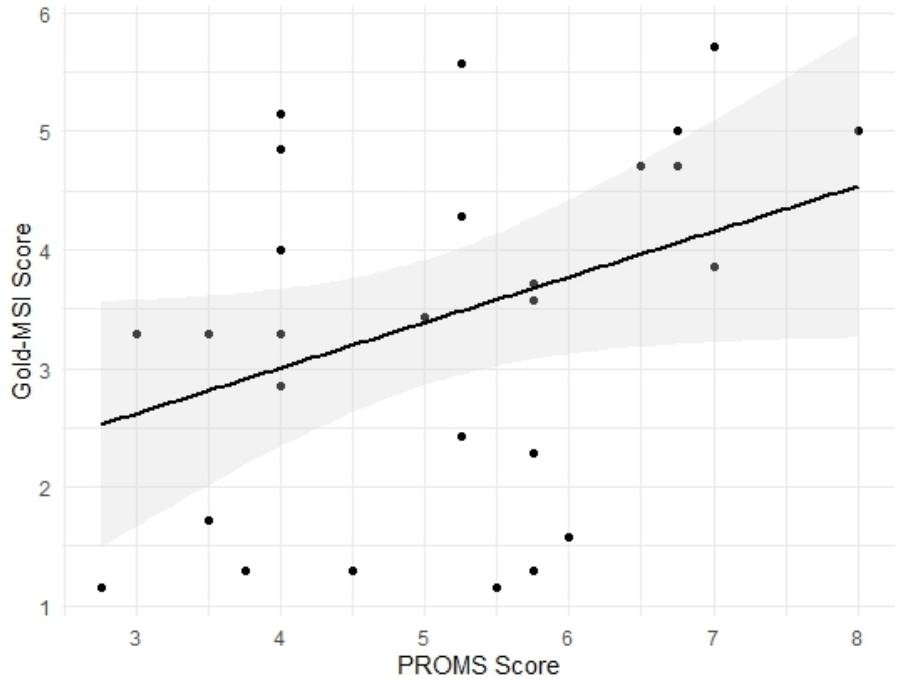


Figure D4

Scatter Plot of the Correlation Between Years of Musical Training (MT) and CA-BAT

