Title:

Auditory, but not audiovisual cues lead to higher neural sensitivity to the statistical regularities of an unfamiliar musical style

Authors:

Ioanna Zioga[a], Peter M. C. Harrison[b], Marcus T. Pearce[b,c], Joydeep Bhattacharya[d], Caroline Di Bernardi Luft[a]

[a] School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom

[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, United Kingdom

[c] Centre for Music in the Brain, Aarhus University, Aarhus, Denmark

[d] Department of Psychology, Goldsmiths, University of London, London SE14 6NW, United Kingdom

Corresponding authors:

Ioanna Zioga, i.zioga@qmul.ac.uk

Caroline Di Bernardi Luft, c.luft@qmul.ac.uk

Abstract

It is still a matter of debate whether visual aids improve learning of music. In a multi-session study, we investigated the neural signatures of novel music sequence learning with or without aids (auditory-only: AO, audio-visual: AV). During three training sessions on three separate days, participants (non-musicians) reproduced (note by note on a keyboard) melodic sequences generated by an artificial musical grammar. The AV group ($N = 20$) had each note colour-coded on screen, whereas the AO group ($N = 20$) had no colour indication. We evaluated learning of the statistical regularities of the novel music grammar before and after training by presenting melodies ending on correct or incorrect notes, and by asking participants to judge the correctness and surprisal of the final note, while EEG was recorded. We found that participants successfully learned the new grammar. While the AV group, as compared to AO group, reproduced longer sequences during training, there was no significant difference in learning between groups. At the neural level, after training, the AO group showed larger N100 response to low-probability compared to high-probability notes, suggesting an increased neural sensitivity to statistical properties of the grammar; this effect was not observed in the AV group. Our findings indicate that visual aids might improve sequence reproduction whilst not necessarily promoting better learning, indicating a potential dissociation between sequence reproduction and learning. We suggest that the difficulty induced by auditory-only input during music training might enhance cognitive engagement, thereby improving neural sensitivity to the underlying statistical properties of the learned material.

Keywords: Visual aids; Artificial music grammar; EEG; Statistical learning; Training.

1. Introduction

Music forms a vital part of the school curriculum in much of the Western world. During the first years of music education, teaching music usually takes the form of a game (Aronoff, 1983; Bowles, 1998): different colours represent different pitches, imaginary stairs symbolize musical scales, and claps represent rhythms. A widely used method is to put colourful stickers on the keys of a piano keyboard (Simpson, 2015) or on the violin fingerboard (Abler, 2002) to indicate finger positions. Guitar Hero (https://www.guitarhero.com/uk/en/), a music computer game, makes people feel empowered by being able to reproduce popular songs on a guitar toy using visual cues; but do they really learn music? There is no research (to the best of our knowledge) testing whether such methods improve learning. Over a multi-session, conducted on separated days, musical training experiment, we examined whether visual aids would lead to better learning of an unfamiliar music grammar, and investigated the respective electrophysiological correlates of statistical music learning.

Musical learning depends not only on developing abilities for singing or playing a musical instrument, but also on learning a musical grammar, i.e. the statistical properties of a particular musical style. Musical experts have typically internalized the rules or probabilistic regularities that govern a specific music style and can form expectations for subsequent events while listening (Jonaitis & Saffran, 2009; Meyer, 1956). The fulfilment or violation of these expectations plays a crucial role in the emotional experience of music (Huron, 2006; Juslin & Västfjäll, 2008). Importantly, the formation of expectations can be used as an index of learning: the greater the knowledge of a learned musical style, the larger the degree of unexpectedness when a rule is violated (Steinbeis, Koelsch, & Sloboda, 2006).

Humans can acquire knowledge of the statistical regularities of auditory structures even after short exposure (Lieberman, Chang, Chiao, Bookheimer, & Knowlton, 2004; Loui, 2012; Misyak, Christiansen, & Tomblin, 2010; Pothos, 2007; Reber, 1993; Rohrmeier & Cross, 2014; Rohrmeier & Rebuschat, 2012; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Saffran, Newport, & Aslin, 1996). Statistical learning and recognition of grammatical patterns through passive exposure has been demonstrated in tone (Saffran, Reeck, Niebuhr, & Wilson, 2005; Saffran et al., 1999) and timbre (Tillmann & McAdams, 2004) sequences, as well as in unfamiliar musical systems (e.g., use of the Bohlen-Pierce scale: Loui & Wessel, 2008; Loui, Wessel, & Kam, 2010). Participants can also perform accurate predictions on other types of stimuli based on their temporal statistics,

such as on sequences of visual stimuli (e.g., abstract visual shapes: Fiser, & Aslin, 2002; Gabor patches: Luft, Baker, Goldstone, Zhang, & Kourtzi, 2016; Luft, Meeson, Welchman, & Kourtzi, 2015; tones in oddball task: Debener, Makeig, Delorme, & Engel, 2005; semantics: Proverbio, Leoni, & Zani, 2004). Studies with infants demonstrate statistical learning of both auditory and visual information, providing evidence for an underlying domain-general mechanism (Kirkham, Slemmer, & Johnson, 2002; Saffran et al., 1996). However, Conway and Christiansen (2006) found that adult participants can simultaneously learn the statistical regularities of two different artificial grammars, one presented with auditory and one with visual stimuli, therefore suggesting modality-specific statistical learning.

The electrophysiological recording, especially the event related potential (ERP) response, has routinely been used to study the neural correlates of learning due to its excellent temporal resolution (Rugg & Coles, 1995), and has been associated with sensory and perceptual processing modulated by expectation and familiarity (Näätänen, Gaillard, & Mäntysalo, 1978; Tremblay & Kraus, 2002). Violation, as compared to fulfilment, of pitch expectations is robustly associated with a larger N100 component, a fronto-central negativity around 100 ms after the onset of a melodically unexpected note (Koelsch & Jentschke, 2010; Pearce, Ruiz, Kapasi, Wiggins, & Bhattacharya, 2010). The N100 has been also used as an index of statistical learning of auditory sequences, with studies showing increased N100 in response to tones with lower transitional probability compared to tones with higher probability (Abla, Katahira, & Okanoya, 2008; Moldwin, Schwartz, & Sussman, 2017; Paraskevopoulos, Kuchenbuch, Herholz, & Pantev, 2012). Koelsch and colleagues (2016) found that the amplitude of this response was negatively related to the probability of an auditory event. It has been suggested that this early component reflects the magnitude of prediction errors in statistical learning contexts (Tsogli, Jentschke, Daikoku, & Koelsch, 2019).

The P200, a positive ERP component peaking around 200 ms after the onset of an event, has been linked to stimulus familiarity. For example, familiar speech variants of syllables (Tremblay & Kraus, 2002) and familiar words (Perfetti, & Wang, 2006; Stuellein, Radach, Jacobs, & Hofmann, 2016) have been associated with a larger P200 than unfamiliar syllables and words. Further, musicians demonstrate larger P200 in auditory tasks compared to non-musicians, which is usually attributed to their long-term musical training, inducing greater familiarity with the stimuli (Atienza, Cantero, & Dominguez-Marin, 2002; Tremblay, Kraus, McGee, Ponton, & Otis, 2001).

The aforementioned studies on statistical learning mostly focused on learning by training a single perceptual modality (e.g., auditory/visual). Previous studies have demonstrated the beneficial effects of multi-modality on learning (e.g., Brünken, Plass, & Leutner, 2004; Cleary, Pisoni, & Geers, 2001; Tierney, Bergeson-Dana, & Pisoni, 2008). There is behavioural and neurophysiological evidence demonstrating that adults are faster at detecting a target when correlated information is presented to multiple sensory modalities than when information is presented unimodally (e.g., Colonius & Diederich, 2006; Molholm, Ritter, Javitt, & Foxe, 2004; Sinnett, Soto-Faraco, & Spence, 2008). The "Simon" task has been widely used to study statistical learning: it uses a game device with four coloured buttons corresponding to different tones. Every time a tone is played, the respective button lights up. Tierney and colleagues (2008) asked participants to reproduce random sequences of coloured lights by pressing the keys on the Simon device. Results showed that longer sequences were reproduced in the audio-visual condition (colour names spoken and buttons lighting up simultaneously) compared to the auditory-only or visual-only condition. Beneficial effects of audio-visual presentation on learning have also been found in other tasks, such as presentation of biological textbook material with vs. without verbal instruction, in addition to pictorial presentation (Brünken et al., 2004).

Our study is the first (to our knowledge) to investigate the effect of visual aids on statistical learning of music with interleaved passive exposure to and active reproduction of music. Alternating different methods is efficient for learning and generalization of knowledge (Richland, Bjork, Finley, & Linn, 2005), as well as more ecologically valid compared to mere passive exposure to the learned material. In contrary to previous studies which assessed learning just after exposure, we performed a 1-day follow up test, in order to ensure we measure learning rather than immediate effects of exposure. We introduced a novel experimental paradigm combining behavioural, electrophysiological, and computational methods. Specifically, non-musicians were trained on an unfamiliar artificial music grammar (taken from Rohrmeier, Rebuschat, & Cross, 2011) through passive exposure and active reproduction of melodic sequences on a sound keyboard with or without visual aids, over three separate days. An artificial music grammar was ideal for our investigation because it represented a completely novel musical style for all participants. Participants' knowledge of the novel grammar was assessed before and after training by taking judgements of the perceived correctness and surprisal of high-probability, low-probability, incorrect, and random notes. The ERPs in response to these notes were also analysed.

We used a computational model of auditory expectation (Information Dynamics of Music, IDyOM: Pearce, 2018) to quantify the conditional probability of each note in every sequence, reflecting the degree of expectedness of a particular note given the preceding musical context. IDyOM uses variable-order Markov models (Begleiter, El-Yaniv, & Yona, 2004) to generate the conditional probability of a note given its preceding context based on the frequency with which each note has followed the context in a given corpus of music. IDyOM embodies the hypothesis that listeners base their expectations on learning the statistical regularities in the musical environment, with listeners perceiving high-probability notes as expected and low-probability notes as unexpected. Previous behavioural, physiological, and EEG studies have demonstrated that IDyOM successfully predicts listeners' expectations (Egermann, Pearce, Wiggins, & McAdams, 2013; Hansen & Pearce, 2014; Omigie et al., 2013; Pearce, Müllensiefen, D., & Wiggins, 2010; Pearce et al., 2010). The probability of each event according to the model can be log-transformed to yield its *information content* (IC), which reflects how unpredictable the model finds a note in a particular context. We used IDyOM to analyse each melodic sequence generated by the artificial music grammar, and manipulated these sequences to construct melodies terminating on high- and low-probability, incorrect, and random notes. Participants' learning was evaluated in terms of their accuracy in recognizing notes belonging to the grammar.

Previous studies have demonstrated a distinction between performance during training and learning (e.g., Kantak & Winstein, 2012; Lee & Genovese, 1988; Schmidt & Bjork, 1992). In a review, Soderstrom and Bjork (2015) argued in favour of differentiating learning from performance during training, as the former refers to a long-term change in behaviour or knowledge that supports retention and transfer, and the latter to temporary fluctuations in behaviour or knowledge which are observed close to the acquisition period. In our study, "performance during training" (according to Soderstrom and Bjork, 2015) corresponds to sequence reproduction in the training sessions, whereas "learning" refers to acquisition of the statistical regularities of the AMG. Based on the aforementioned studies, our hypothesis is twofold. First, we hypothesized that multimodality would aid sequence reproduction since the visual cues would signal to the participants which exact keys to press. Second, we expected that the presence of visual aids would have a negative impact on learning as the auditory input is modality-appropriate for learning music, and visual information in this context might work as a distractor for better encoding. At the neural level, we predicted that the N100 component would be higher in response to low-probability and incorrect notes (compared to high probability notes) after training. Since larger N100 in response to low-probability notes

indicates better learning of the statistical regularities of the grammar, we hypothesised this would be higher for the auditory-only group. Further, we expected that the P200 component, as an index of familiarity (Tremblay & Kraus, 2002), would be enhanced after training in both groups. Finally, we explored how the early right anterior negativity (ERAN), an ERP component previously associated with syntactical violations in music (Koelsch, Gunter, Friederici, & Schröger, 2000; Pearce & Rohrmeier, 2018), would be modulated in our statistical learning paradigm.

2. Methods

2.1. Participants

Forty neurologically healthy human adults (24 female) aged between 20 and 32 years old (mean ± s.d. age of 22.42 ± 3.04 years) participated in the experiment. Participants were randomly assigned to one of two groups which differed in the training method: audio-visual group, AV (N = 20, 12 female, age range 20 – 32 years, mean ± s.d. 22.25 ± 3.37 years), and auditory-only group, AO (N = 20, 12 female, age range 20 – 30 years, mean ± s.d. 22.60 ± 3.52 years). All participants self-reported that they were non-musicians, and this was validated by the 'Goldsmiths Musical Sophistication Index' (Gold-MSI) questionnaire (Müllensiefen, Gingras, Musil, & Stewart, 2014): mean Gold-MSI Musical Training scores ± s.d. were 12.08 ± 3.63 for the AV group and 12.10 ± 5.51 for the AO group from a possible range of 7-49 points (higher values indicating more musical training). The scores were not significantly different between groups ($t(38) = .017$, $p = .987$). Two participants were excluded because they did not sufficiently engage with the task (gave the same response throughout the pre- and post-test), leaving 19 participants per group. All participants reported normal hearing and normal or corrected-to-normal vision. Participants gave written informed consent and received financial compensation at a rate of £7 per hour for their participation. The study was approved by the Ethics Board at Queen Mary University of London.

2.2. Materials

Gold-MSI Musical training questionnaire: The Musical Training factor (Dimension 3) of the Gold-MSI comprises a self-report measure including seven statements regarding formal musical training experience and musical skill. Each statement (e.g., *'I have never been complimented for my talents as a musical performer'*) requires a response from 1 (*Completely*

Disagree) to 7 (*Completely Agree*). This measure was used to validate that all participants were non-musicians.

Artificial music grammar (AMG): Our melodic stimuli were sequences generated by an artificial music grammar (AMG) taken from Rohrmeier and colleagues (2011) (Figure 1A). This grammar consists of 8 different tone pairs, and the tones belong to the Western diatonic major scale (C4, D4, E4, F4, G4, A4, B4). The AMG generated 18 different melodic sequences, ranging from 8 to 22 notes long (mean length ± s.d. = 14.56 ± 3.87). Melodic sequences with circular paths were excluded, as they were too long to be used in our paradigm. Twelve of these sequences were used for the training and test sessions ('old-grammatical'), while the remaining six were only presented in the last session to test generalization to unheard melodies of the grammar ('new-grammatical'). Please refer to the Supplementary materials for the 18 melodic sequences in musical notation (Figure S1).

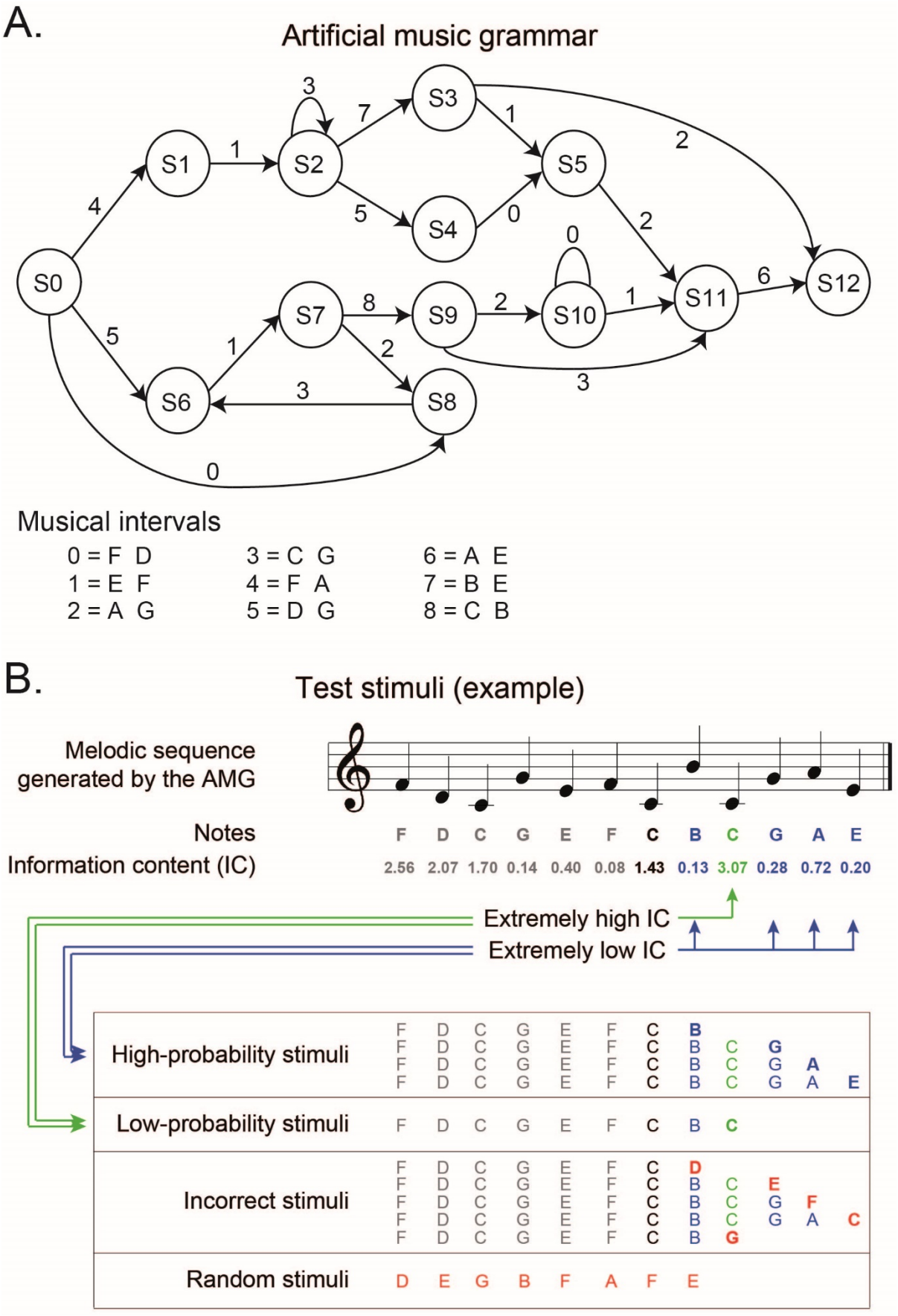


Figure 1 - A. Top: Schematic illustration of the artificial music grammar (AMG) by Rohrmeier, Rebuschat, and Cross (2011). Musical intervals are numbered from 0 to 8. Symbols starting with 'S'

constitute the grammar nodes. Bottom: Each terminal corresponds to a pair of musical notes. The musical notes range from C4 to B4; **B.** An example of how the stimuli for the test sessions were generated. For each melodic sequence generated by the AMG, notes with extremely high (green) and extremely low (blue) information content (IC) were identified. Notes only after the 6th note were identified, in order for the stimuli to have a considerable length. High-probability (HP) notes corresponded to low IC, while low-probability (LP) notes corresponded to high IC. HP stimuli (LP stimuli) were constructed by interrupting the melodic sequences of the AMG on the identified notes with low IC (high IC). Incorrect stimuli were generated by replacing the last note of HP and LP stimuli with an incorrect note (i.e. a note that never existed in the AMG at that particular place). Random stimuli were constructed by generating random note sequences. This procedure was done separately for the pre- and post-tests using the 12 old-grammatical sequences, as well as for the generalisation session using the 6 new-grammatical sequences.

IDyOM analyses of melodic sequences: An information theoretic model of music expectation, IDyOM (Pearce, 2018; 2005) was used to analyse the statistical properties of the melodic sequences generated by the AMG. We conducted leave-one-out cross-validations, while IDyOM generated predictions for each sequence after pretraining on the other 17 sequences. IDyOM uses *viewpoints* to generate predictions. We evaluated different sets of viewpoints and selected the viewpoint chromatic pitch and chromatic interval (cross-entropy = 0.986), which outperformed the single viewpoint chromatic pitch (cross-entropy = 1.007), and the viewpoint set chromatic pitch, chromatic interval, and contour (cross-entropy = 1.043).

Further, IDyOM was used to make predictions combining a long-term model, which was pretrained on the 17 other melodies, and incrementally on the current melody, as well as a short-term model, that was only trained incrementally on the current melody. This combination of long- and short-term models has been found to reflect listeners' expectations well (Pearce, 2005). IDyOM estimates the probability for each note in each of the 18 AMG melodies. We calculated *information content* (IC) by taking the negative logarithm (base 2) of this probability estimate. Low IC corresponds to high-probability (i.e. predictable) notes, while high IC corresponds to low-probability (i.e. unpredictable) notes based on a given grammar.

Melodic stimuli for the judgement sessions: The melodies were interrupted after a target note, and participants were prompted to judge if the last note was correct or incorrect, and surprising or not surprising. For the pre-test and post-test, we used 280 melodies, terminating with "target" notes of different levels of note probability: 70 high-probability (HP), 70 low-probability (LP), 70 incorrect (INC), and 70 random (Figure 1B). For the generalisation session we used 105 melodies: 35 HP, 35 LP, and 35 INC. The melodies for the test sessions

were generated from the 12 old-grammatical sequences, whereas for the generalisation session they were generated from the 6 new-grammatical sequences.