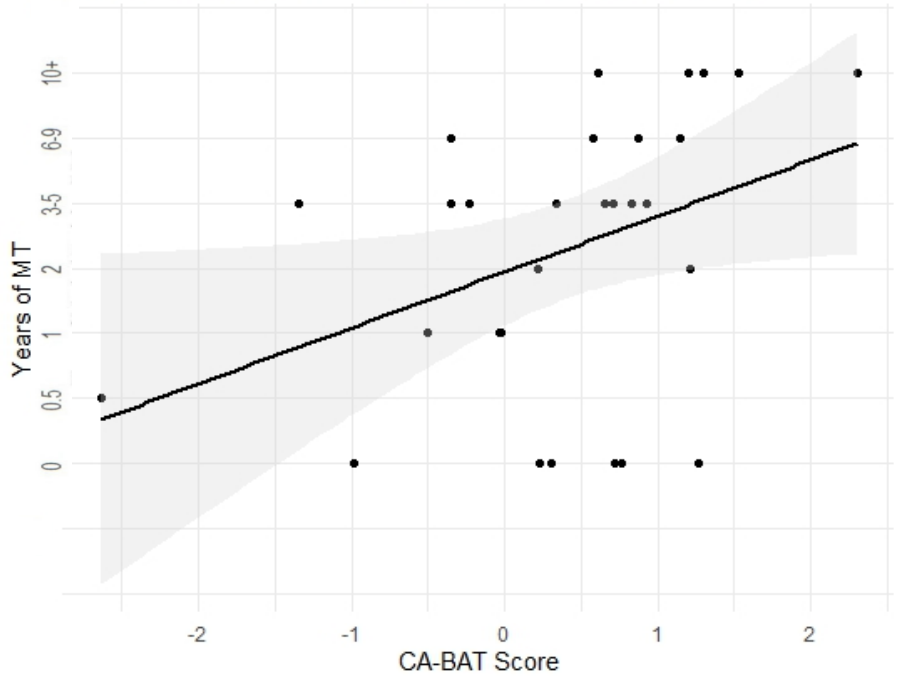


Figure D5

Scatter Plot of the Correlation Between Years of Musical Training (MT) and PROMS

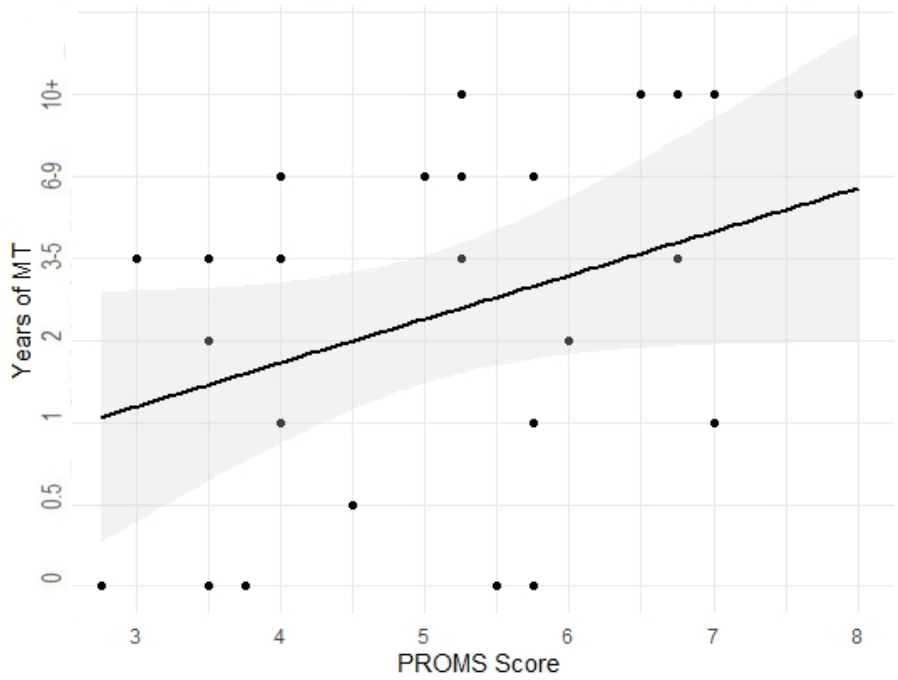
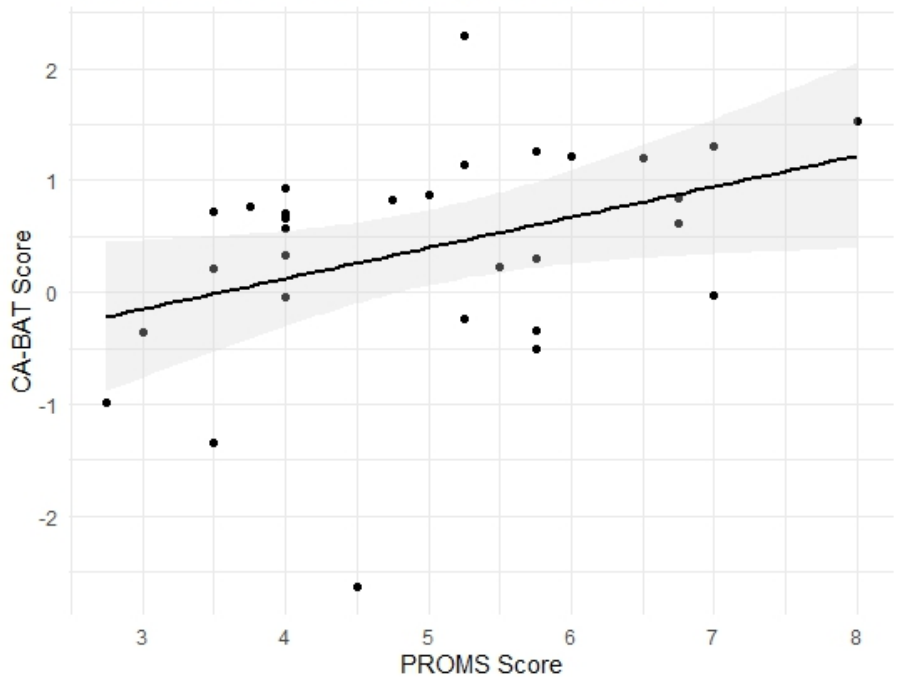