To generate the melodies ending on HP and LP notes, we first identified those with the lowest 30% information content (IC) (extreme HP) and those with the highest 30% (extreme LP) out of all the notes of the 18 AMG sequences. The probability values of the identified HP notes ranged from 0.83-0.94 (M = 0.90, SD = 0.03), while the probability of the LP notes ranged from 0.01-0.37 (M = 0.21, SD = 0.10). There were 79 notes with extreme probabilities: 55 belonged to the old-grammatical sequences, and 24 to the new-grammatical sequences. Of the 55 ones, 36 notes were HP and 19 LP. To reach 70 trials per condition, 34 (randomly picked) of the 36 HP melodies were repeated once, while all 19 LP were repeated three times (giving 57 melodies), and 13 (randomly selected from the middle 40% of the distribution) were added (total of 70). The same was applied for the new-grammatical sequences. The 16 HP melodies were repeated once (32) and 3 more (randomly picked) were added (35 in total). The 8 LP melodies were repeated four times (32) and 3 more (randomly picked) were added (35).

The incorrect melodies (INC) were generated by replacing the last note of the HP and LP melodies with a note that never appeared in that context in the AMG. Three different sets of INC melodies were created, one for the pre (70), one for the post test (70), and one for the generalisation (35). We also generated two different sets of 70 random melodies, presented in the pre- and post-test. The random melodies had similar length to the rest of the melodies, by producing 5 random melodies for each of the possible lengths (7 to 20 notes). The melodies were played through speakers located to the left and right of participants. Notes had a duration of 330 ms, with the next note beginning immediately after the end of the previous note, and were played with a piano timbre. All notes had a 100-ms fade-out time. Psychtoolbox (Brainard & Vision, 1997) was used for stimuli presentation. Examples of the stimuli are now included as audio files in Supplementary materials.

2.3. Procedure

Participants came to the lab on four separate days with a maximum two-day gap between any of the days (Figure 2A). Participants received training on the melodies generated by the AMG, through active reproduction on a keyboard with (AV group) or without visual cues (AO group) across three sessions (days 1-3). Learning of the AMG was assessed before and after training (days 1 and 4). Participants were presented with melodies and were

prompted to judge if the final note was correct or incorrect and surprising or not surprising, while their EEG was recorded. In the last generalisation session, participants were asked to judge if the final note of previously unheard sequences was surprising or not surprising. As the primary aim of our study was to the efficacy of visual aids in music (but not visual) learning, participants from both groups were tested only in the auditory domain (auditory without visual stimuli) in the pre-test, post-test, and generalization sessions. On days 2 and 3, after a brief (5 min) passive exposure to all the old-grammatical sequences three times (36 in total), participants were then asked to complete a short surprisal (yes or no) judgement task of melodies ending with high-probability or low-probability notes (intermediate surprisal sessions). After each training session, participants were asked to compose and perform a musical composition based on the learned materials, but this part is outside the scope of this paper.

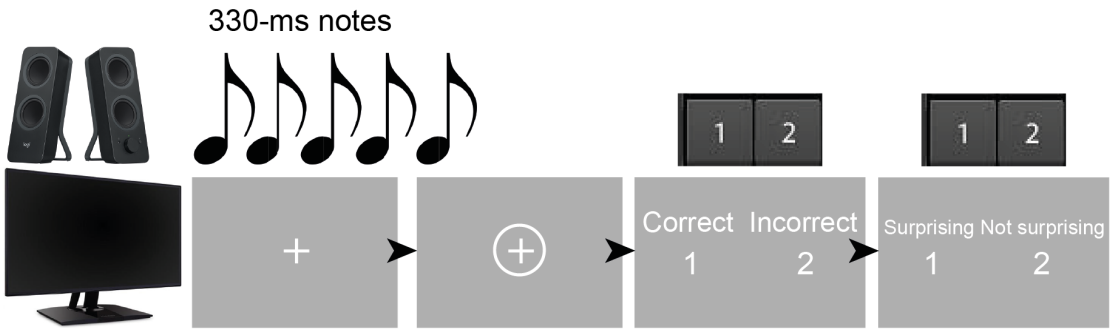
A.

Procedure

Day 1	Day 2	Day 3	Day 4
Pre-test + EEG 40 mins Correctness and surprisal judgments	Passive exposure 1 5 mins	Passive exposure 2 5 mins	Post-test + EEG 40 mins Correctness and surprisal judgments
	Intermediate 1 5 mins Surprisal judgments	Intermediate 2 5 mins Surprisal judgments	
Training 1 30 mins Active reproduction	Training 2 30 mins Active reproduction	Training 3 30 mins Active reproduction	Generalisation 20 mins Surprisal judgments

B.

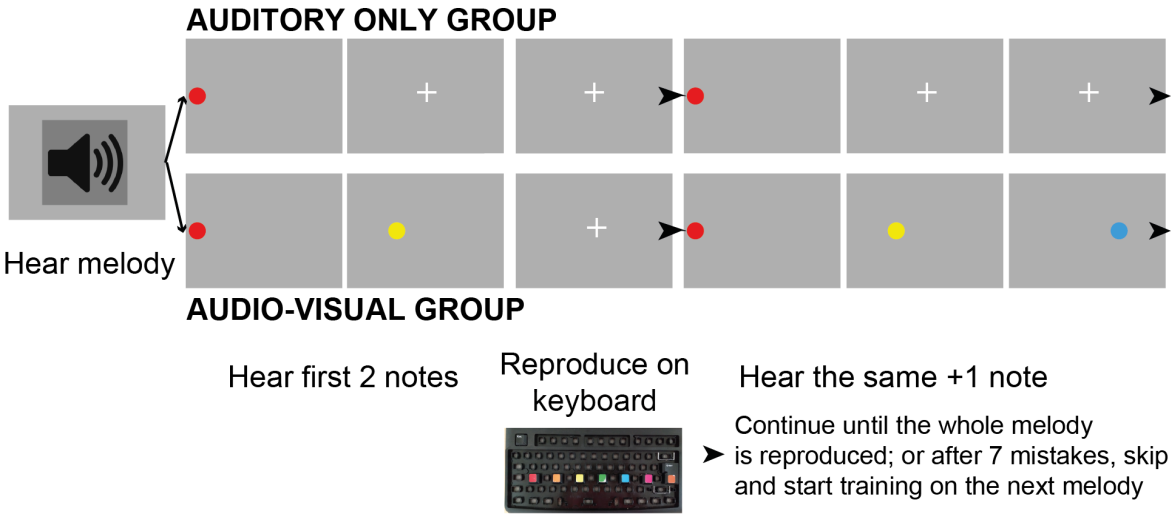
Trial structure
Judgement sessions



Hear melody | Onset of last note | Button press *(pre-, post-test)* | Button press *(pre-, post-test, intermediate, generalisation)*

C.

Training sessions



Hear first 2 notes → Reproduce on keyboard → Hear the same +1 note

Continue until the whole melody is reproduced; or after 7 mistakes, skip and start training on the next melody

Figure 2. A. Schematic representation of the experimental procedure; **B.** Trial structure of the test sessions. Participants were presented with a note sequence and were prompted to judge whether the last note was correct or incorrect (pre- and post-test) and surprising or not surprising (pre-, post-test, intermediate, and generalisation sessions), by pressing 1 or 2 on a computer keyboard; **C.** Trial structure of the three training sessions. Participants listened to a melodic sequence generated by the AMG. Then, they heard the first two notes and needed to reproduce them on the keyboard. If they were correct, the next sequence would increase by one note. If they made a mistake, they could try again. The audio-visual (AV) group was presented with the visual cues of all the notes that they needed to reproduce on screen, whereas the auditory-only (AO) group was only given the first visual cue as a reference.

Training sessions

Participants received training on a computer keyboard which was adjusted to serve as a sound keyboard. A red, an orange, a yellow, a green, a blue, a pink, and a brown sticker were put on keys A, D, G, J, L, ‘, and ENTER, respectively (see Figure 2C). Before the first session only, participants had some familiarisation time with the keyboard. First, they listened to the whole scale ascending three times, while the visual cue corresponding to each note was simultaneously presented on screen. The cues were spatially positioned on the screen in the same configuration as the stickers on the keyboard, i.e. lower notes on the left and higher on the right. Participants were then allowed three minutes to familiarise themselves with the keyboard. To confirm they had basic understanding of the tones, they took a short discrimination test: they listened to pairs of notes for which they were presented with only the first visual cue. They were required to identify the second note and reproduce the note pair on the keyboard. After three attempts, the solution was presented on screen. There were 42 pairs in total, covering all possible note combinations, e.g., C-D, C-E, C-F, etc. All participants passed an arbitrary threshold of 70% correct and proceeded with the training.

Participants attended three 25-minute training sessions on three separate days. The training proceeded as follows. Participants began by hearing a melodic sequence. Then the first 2 notes of the melody were presented. Only after participants reproduced them correctly, the next segment was increased by a note and so on. If they made a mistake, the melodic segment would repeat for further (max. 7) attempts. The difference between the AV and the AO group was that the former was presented on screen with the visual cues of all the notes, whereas the latter was only given the first cue as a reference (to indicate the first note of the sequence), but relied only on the auditory information to reproduce the rest of the sequence.

In the generalisation session, participants were presented with unheard sequences in randomized order, and were asked to judge if the last note was surprising or not surprising. There were 105 trials in total and the session lasted around 20 mins.

Passive exposure sessions

Following the statistical learning literature (e.g., Loui et al., 2010; Rohrmeier et al., 2011), participants attended two (days 2 and 3) passive exposure sessions to three repetitions of the grammatical sequences in randomised order. They were instructed to listen attentively to the melodies. There was a total of 36 sequences and the session lasted approximately 5 minutes.

Intermediate surprisal sessions

After each exposure session (days 2 and 3), participants were presented with sequences terminating on high-probability and low-probability notes and were asked to judge if the last note was surprising or not surprising. There were 36 trials, lasting around 7 minutes in total.

Test sessions

To assess learning, test sessions were conducted before (pre-test: day 1) and after (post-test: day 4) training. Participants were seated in front of a computer, while their EEG was recorded. In the pre-test only, they were informed that they would listen to melodies of an unfamiliar music grammar governed by a set of rules. They were instructed to attend as the melodies would stop at random points and were asked to make two judgements on the last note: 1. correct or incorrect, and 2. surprising or not surprising (Figure 2B). The distinction between correct and incorrect notes is related to the grammar rules, while the probability refers to the information content. Specifically, correctness refers to whether a note is allowed or disallowed by the grammar: a correct note is grammatical, whereas an incorrect note is ungrammatical. Within the correct notes, some have a low-probability while others have a high-probability. The surprisal ratings add to the correctness judgement, as some notes can be surprising but also correct. Therefore, the surprisal ratings were used as a measure of perceived expectedness of the stimuli, which would reflect successful internalized acquisition of the statistical rules.

Furthermore, in the two intermediate sessions we used only high-probability and low-probability stimuli to test participants, as we did not want to expose participants to incorrect stimuli during the "training days" (see the Procedure section under "Intermediate surprisal sessions", p.15). We thus needed to use surprisal judgements as a measure of learning during training as correctness judgements would not be appropriate on those sessions (both high- and low-probability stimuli are correct). Similar studies on implicit sequence learning of melodic sequences have used two-alternative forced-choice recognition tasks that use other ratings apart from correctness to assess learning of an artificial musical system (e.g., Loui & Wessel, 2008; Loui, Wessel, & Kam, 2010). For example, Loui and Wessel (2008) used familiarity ratings, i.e. presented participants with two melodies and asked them to indicate which one is more familiar.

Three practice trials familiarised participants with the task. Across participants the presentation order of the trials was randomised. There were 280 trials in total, and each session lasted around 40 mins.

Working memory task

Participants also completed a working memory span task (WM) (adjusted from the Wechsler Adult Intelligence Scale, WAIS: Wechsler, 1955). In this task, participants were presented with sequences of random numbers from 1 to 9 and had to replicate them on a number pad. Starting from length three, the number of digits was increased by one every time a correct response was made, otherwise the number of digits of the next sequence was reduced by one. This lasted 10 minutes, and the working memory span was calculated as the mean length of the correctly reproduced sequences. Due to technical problems, data from only 31 participants remained for the WM task.

2.4. EEG recording and preprocessing

EEG was recorded from 64 Ag-AgCl electrodes attached to the EGI geodesic sensor net system (HydroCel GSN 64 1.0; EGI System 200; Electrical Geodesic Inc., OR, USA; https://www.egi.com/) and amplified by an EGI Amp 300. The sampling frequency was 500 Hz. The MATLAB Toolbox EEGLAB (Delorme & Makeig, 2004) was used for data preprocessing, and FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011) for data analysis. Data were recorded with an online reference at the right mastoid and re-referenced

to the average of the left and right mastoids. Continuous data were high-pass filtered at 0.5 Hz and then epoched from -0.2 to 0.6 sec after the onset of the last note. Data from electrodes with consistently poor signal quality, as observed by visual inspection and by studying the topographical maps of their power spectra, were removed and replaced by interpolating neighbouring electrodes. Artefact rejection was conducted in a semi-automatic fashion: first, artefactual epochs containing movement, muscle artefacts and saccades were removed after visual inspection, and, second, independent component analysis was used to correct for eye-blink related artefacts. Subsequently, data were detrended, i.e. from each data point of the averaged ERP of each participant, we subtracted the average ERP value. The epoched data was low-pass filtered at 30 Hz and baseline corrected from -0.2 to 0 sec. Five participants were removed due to poor EEG data quality (more than 30% of the trials rejected in at least one of the test sessions) (N_{AO} = 15; N_{AV} = 18).

2.5. Statistical analysis

2.5.1 Behavioural analysis

Behavioural data

Participants' learning was assessed throughout, including pre-test and post-test, intermediate surprisal sessions, and the generalisation session. For the pre-test and post-test sessions, a response was considered correct if a high-probability (HP) or low-probability (LP) note was judged as correct, and if an incorrect note (INC) was judged as incorrect. We performed a 2 (*session:* pre, post) x 2 (*group:* AV, AO) mixed factorial ANOVA on accuracy.

To assess sensitivity to the statistical probabilities of the artificial music grammar, we calculated the percentage of notes judged as surprising within each note probability category in the pre- and post-test, as well as in the intermediate sessions. For the pre- and post-test sessions, we conducted a 3 (*note probability:* HP, LP, INC) x 2 (*session:* pre, post) x 2 (*group:* AV, AO) mixed ANOVA with percentage judged as surprising as the dependent variable. For the intermediate surprisal sessions, we conducted a 3 (*note probability:* HP, LP, INC) x 2 (*intermediate session:* 1, 2) x 2 (*group:* AV, AO) mixed ANOVA with the same dependent variable.

We evaluated sequence reproduction performance at the training sessions by calculating the mean length of correctly reproduced sequences (in number of notes). Due to technical problems with saving the results, four participants were excluded from this analysis

(N_{AO} = 15, N_{AV} = 19). A 3 (*training session:* 1, 2, 3) x 2 (*group:* AV, AO) mixed ANOVA on sequence length.

Finally, we investigated whether sequence reproduction performance predicts learning by performing a multiple linear regression with average length of reproduced sequences in the third training session and group as predictors, and accuracy in the post-test as the dependent variable. In order to test whether sequence reproduction performance or learning depended on working memory skills, as assessed from the digit span task, we conducted two linear regressions: group and working memory were the predictors, and (i) sequence reproduction performance and (ii) learning was the dependent variable.

ERP data

Regions of interest (ROI) analysis. The following regions of interest (ROIs) were used for the analysis, based on previous literature (Carrus, Pearce, & Bhattacharya, 2013; A. R. Halpern et al., 2017) and visual inspection of the ERPs: N100 (80-145 ms) and P200 (150-225 ms) in fronto-central regions (E8, E6, E4, E9, E3, E7, E54, E47 in the EGI configuration, corresponding to: AFz, Fz, FCz, F1, F2, FC1, FC2, Cz in the standard 10-20 system). For each ROI, the mean ERP amplitude, as well as the peak latencies of the N100 and P200 components, were calculated. Two 3 x 2 x 2 mixed, repeated measures ANOVAs were performed (one for N100 and one for P200) with the following factors: *note probability* (HP, LP, INC), *session* (pre-test, post-test), and *group* (AV, AO).

The ERAN was also analysed from 0.140 to 0.220 sec (based on Koelsch, Kilches, Steinbeis, & Schelinski, 2008) at Fpz. The ERAN was identified as the difference in response to LP minus HP notes, and to INC minus HP notes. Two 2 (*session:* pre vs. post) x 2 (*group:* AO vs. AV) mixed ANOVAs with LP-HP and INC-HP as the dependent variables, respectively, were performed.

Non-parametric cluster permutation. In order to explore potential differences of auditory-only vs. audio-visual training on brain responses, we further conducted a non-parametric cluster permutation test (Maris & Oostenveld, 2007). This test first performs independent *t* tests at each data point, and then identifies clusters of electrodes that exceed a defined threshold and have the same sign. Subsequently, the cluster-level statistic is calculated as the sum of the *t* values of the cluster. Finally, the maximum value of the cluster-level statistic is evaluated by

calculating the probability that it would be observed under the assumption that the two compared conditions are not significantly different.

Specifically, we compared AV vs. AO on their brain responses to low-probability minus high-probability notes (LP-HP) within the first 500 ms after note onset. The permutation distribution was extracted from the statistic values of independent samples *t*-tests based on 500 random permutations. The probability threshold was set at $p = .05$. Subsequently, we performed independent samples *t*-tests on the average values of the identified clusters.

3. Behavioural results

3.1. Pre- and post-test sessions

A 2 (*session*: pre-test, post-test) x 2 (*group*: AO, AV) mixed ANOVA showed that participants successfully learned the grammar (main effect of *session*: $F(1,36) = 67.751$, $p < .001$, $\eta^2 = .653$) (Figure 3A). There was no effect of group or interaction between the variables ($p > .7$).

As the random notes were neither correct nor incorrect based on the grammar, we did not expect to see any difference in the percentage judged as correct in the pre-test vs. the post-test. This was confirmed by a 2 (*session*) x 2 (*group*) ANOVA which showed no significant main effects or interaction between the variables ($p > .2$).

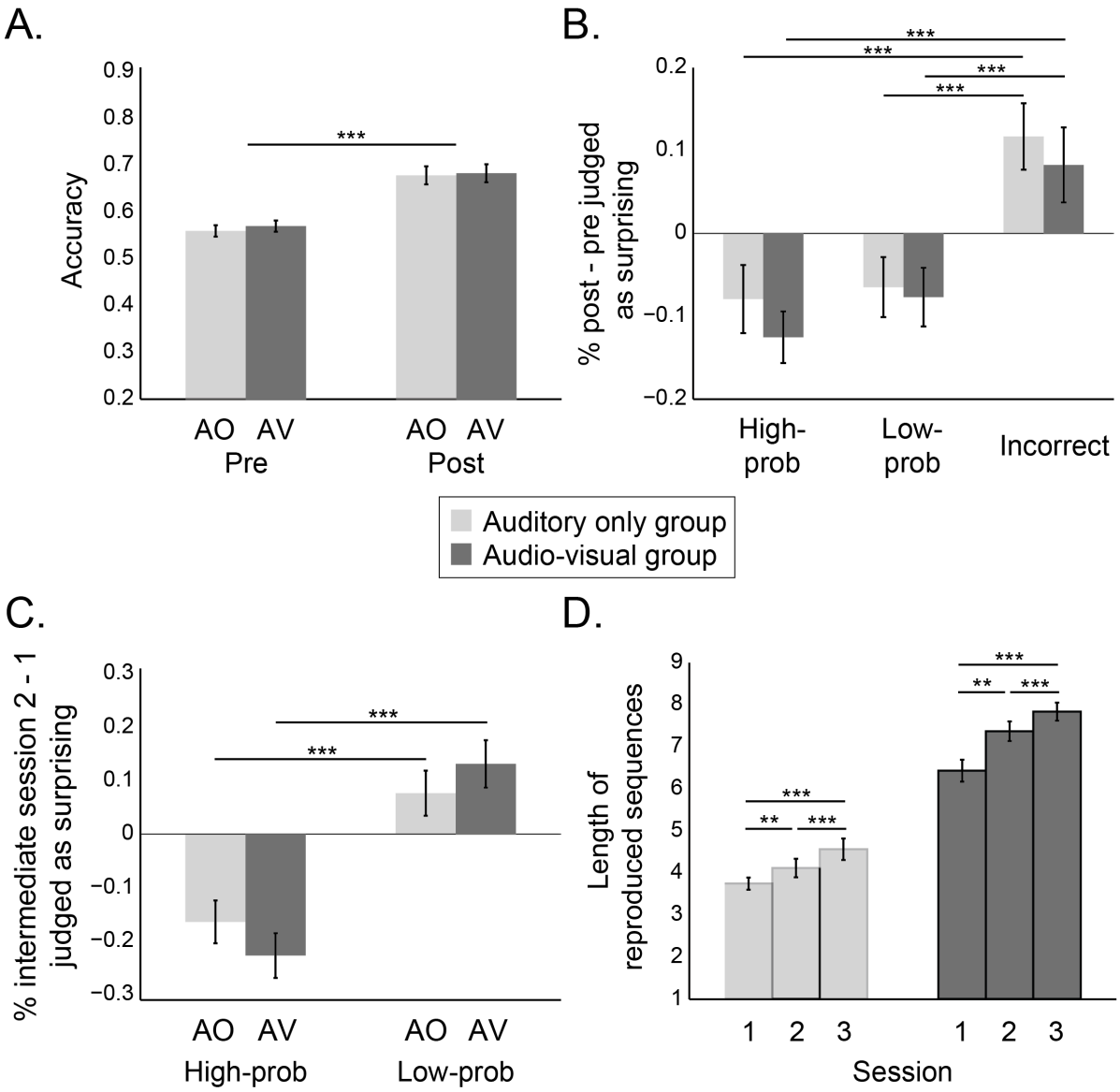


Figure 3 - A. Mean accuracy for the auditory-only (AO) and the audio-visual (AV) training groups, separately in the pre-test and post-test sessions; **B.** Post- minus pre-test differences between mean percentage of notes judged as surprising in the auditory-only (AO) vs. the audio-visual (AV) training groups, separately for high-probability (HP), low-probability (LP), and incorrect (INC) notes; **C.** Intermediate session 2 minus session 1 differences between percentage of notes judged as surprising, separately for each group and note probability type; **D.** Mean length (number of notes) of correctly reproduced sequences across the three training sessions, separately for the AO and the AV group. Error bars represent ± 1 standard error mean (*SEM*). * $p < .050$, ** $p < .010$, and *** $p < .001$.