Figure D6

Scatter Plot of the Correlation Between CA-BAT and PROMS



Appendix E: Likelihood Ratio Test Results and Model Coefficients

Table E1

Overview of Likelihood Ratio Test Results of Various Models Across Different Data Sets.

Data	Dependent variable	Model name	Parameters	*logLik*	*df*	χ^2/*LR.stat*	*p*	Added predictor compared to model 0	Model of comparison	Action
Rating task data	Familiarity rating	Model 0	5	-566.46						Kept
		Model 1	7	-560.27	2	12.37	.002*	wordtype	Model 0	Kept
		Model 2	13	-555.20	6	10.16	.12	wordtype + Years_MT	Model 1	Discarded
		Model 2_cat	8	-559.98	1	0.58	.45	wordtype + MT_cat	Model 1	Discarded
		Model 3	8	-559.39	1	1.77	.18	wordtype + CA.BAT	Model 1	Discarded
		Model 4	8	-560.24	1	0.06	.80	wordtype + PROMS	Model 1	Discarded
	Accuracy score	Model 0	3	-284.21						Kept
		Model 1	5	-272.62	2	23.18	< .001*	wordtype	Model 0	Kept
		Model 2	11	-266.00	6	13.25	.04*	wordtype + Years_MT	Model 1	Kept
		Model 2_cat	8	-269.62	3	6.00	.11	wordtype + MT_cat	Model 1	Discarded
		Model 3	12	-265.38	1	1.25	.26	wordtype + Years_MT + CA.BAT	Model 2	Discarded
		Model 4	12	-265.99	1	0.03	.87	wordtype + Years_MT + PROMS	Model 2	Discarded

	Accuracy percentage	Model 0	4	5399.30						
		Model 1	10	5301.90	6	0	1	Years_MT	Model 0	Discarded
		Model 2	5	5030.30	1	0	1	CA.BAT	Model 0	Discarded
		Model 3	5	4895.00	1	0	1	PROMS	Model 0	Discarded
EEG data	WLI	Model 0	3	26.90						Kept
		Model 1	4	30.25	1	6.70	.01*	condition	Model 0	Kept
		Model 2	5	30.27	1	0.03	.87	condition + Years_MT	Model 1	Discarded
		Model 2_cat	5	30.33	1	0.16	.69	condition + MT_cat	Model 1	Discarded
		Model 3	5	32.36	1	4.21	.04*	condition + CA.BAT	Model 1	Kept
		Model 4	6	33.93	1	3.15	.08	condition + CA.BAT + PROMS	Model 3	Discarded
	Syllable ITC	Model 0	3	34.46						Kept
		Model 1	4	35.06	1	1.19	.27	Years_MT	Model 0	Discarded
		Model 1_cat	4	1145.2	1	1.14	.29	cat_MT	Model 0	Discarded
		Model 2	4	1144.7	1	0.10	.76	CA.BAT	Model 0	Discarded
		Model 3	4	1146.8	1	4.35	.04*	PROMS	Model 0	Kept

Note. * $p < .05$.