Furthermore, participants judged low-probability (LP) and incorrect notes (INC) as more surprising than notes of high probability (HP) after training, showing that learning made them more sensitive to the statistical probabilities of the grammar (Figure 3B). A 3 x 2 x 2 mixed ANOVA revealed significant main effects of *note probability* ($F(2,72) = 45.566$, $p <$

.001, η^2 = .559), as well as a *note probability* * *session* interaction (F(2,72) = 28.081, p < .001, η^2 = .438). Post hoc contrasts revealed that participants judged HP and LP notes as less surprising in the post-test compared to the pre-test session (HP: t(37) = -3.982, p < .001, *Cohen's d* = -.650; LP: t(37) = -2.841, p = .007, *Cohen's d* = -.472), whereas the opposite was found for the INC notes, i.e. participants judged them more as surprising in the post-session (t(37) = 3.331, p = .002, *Cohen's d* = .559). In both sessions, INC notes were judged as surprising significantly more often than the HP notes (pre: t(37) = 3.913, p < .001, *Cohen's d* = .635; post: t(37) = 7.108, p < .001, *Cohen's d* = 1.159), as well as than the LP notes (pre: t(37) = 2.623, p = .013, *Cohen's d* = .428; post: t(37) = 6.741, p < .001, *Cohen's d* = 1.102). The LP notes were more often judged to be surprising than the HP notes (pre: t(37) = 2.362, p = .024, *Cohen's d* = .387; post: t(37) = 4.926, p < .001, *Cohen's d* = .801). There was no effect of group or any other effect or interaction between the variables (p > .3). There was no difference in the percentage of random notes judged as surprising in the pre-test vs. the post-test (p > .2).

3.2. Intermediate sessions

Participants' surprisal judgements in the two intermediate sessions were also evaluated by a 2 (*note probability:* HP, LP) x 2 (*intermediate session:* 1, 2) x 2 (*group:* AO, AV) mixed ANOVA (Figure 3C). Results revealed significant main effects of *session* (F(1,36) = 4.860, p = .034, η^2 = .119) and *note probability* (F(1,36) = 5.135, p = .030, η^2 = .125). There was also a significant *note probability* * *session* interaction (F(1,36) = 49.013, p < .001, η^2 = .577). Post hoc analysis revealed that participants judged HP notes as significantly less surprising in the second compared to the first session (HP: t(37) = -6.741, p < .001, *Cohen's d* = -1.094), whereas the opposite was found for the LP notes, i.e. participants judged them as more surprising in the second session (t(37) = 3.411, p = .002, *Cohen's d* = .554). Further, in the first session, HP notes were judged as surprising significantly more often than the LP notes (t(37) = -2.080, p = .045, *Cohen's d* = -.340), whereas the opposite effect was observed in the second session (t(37) = 6.103, p < .001, *Cohen's d* = .996). There was no effect of group or interaction between the variables (p > .2). Therefore, HP notes were judged as less surprising in the second compared to the first session, whereas the opposite was found for the LP notes, and this effect did not differ between groups.

3.3. Generalisation test

A 3 (*note probability*: HP, LP, INC) x 2 (*group:* AO, AV) mixed ANOVA demonstrated that participants successfully differentiated between the statistical probabilities of unheard sequences (main effect of *note probability:* $F(2,72) = 10.166$, $p < .001$, $\eta^2 = .220$). Planned contrasts revealed that participants judged LP notes as more surprising than HP ones ($t(37) = 2.362$, $p = .024$, *Cohen's d* = .383), and INC notes more surprising than both HP ($t(37) = 3.913$, $p < .001$, *Cohen's d* = .635) and LP ones ($t(37) = 2.623$, $p = .013$, *Cohen's d* = .426). There was no effect of group or interaction between the variables ($p > .3$).

3.4. Training

Performance during training improved incrementally and the AV group did substantially better in all sessions, as confirmed by a 3 (*session:* 1, 2, 3) x 2 (*group:* AO, AV) mixed ANOVA (Figure 3D). In particular, results revealed a main effect of *session:* $F(2,64) = 37.676$, $p < .001$, $\eta^2 = .541$. Further, the AV group was able to reproduce longer sequences overall compared to the AO group (main effect of *group:* $F(1,32) = 111.335$, $p < .001$, $\eta^2 = .777$). Results also revealed a significant *session * group* interaction ($F(2,64) = 3.369$, $p = .041$, $\eta^2 = .095$). Planned contrasts showed that the AV group were significantly better in all sessions compared to AO (session 1: $t(32) = 8.407$, $p < .001$, *Cohen's d* = .691; session 2: $t(32) = 9.893$, $p < .001$, *Cohen's d* = .621; session 3: $t(32) = 9.867$, $p < .001$, *Cohen's d* = .640). Paired *t*-tests showed that both groups performed better in the third session compared to the first (AO: $t(14) = 5.140$, $p < .001$, *Cohen's d* = 1.327; AV: $t(18) = 5.762$, $p < .001$, *Cohen's d* = 1.322) and second session (AO: $t(14) = 4.392$, $p = .001$, *Cohen's d* = 1.134; AV: $t(18) = 4.055$, $p = .001$, *Cohen's d* = .930), and better in the second compared to the first session (AO: $t(14) = 3.277$, $p = .006$, *Cohen's d* = .846; AV: $t(18) = 4.062$, $p = .001$, *Cohen's d* = .932).

To further investigate whether one of the two groups improved more from training session 1 to training session 3, we conducted a paired samples *t*-test on training performance (i.e. length of the replicated sequences) between the training session 3 minus training session 1 differences of the AV vs. the AO group. Results revealed that the AV group improved more ($M = 1.404$, $SD = 1.061$) compared to the AO group ($M = .815$, $SD = .614$), but that was only marginally statistically significant ($t(32) = -1.906$, $p = .066$).

3.5. Training predicting learning

In order to investigate whether better sequence reproduction performance during training leads to better learning, a multiple linear regression analysis was performed to predict learning (accuracy in the post-test) with the predictors group (AO, AV) and sequence reproduction (average length of reproduced sequences in the third training session). Results showed that neither group nor sequence reproduction significantly predicted learning ($p > .1$), and the model was not significant overall ($F(2,31) = 1.506$, $p = .238$, $R^2 = .089$), suggesting that sequence reproduction during training does not necessarily ensure successful learning in either of the groups.

3.6. Working memory predicting sequence reproduction and learning

We tested whether sequence reproduction or learning depended on working memory capacity, as assessed by the digit span task. We conducted two linear regression analyses with group (AO, AV) and working memory performance as predictors, and (i) sequence reproduction and (ii) learning as the dependent variable. In the first regression, group was a significant predictor of sequence reproduction ($p < .001$), but not working memory ($p = .880$); the model was overall significant ($F(2,20) = 20.043$, $p < .001$, $R^2 = .667$). In the second regression, neither group nor working memory were significant predictors of learning ($p > .1$), and the model was not significant overall ($F(2,20) = 1.269$, $p = .303$, $R^2 = .113$).

4. ERP results

4.1. N100 time window (80 – 145 ms)

As shown in Figure 4A, C, D, neural sensitivity to the statistical properties of the grammar was reflected in the N100 amplitude of the auditory-only group (AO) as the N100 was higher in response to low-probability (LP) than high-probability (HP) notes in the post-test; the audio-visual group (AV) did not show a similar differentiation. Confirming this, a mixed ANOVA yielded a significant *session * note probability * group* interaction ($F(2,62) = 4.290$, $p = .018$, $\eta^2 = .122$), as well as a significant *session * group* interaction ($F(1,31) = 4.140$, $p = .050$, $\eta^2 = .118$). Planned contrasts showed that the N100 was higher in response to LP compared to HP notes in the post-test in the AO group ($t(14) = -2.319$, $p = .036$, *Cohen's d* = .599), whereas that was not significant in the pre-test ($t(14) = .979$, $p = .344$, *Cohen's d* =

.253). The contrast between LP and HP was not significant in the AV group for neither of the sessions (pre: $t(17) = -.911$, $p = .375$, *Cohen's d* = .215; post: $t(17) = 1.617$, $p = .124$, *Cohen's d* = .381). Further, the N100 amplitude in response to HP notes became less negative from pre- to post-test in the AO group ($t(14) = 3.193$, $p = .007$, *Cohen's d* = .866). No effect was found for the AV group ($p >$. 1) (Figure 4B, D). There was no other significant effect nor interaction between the factors ($p > .4$).

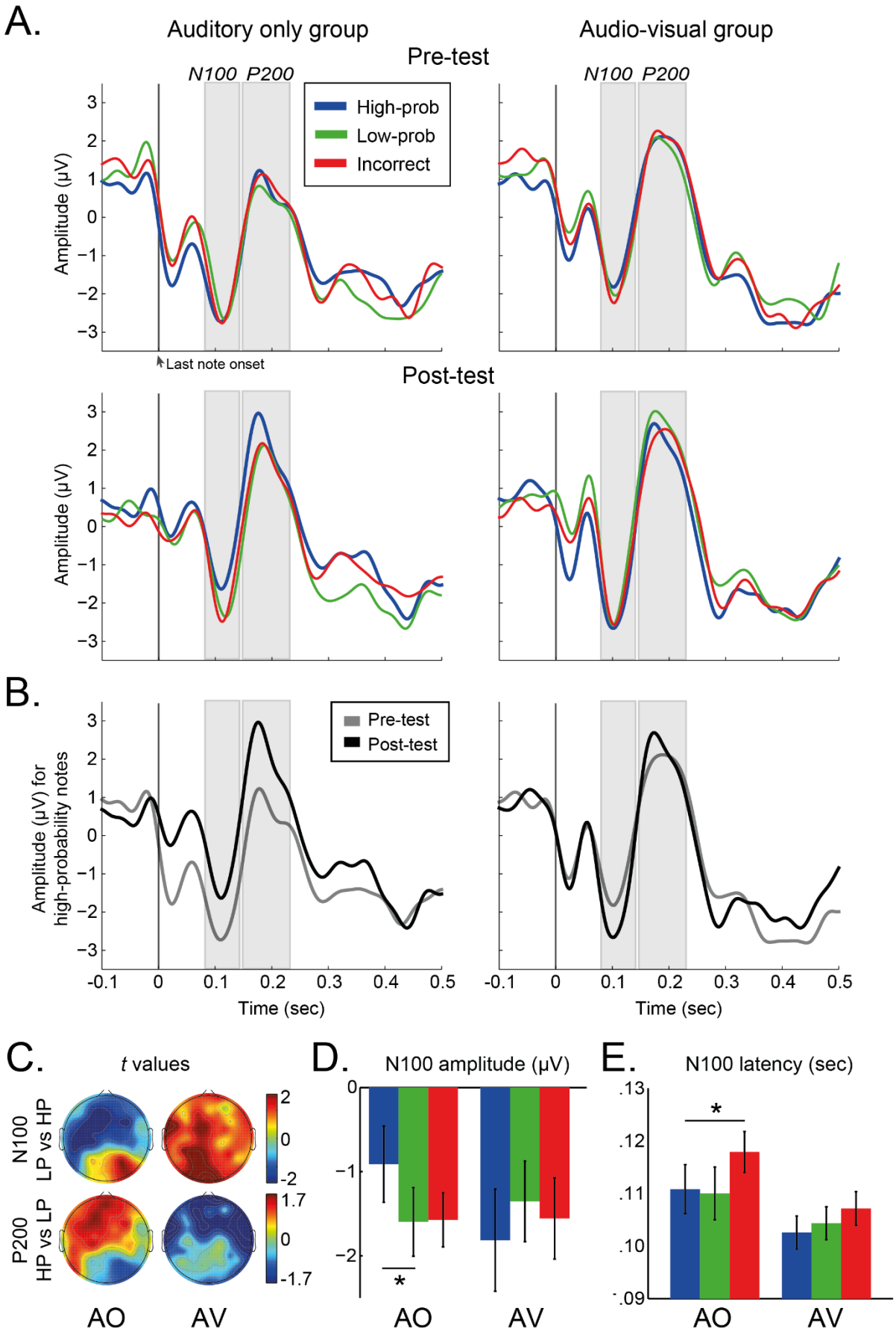

A.
Auditory only group
Audio-visual group
Pre-test
N100 P200
High-prob
Low-prob
Incorrect
Amplitude (μV)
Last note onset
Post-test
B.
Pre-test
Post-test
Amplitude (μV) for high-probability notes
Time (sec)
C.
t values
N100 LP vs HP
P200 HP vs LP
AO
AV
D.
N100 amplitude (μV)
E.
N100 latency (sec)

Figure 4 – A. Grand average ERPs in response to high-probability, HP (blue), low-probability, LP (green), and incorrect, INC (red) notes in fronto-central brain regions, for pre- (top) and post-test (bottom) sessions, separately for the auditory-only, AO and the audio-visual group, AV; **B.** Grand average ERPs in response to high-probability notes for the AO (left) and the AV group (right), for pre- (gray) and post-test (black); **C.** T-value topoplots between LP vs. HP notes for the N100 (top), and between HP vs. LP notes for the P200 (bottom), in the AO (left) and AV (right) groups; **D.** Mean amplitudes for the N100 (80-145 ms after the onset of the last note) in the post-test separately for the AO and the AV group, and for HP (blue), LP (green), and INC (red) notes. **E.** Latency of the N100 in the post-test for AO (left) and AV (right), for HP, LP, and INC notes. Error bars represent ± 1 standard error mean (*SEM*). * $p < .050$, ** $p < .010$, and *** $p < .001$.

The enhanced neural sensitivity of the AO group to the statistical regularities of the grammar was also reflected on the latency of the N100 (Figure 4E). A mixed ANOVA revealed a significant *session* * *note probability* interaction ($F(2,62) = 4.076$, $p = .022$, $\eta^2 = .116$). In the AO group, the latency of the N100 in response to INC notes was longer in the post compared to the pre session ($t(14) = 2.620$, $p = .020$, *Cohen's d* = .676). Further, in the pre-test, the AO group showed shorter latencies to INC than LP notes ($t(14) = -2.300$, $p = .037$, *Cohen's d* = .734), while in the post-test they showed longer latencies to INC than HP notes ($t(14) = 2.567$, $p = .022$, *Cohen's d* = .663). There was no difference in the post vs. pre latency in the AV group ($p > .3$). There was no other significant effect nor interaction between the variables ($p > .1$).

4.2. P200 time window (150 – 225 ms)

There was a marginal effect of *session* $F(1,31) = 11.936$, $p = .002$, $\eta^2 = .278$) (Figure 4A), as the P200 amplitudes increased from the pre-test ($M = 1.263$, $SD = 2.258$) to the post-test session ($M = 2.012$, $SD = 2.214$) There was no other effect or interaction between the variables ($p > .2$).

There was no significant effect nor interaction between the variables on the P200 latencies ($p > .1$).

4.3. ERAN time window (140 – 220 ms)

Two 2 (*session:* pre vs. post) x 2 (*group:* AO vs. AV) mixed ANOVAs with LP-HP and INC-HP as the dependent variables, respectively, were performed on the ERAN (see Figure S2 in Supplementary materials). The LP-HP ANOVA did not reveal any significant

main effect or interaction ($p > .1$). However, the INC-HP difference increased from pre-test to post-test (main effect of *session*: $F(1,31) = 5.836$, $p = .022$, $\eta^2 = .158$), but there was no other effect or interaction between the variables ($p > .2$). The analysis in detail and a figure are included in Supplementary materials.

4.4. Non-parametric cluster permutation analysis

We compared brain responses to low-probability minus high-probability (LP-HP) notes between the AV vs. the AO group with a non-parametric cluster permutation test. Results revealed two fronto-central clusters, the first from 100-200 ms ($p = .023$) and the second from 250-350 ms ($p = .031$) (see Figure S3A for the topography in the Supplementary materials). In both clusters, the AO group showed an enhanced negativity in response to LP compared to HP notes, which was not the case for the AV group (Figure S3C). This is also evident from the difference ERP plots in Figure S3B, in which there is a negative-going wave in the AO group, but not in the AV group, in both time windows. Both clusters were further statistically tested by independent samples *t*-tests confirming the group difference (0.1-0.2: $t(28) = 3.484$, $p = .002$, *Cohen's d* = 1.272; 0.1-0.2: $t(28) = 3.048$, $p = .005$, *Cohen's d* = 1.113). Considering the time windows and the midfrontal topography of both clusters, they might represent the N100-P200 effects we observed, and a later negative-going component resembling an N200, respectively. These findings can be visualised in Figure S3 (Supplementary).

5. Discussion

Our main goal was to investigate the effect of multisensory music learning using visual aids, and on the respective neural correlates, as well as examine the distinction between sequence reproduction and learning of the statistical regularities of an unfamiliar music grammar. Our study was the first, to our knowledge, to investigate statistical learning over multiple sessions conducted on separate days under different training regimes. In contrast to previous studies showing that multimodality is beneficial for learning (e.g., Brünken, Plass, & Leutner, 2004; Cleary, Pisoni, & Geers, 2001; Tierney, Bergeson-Dana, & Pisoni, 2008), we found that visual aids boosted sequence reproduction but did not improve

statistical learning, suggesting that performance during training and actual learning are two distinct or relatively independent processes. This was also reflected in the neural correlates, as training without visual aids was associated with increased sensitivity to the statistical properties of the musical style.

As expected, participants who received musical training with visual aids were able to reproduce considerably longer sequences compared to those without visual aids. Previous studies have demonstrated that visual cues engage working memory resources in visuo-spatial (e.g., Gathercole & Alloway, 2008) and arithmetic (e.g., St Clair-Thompson, Stevens, Hunt, & Bolder, 2010) domains. Thus, it might have been easier for the audio-visual group to reproduce the sequences by relying on short-term memory of the visual cues. However, this mechanism might only be efficient for immediate reproduction, and not necessarily beneficial for longer-term acquisition of knowledge nor for developing an enhanced sensitivity to the underlying rules. Future studies are needed to investigate the efficacy of visual aids in longer-term learning over periods of weeks, months and years, of the statistical regularities of a certain musical style.

Interestingly, the superior sequence reproduction using visual aids was not reflected in greater knowledge of the grammar, as the audio-visual and the auditory-only groups performed equally well in the test after training. This is in line with previous studies demonstrating that performance during training and learning are distinct processes (e.g., Kantak & Winstein, 2012; Lee & Genovese, 1988; Schmidt & Bjork, 1992). According to the Desirable Difficulties theory (Bjork & Bjork, 2011; Bjork, 1994), learning and retention can improve with the use of more difficult and challenging material during acquisition. That is, learning can be substantially improved by superficial changes in the presentation of the material, e.g., using a letter format that is harder to read (Diemand-Yauman, Oppenheimer, & Vaughan, 2011), due to an increasing level of cognitive engagement during learning, which enhances subsequent recall. For example, when a response word associated with a stimulus word is generated rather than read by the subject, later recall is improved, attributed to strengthening of memory associating the two items (Hirshman, & Bjork, 1988). Further, learning and generalization of knowledge beyond specific recall improves when two sets of information are interleaved rather than grouped into separate blocks (Richland, Bjork, Finley, & Linn, 2005). Likewise, participants exhibit better understanding of paragraphs with deleted letters than paragraphs with intact letters (Maki, Foley, Kajer, Thompson, & Willert, 1990). A beneficial effect of using a hard-to-read letter font on memory recall has been also demonstrated in classroom settings (Diemand-Yauman et al., 2011). Our findings suggest that

participants' reliance on the auditory information constituted an extra difficulty in the sequence reproduction task, which may have led to enhanced neural sensitivity to the statistical properties of the learned material (Craik, & Tulving, 1975). In contrary, visual cues made the task much easier, thus requiring less cognitive engagement, and providing no enhancement of subsequent retrieval.

Another explanation might lie on the modality specificity of statistical learning (see Frost, Armstrong, Siegelman, & Christiansen, 2015 for a review). In particular, it is possible that training with visual cues could have led to visual learning dominating over auditory learning, thus producing a deficit in the test sessions. Transfer of learning across modalities has been found to be limited (Redington & Chater, 1996; Tunney & Altmann, 1999), while there are qualitative learning biases among the auditory, visual, and tactile modalities (Conway & Christiansen, 2005; Emberson, Conway, & Christiansen, 2011). Further, there is evidence suggesting that sometimes multimodality results in cross-modal competition (e.g., Robinson & Sloutsky, 2013; Sinnett, Spence, & Soto-Faraco, 2007; see Spence, 2009 for a review), where the more salient a stimulus representation, the more it dominates the competition (Rapp & Hendel, 2003). Furthermore, visual dominance effects have been demonstrated, where participants failed to detect a tone when it was simultaneously presented with a light (Colavita effect: Spence, 2009). Visual dominance can be modulated but not completely reversed with selective attention (Sinnett et al., 2007), however, in our study, there were no instructions on direction of attention. Contrary to our findings, previous studies have shown that combining auditory with visual information can be beneficial for sequence learning (e.g., Brünken et al., 2004; Cleary, Pisoni, & Geers, 2001; Coull, Tremblay, & Elliott, 2001; Pisoni & Cleary, 2004; Seitz, Kim, & Shams, 2006; Shams & Seitz, 2008; Tierney et al., 2008). For example, participants presented with audio-visual stimuli have been found to reproduce longer sequences compared to when they were presented with audio or visual only cues (Simon task: Cleary, Pisoni, & Geers, 2001; Tierney et al., 2008). Likewise, in our study, the participants trained with audio-visual stimuli were able to reproduce longer sequences during training (compared to participants trained with auditory stimuli only) but this benefit did not result in better learning of the statistical regularities of the artificial music grammar.