Table E2

Tukey HSD Corrected Emmeans and Contrast for Word Type Comparisons Based on the Familiarity Rating in the Rating Task

Contrast	*β*	*SE*	*df*	*z-ratio*	*p-value*
foil - part-word	-0.20	0.24	Inf	-0.83	.68
foil - word	-0.95	0.23	Inf	-4.08	< .001***
part-word - word	-0.76	0.19	Inf	-4.03	< .001***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table E3

Tukey HSD Corrected Emmeans and Contrast for Word Type Comparisons Based on the Accuracy Score in the Rating Task

Contrast	*β*	*SE*	*df*	*z-ratio*	*p-value*
foil - part-word	-0.20	0.35	Inf	0.57	.84
foil - word	-2.41	0.47	Inf	-5.12	< .001***
part-word - word	-2.61	0.43	Inf	-6.12	< .001***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix F: Additional Figures for the Rating Task Data

Figure F1

Mean Familiarity Ratings (1-4) per Participant

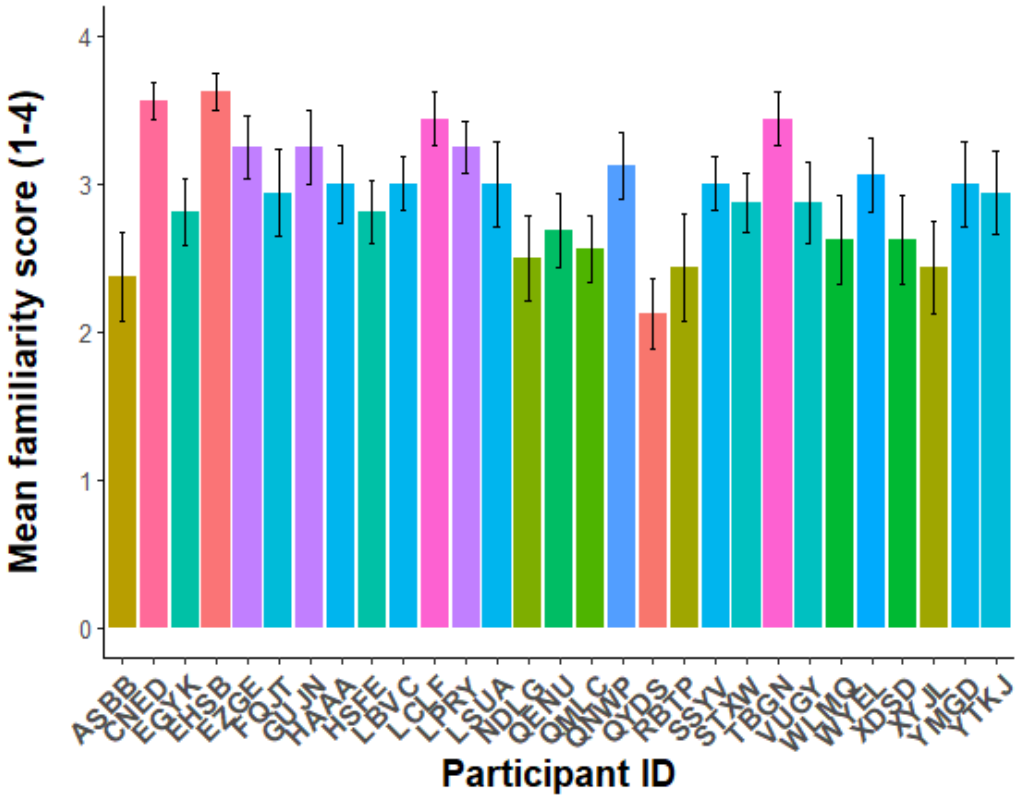
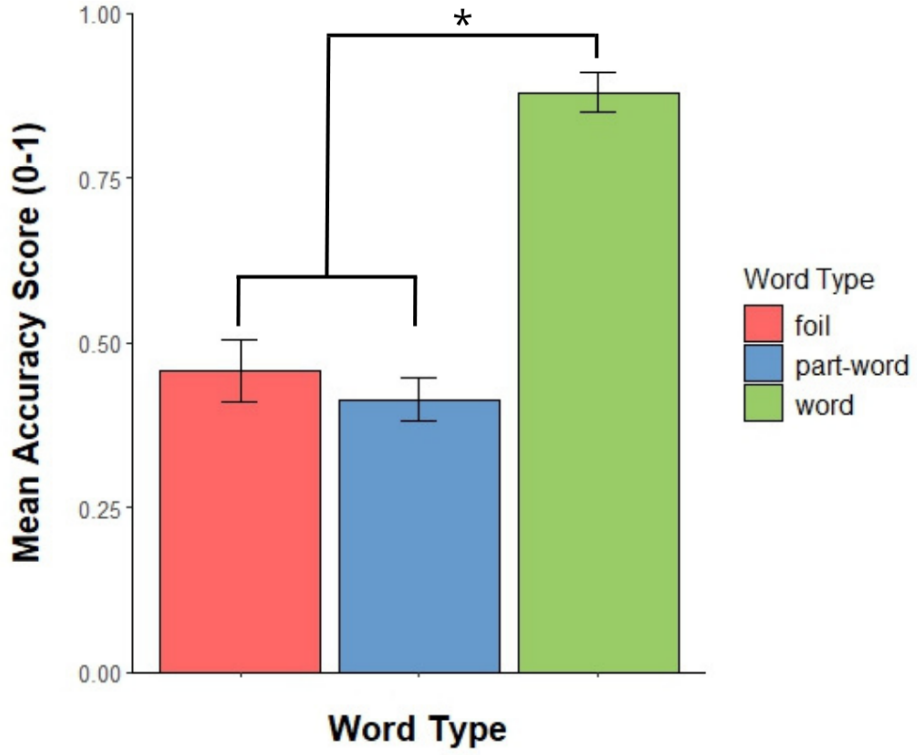