Importantly, however, the aforementioned studies used the same stimulus modalities for the learning and testing phases (auditory-only, visual-only, or audio-visual). Here we need to note that, because the multi-modality advantage is often used as a justification for visual aids in music learning, we tested both groups without the visual cues (even if they were used during training). The reason we did not assess learning in the visual or audio-visual modality

was that our primary focus was on music learning, i.e. we aimed to study whether visual aids improve music learning. Nevertheless, it would be interesting to see how the AV group would perform if tested with audio-visual stimuli. Future studies are needed to investigate the effect of audio-visual training on multi-modal learning vs. modality-specific learning. In our study, visual cues improved sequence reproduction but not statistical learning, thus providing corroborating evidence for modality-specific learning.

Furthermore, our study differs from the aforementioned studies in additional aspects. While the auditory cues of the Simon task (Cleary, Pisoni, & Geers, 2001; Tierney et al., 2008) consisted of the names of the colours (i.e. participants heard the names of the colours they needed to reproduce), our participants heard tones and were required to find the corresponding keys – a much more difficult task, especially for non-musicians. Further, there were differences in the method of testing the knowledge after training. For example, Cleary and colleagues (2001) assessed performance based on the length of the sequences participants accurately reproduced during training, while Tierney and colleagues (2008) used familiarity ratings on a scale from 1 (least familiar) to 7 (most familiar). In contrast, we asked for grammaticality judgements of specific notes (correct or incorrect). We could speculate here that the latter requires better learning of the statistical regularities of the music grammar since it cannot rely purely on working memory (reproduction) or just familiarity. This, however, did not occur for the audio-visual group since the multimodal nature of the stimuli made the immediate reproduction task so much easier to achieve.

Participants showed generalization of their knowledge to new melodies, and there was no difference between the groups. Previous studies have demonstrated generalization effects after a brief exposure to novel music (Loui & Wessel, 2008; Loui et al., 2010). This suggests that both groups internalized the underlying rules of the new grammar and were able to extrapolate their knowledge to unheard melodies. Participants also exhibited sensitivity to notes with different levels of predictability, as they scored incorrect, low-probability, and high-probability notes as increasingly more surprising.

On the neural level, after training, the auditory-only group showed enhanced N100 in response to low-probability compared to high-probability notes, but this effect was not present in the audio-visual group, which exhibited no differences between probability types. The N100 has been previously linked to expectation (e.g., Daikoku, Yatomi, & Yumoto, 2015; Stefan Koelsch & Jentschke, 2010; Omigie et al., 2013). For example, Omigie and colleagues (2013) found that a similar enhanced early frontal negativity was elicited in response to unpredictable notes only in controls, but not in amusic patients, which showed

impaired explicit knowledge of the music for the latter group. Abla and colleagues (2008) found that, after first exposure to a novel grammar, the N100 was increased in response to unexpected words in high-learners only. In our study, the auditory-only group exhibited an increased sensitivity to the statistical properties of the artificial music grammar, while the audio-visual group had a less robust representation of the material. Thus, the N100 in response to unexpected notes could reflect the strength of the prediction error: better learning would lead to formation of strong predictions, which, if violated, would elicit an increased N100 amplitude. In contrary, there is not much prediction error when the predictions are weak. Based on predictive coding, the brain inhibits the neural responses to predictable stimuli in order to achieve efficient processing (Friston, 2005). Auditory, task-relevant training might have led to the auditory-only group forming more specific expectations of how music should unfold due to better knowledge of the statistical properties of the grammar, thus creating stronger prediction error signals when those expectations were violated. On the other hand, the audio-visual group might have had less sensitivity to the subtle statistical regularities of the grammar due to modality-specific learning.

The P200 component was larger in the post-test compared to the pre-test session and this was not different between groups. This early positive deflection is reported to be enhanced after a prolonged training (e.g., Bosnyak, Eaton, & Roberts, 2004; Reinke, He, Wang, & Alain, 2003). Lexical processing with familiar words induces larger P200 amplitudes than unfamiliar words (Perfetti, & Wang, 2006; Stuellein et al., 2016). Liu and colleagues (2006) observed larger P200 amplitudes in response to familiar compared to unfamiliar Chinese characters and English words, proposing a potential link between P200 and processing speed. In Stuellein and colleagues (2016), recently seen words were associated with larger P200 and faster response times compared to unseen words during the experiment, suggesting quicker lexical access and semantic integration in memory. In the auditory domain, participants' P200 amplitudes showed a robust increase after training associated with learning to distinguish two synthetic speech variants of the syllable /ba/ (Tremblay & Kraus, 2002). The authors suggested that this component reflects a pre-attentive mechanism linked to enhanced perception as a result of learning.

The latency of the N100 was delayed for unpredictable notes compared to predictable notes in the auditory-only group, but not in the audio-visual group. The latency of this component has been previously associated with processing speed (Polich, Ellerson, & Cohen, 1996), and correlated with task difficulty (Goodin, Squires, & Starr, 1983). Therefore, in our study, the N100 latency effect could reflect that the auditory-only group processed faster and

easier the expected events compared to unexpected, whereas the audio-visual group did not differentiate the varying types of expectancy. As manifested by the early neural responses, the results provide evidence for successful early discrimination of subtle statistical differences and, therefore, increased neural sensitivity in the auditory-only group, which was not apparent in the audio-visual group.

Our results revealed no substantial modulation of the ERAN by training method or clear effects of learning. This is an unexpected finding, considering that the ERAN has previously been associated with violation of syntax in Western tonal music in both chord and melodic sequences (Koelsch et al., 2000, 2008; Loui, Grent, Torpey, & Woldorff, 2005; Pearce & Rohrmeier, 2018; Steinbeis & Koelsch, 2008). Specifically, ERAN is increased in response to completely ungrammatical or stylistically unpredictable (but not ungrammatical) chords (Leino, Brattico, Tervaniemi, & Vuust, 2007; Steinbeis et al., 2006), but diminished in response to expected elements, after a certain context has been established (Leino et al., 2007). One explanation could be that, in our study, participants were not able to infer harmony from the presented melodies, which participants in studies using Western music are potentially performing. Koelsch et al. (2016) supported that ERAN does not necessarily reflect processing of local dependencies, as local irregularities confound with hierarchical structure (Kim, Kim, & Chung, 2011; Villarreal, Brattico, Leino, & Østergaard, 2011). In our study, we tracked the statistical properties of the melodic sequences with the IDyOM computational model. This model's probability estimates are long-term (based on prior learning of the grammar) and contextual (the probabilities are conditional on the entire preceding melodic sequence). Therefore, it is unexpected that the ERAN did not capture robustly the statistical learning process, and future studies are needed to investigate this further. In contrary, the N100 is more sensitive to local expectancy violation in various modalities, such as auditory, visual, temporal (Duzcu, Özkurt, Mapelli, & Hohenberger, 2019; Michalski, 2000). This component has been especially reflective of statistical learning, with larger N100 in response to tones with lower transitional probability compared to tones with higher probability (Abla et al., 2008; Halpern et al., 2017; Moldwin, Schwartz, & Sussman, 2017; Paraskevopoulos et al., 2012; Zioga, Harrison, Pearce, Bhattacharya, & Luft, 2020).

In line with the ROI analysis, a whole-head cluster permutation analysis provided corroborating evidence for the increased sensitivity of the auditory-only group to the subtle statistical properties of the musical grammar. Besides the N100 and P200 time windows, a later, negative-going wave was revealed, which was increased in response to low- compared to

high-probability notes in the auditory-only group, but not in the audio-visual group. The topography and time of this resemble the N200 component observed in prediction error studies (Ferdinand, Mecklinger, & Kray, 2008; Hajihosseini & Holroyd, 2013; Kopp & Wolff, 2000; Oliveira, McDonald, & Goodman, 2007). Both the N100 and N200 are sensitive to prediction errors, the former as a more immediate response, whereas the latter represents more top-down, later processes. Specifically, the N200 is elicited by deviant stimuli (Hoffman, 1990). This was initially identified in oddball paradigms, where a continuously-presented stimulus is interrupted by infrequent stimuli (Näätänen & Picton, 1986). The N200 is evoked to prediction errors when a mismatch between an expected and the sensory input is detected (Ferdinand et al., 2008; Hajihosseini & Holroyd, 2013; Kopp & Wolff, 2000; Oliveira et al., 2007). For example, in a music performance study (Maidhof, Vavatzanidis, Prinz, Rieger, & Koelsch, 2010), pianists showed an N200 component following unexpected notes, which was enhanced during performance than during perception of musical sequences. Therefore, our findings suggest enhanced sensitivity to statistical regularities as evidenced from both early (unconscious) and late (conscious) neural responses after musical training with auditory only cues.

Previous studies have demonstrated multisensory neuroplastic changes in the auditory cortex after multisensory music training (Kuchenbuch, Paraskevopoulos, & Herholz, 2014; Pantev, Paraskevopoulos, Kuchenbuch, Lu, & Herholz, 2015; Paraskevopoulos, Kraneburg, Herholz, Bamidis, & Pantev, 2015; Paraskevopoulos, Kuchenbuch, Herholz, & Pantev, 2014; Paraskevopoulos, Kuchenbuch, Herholz, Foroglou, et al., 2014; Paraskevopoulos, Kuchenbuch, Herholz, & Pantev, 2012a). Long-term musical training is associated with enhanced multisensory, audio-visual integration and neuroplastic changes in the auditory cortex, whereas short-term training affects the processing of each modality separately (Pantev et al., 2015). In an MEG audio-tactile mismatch paradigm (Kuchenbuch et al., 2014), musicians showed enhanced higher-order audio-tactile integration as evidenced by their brain responses to multisensory deviant stimuli, whereas non-musicians demonstrated only bottom-up processing driven by tactile stimuli. In an audio-visual integration study on musicians, Paraskevopoulos and colleagues (2015) musicians showed increased connectivity in areas relying on the contribution of the left inferior frontal cortex in response to auditory pattern violations, which was interpreted as better audio-visual cortical integration. In contrary, non-musicians had more sparse integration of visual and auditory information and relied more on the visual information. Considering that our participants were non-musicians, it could be that

training with visual cues might have triggered an over-reliance on these cues, which then distracted them from the statistical regularities of the music.