Figure F2

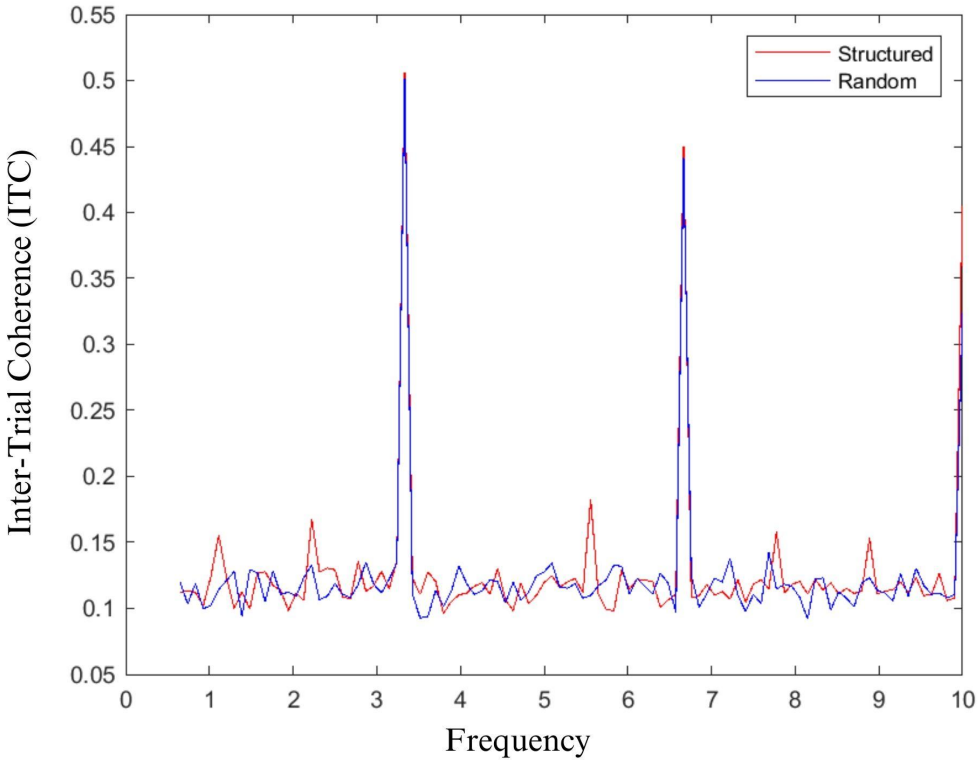
Mean Accuracy Score (0-1) per Word Type



Appendix G: Additional Figures and Plots for the EEG Data

Figure G1

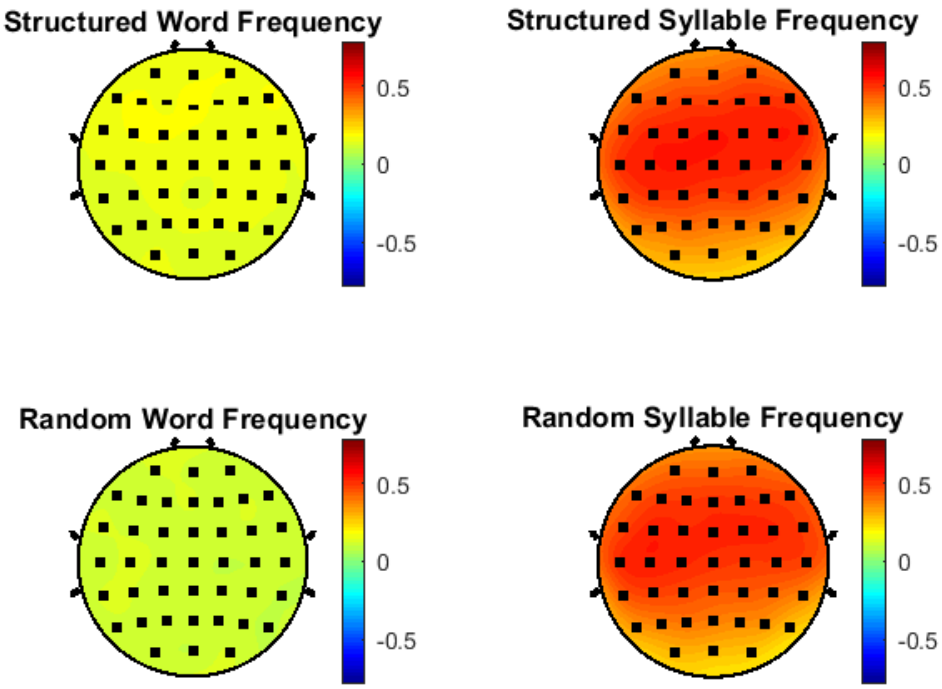
Frequency Peaks for the Structured (Red) and Random Condition (Blue) (0-10 Hz).



Note. In both conditions, there is a peak at 3.3 Hz (word frequency), and also at 6.6 Hz (harmonic frequency of 3.3 Hz). In the structured condition, there is also a peak at 1.1 Hz (word frequency).

Figure G2A

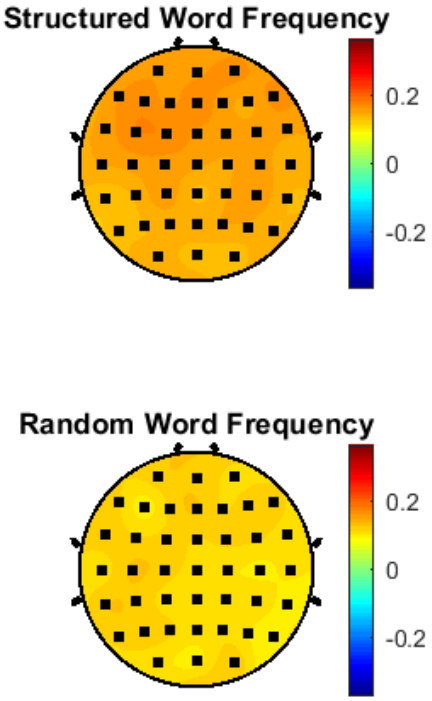
Topographical Plots Visualising ITC Value Distribution Across the Scalp, as a Function of Condition and Frequency (Word, Syllable).



Note. Different scales are used for word versus syllable frequencies.

Figure G2B

Topographical Plots Visualising ITC Value Distribution Across the Scalp for the Word Frequency, as a Function of Condition.



Note. The ITC of word frequencies is generally lower than that of syllable frequencies. To highlight the (significant) difference in ITC value between the structured and random condition, this plot uses a lower scale than the plot in Figure G2A.

Figure G3

Scatter Plot of the Correlation Between WLI and Accuracy Scores