This analysis did not reveal an effect of training with visual aids in visual processing or other posterior regions. This is in contrast with previous neuroscientific work on multisensory learning (Pantev et al., 2015; Paraskevopoulos, Chalas, Kartsidis, Wollbrink, & Bamidis, 2018; Paraskevopoulos et al., 2015). For example, in a multisensory oddball paradigm, Paraskevopoulos et al. (2018) demonstrated that deviant visual stimuli were associated with activation of middle temporal and visual association areas, and that was not different between musicians and non-musicians. It could be thus expected that, in our study, unexpected notes would elicit a response in visual areas in the audio-visual training group. However, this could not be directly investigated as our participants were presented with the auditory stimuli only during the post-test session, which means there was no incongruency in relation to the visual signals used for training as they were not present in the test sessions (pre and post). In other words, in our experiment, the prediction error was always auditory, rather than an incongruent audio-visual stimulus pair as in the aforementioned studies (Pantev et al., 2015; Paraskevopoulos et al., 2015). Finally, a study on visual processing found that the amplitude of the N170 component varied depending on reference method, while latency was independent across methods (Joyce & Rossion, 2005), which might suggest that the mastoid references used here could potentially contribute for the lack of effects in visual areas.

Our experimental design is not without limitations. First, it is possible that the effects we observed are specific to the artificial grammar used, which is necessarily limited in scope and may not have provided sufficient challenge to distinguish performance between the groups. However, the behavioural findings do not suggest a ceiling effect, which speaks against this possibility. Second, because the visual modality tends to dominate when attentional resources are depleted (Robinson, Chandra, & Sinnett, 2016), it could be that visual cues disrupted learning by increasing task demands, requiring participants to make associations between the sounds, the keys, and the visual cues, while the auditory-only group needed to only map the keys with the sounds. Furthermore, it would be interesting to examine potential super-additivity effects of the audio-visual integration, i.e. whether multisensory stimulation elicits higher neural activation than the sum of the unisensory stimuli (Stanford & Stein, 2007). Super-additivity effects have been previously demonstrated in various domains, such as the audio-visual (Nichols & Grahn, 2016; Paraskevopoulos et al., 2018) and the audio-tactile (Hoefer et al., 2013). The absence of a visual-only condition comprises a limitation of our study, which could be investigated in future studies. Finally, we

acknowledge that we might have potentially missed effects around the temporal and visual areas due to the average mastoid EEG re-referencing. Future studies are necessary to more appropriately explore the effect of visual aids on music learning on ERPs at temporal sites.

We conclude that musical training with visual aids is not necessarily beneficial for learning; rather it might serve as a distraction from encoding the main material. On the other hand, training without visual aids can lead to an enhanced understanding of the statistical subtleties of an unfamiliar music grammar, as evidenced by an increased sensitivity to statistical regularities at the neural level. Therefore, adding visual cues might give the illusion of learning as we can reproduce long sequences, however, it impairs actual learning of the material, as indexed by neural response properties.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

We would like to thank Khadija Rita Khatun for helping with data collection. Ioanna Zioga is supported by a doctoral studentship from the Department of Biological and Experimental Psychology at the School of Biological and Chemical Sciences, Queen Mary University of London. Peter Harrison is supported by a doctoral studentship from the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1). We also thank the anonymous reviewers for their constructive comments towards the improvement of the paper.

References

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

Abler, W. (2002). Just Say a System for Remembering Finger Placement in Various Keys on the Violin. *American String Teacher*, *52*(3), 56–62. doi:10.1177/000313130205200311

Aronoff, F. (1983). Dalcroze Strategies for Music Learning in the Classroom. *International Journal of Music Education*, *os-2*(1), 23–25. doi:10.1177/025576148300200105

Atienza, M., Cantero, J. L., & Dominguez-Marin, E. (2002). The time course of neural changes underlying auditory perceptual learning. *Learning & Memory*, *9*(3), 138–150.

Begleiter, R., El-Yaniv, R., & Yona, G. (2004). On Prediction Using Variable Order Markov

Models. *Journal of Artificial Intelligence Research*, *22*, 385–421.
Bjork, E., & Bjork, R. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, *2*, 59–68.
Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: The MIT Press.
Bosnyak, D., Eaton, R., & Roberts, L. (2004). Distributed auditory cortical representations are modified when non-musicians are trained at pitch discrimination with 40 Hz amplitude modulated tones. *Cerebral Cortex*, *14*(10), 1088–1099.
Bowles, C. (1998). Music activity preferences of elementary students. *Journal of Research in Music Education*, *46*(2), 193–207. doi:10.2307/3345623
Brainard, D., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.
Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science*, *32*(1/2), 115–132.
Carrus, E., Pearce, M. T., & Bhattacharya, J. (2013). Melodic pitch expectation interacts with neural responses to syntactic but not semantic violations. *Cortex*, *49*(8), 2186–2200.
Cleary, M., Pisoni, D. B., & Geers, A. E. (2001). Some measures of verbal and spatial working memory in eight-and nine-year-old hearing-impaired children with cochlear implants. *Ear and Hearing*, *22*(5), 395.
Colonius, H., & Diederich, A. (2006). The race model inequality: Interpreting a geometric measure of the amount of violation. *Psychological Review*, *113*(1), 148.
Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science*, *17*(10), 905–912.
Conway, C., & Christiansen, M. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24.
Coull, J., Tremblay, L., & Elliott, D. (2001). Examining the specificity of practice hypothesis: Is learning modality specific? *Research Quarterly for Exercise and Sport*, *72*(4), 345–354. doi:10.1080/02701367.2001.10608971
Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.
Daikoku, T., Yatomi, Y., & Yumoto, M. (2015). Statistical learning of music-and language-like sequences and tolerance for spectral shifts. *Neurobiology of Learning and Memory*, *118*, 8–19.
Debener, S., Makeig, S., Delorme, A., & Engel, A. K. (2005). What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis. *Cognitive Brain Research*, *22*(3), 309–321.
Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.
Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the (): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115.
Duzcu, H., Özkurt, T., Mapelli, I., & Hohenberger, A. (2019). N1-P2: Neural markers of temporal expectation and response discrimination in interval timing. *Acta Neurobiol Exp*, *79*, 193–204.
Egermann, H., Pearce, M. T., Wiggins, G. A., & McAdams, S. (2013). Probabilistic models

of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(3), 533–553.
Emberson, L., Conway, C., & Christiansen, M. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *The Quarterly Journal of Experimental Psychology*, *64*(5), 1021–1040. doi:10.1080/17470218.2010.538972
Ferdinand, N. K., Mecklinger, A., & Kray, J. (2008). Error and deviance processing in implicit and explicit sequence learning. *Journal of Cognitive Neuroscience*, *20*(4), 629–642. doi:10.1162/jocn.2008.20046
Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. doi:10.1098/rstb.2005.1622
Frost, R., Armstrong, B., Siegelman, N., & Christiansen, M. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125.
Gathercole, S., & Alloway, T. (2008). *Working memory and learning: A practical guide for teachers*. Sage.
Goodin, D., Squires, K., & Starr, A. (1983). Variations in early and late event-related components of the auditory evoked potential with task difficulty. *Electroencephalography and Clinical Neurophysiology*, *55*(6), 680–686.
Hajihosseini, A., & Holroyd, C. B. (2013). Frontal midline theta and N200 amplitude reflect complementary information about expectancy and outcome evaluation. *Psychophysiology*, *50*(6), 550–562. doi:10.1111/psyp.12040
Halpern, A. R., Zioga, I., Shankleman, M., Lindsen, J., Pearce, M. T., & Bhattarcharya, J. (2017). That note sounds wrong! Age-related effects in processing of musical expectation. *Brain and Cognition*, *113*, 1–9.
Halpern, A., Zioga, I., Shankleman, M., Lindsen, J., Pearce, M. T., & Bhattacharya, J. (2017). That note sounds wrong! Age-related effects in processing of musical expectation. *Brain and Cognition*, *113*, 1–9.
Hansen, N. C., & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology*, *5*, 1052.
Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 484.
Hoefer, M., Tyll, S., Kanowski, M., Brosch, M., Schoenfeld, M., Heinze, H., & Noesselt, T. (2013). Tactile stimulation and hemispheric asymmetries modulate auditory perception and neural responses in primary auditory cortex. *Neuroimage*, *79*, 371–382.
Hoffman, J. E. (1990). Event-related potentials and automatic and controlled processes. In Rohrbaugh J W, Parasuraman R, & Johnson R Jr (Eds.), *Event Related Brain Potentials* (pp. 145–157). New York: Oxford University Press.
Huron, D. B. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press.
Jonaitis, E. M., & Saffran, J. R. (2009). Learning harmony: The role of serial statistics. *Cognitive Science*, *33*(5), 951–968.
Joyce, C., & Rossion, B. (2005). The face-sensitive N170 and VPP components manifest the same brain processes: the effect of reference electrode site. *Clinical Neurophysiology*, *116*(11), 2613–2631.
Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *The Behavioral and Brain Sciences*, *31*, 559–575; discussion

575-621. doi:10.1017/S0140525X08006079
Kantak, S., & Winstein, C. (2012). Learning–performance distinction and memory processes for motor skills: A focused review and perspective. *Behavioural Brain Research*, *228*(1), 219–231.
Kim, S., Kim, J., & Chung, C. (2011). The effect of conditional probability of chord progression on brain response: an MEG study. *PloS One*, *6*(2).
Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42.
Koelsch, S, Busch, T., Jentschke, S., & Rohrmeier, M. (2016). Under the hood of statistical learning: A statistical MMN reflects the magnitude of transitional probabilities in auditory sequences. *Scientific Reports*, *6*, 19741.
Koelsch, S, Gunter, T., Friederici, A., & Schröger, E. (2000). Brain indices of music processing:"nonmusicians" are musical. *Journal of Cognitive Neuroscience*, *12*(3), 520–541. doi:10.1162/089892900562183
Koelsch, S, Kilches, S., Steinbeis, N., & Schelinski, S. (2008). Effects of unexpected chords and of performer's expression on brain responses and electrodermal activity. *PLoS One*, *3*(7).
Koelsch, Stefan, Busch, T., Jentschke, S., & Rohrmeier, M. (2016). Under the hood of statistical learning: A statistical MMN reflects the magnitude of transitional probabilities in auditory sequences. *Scientific Reports*, *6*(1), 19741. doi:10.1038/srep19741
Koelsch, Stefan, & Jentschke, S. (2010). Differences in electric brain responses to melodies and chords. *Journal of Cognitive Neuroscience*, *22*(10), 2251–2262.
Kopp, B., & Wolff, M. (2000). Brain mechanisms of selective learning: Event-related potentials provide evidence for error-driven learning in humans. *Biological Psychology*, *51*(2–3), 223–246.
Kuchenbuch, A., Paraskevopoulos, E., & Herholz, S. C. (2014). Audio-tactile integration and the influence of musical training. *PloS One*, *9*(1).
Lee, T., & Genovese, E. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport*, *59*(4), 277–287. doi:10.1080/02701367.1988.10609373
Leino, S., Brattico, E., Tervaniemi, M., & Vuust, P. (2007). Representation of harmony rules in the human brain: Further evidence from event-related potentials. *Brain Research*, *1142*, 169–177.
Lieberman, M. D., Chang, G. Y., Chiao, J., Bookheimer, S. Y., & Knowlton, B. J. (2004). An event-related fMRI study of artificial grammar learning in a balanced chunk strength design. *Journal of Cognitive Neuroscience*, *16*(3), 427–438.
Loui, P. (2012). Learning and liking of melody and harmony: Further studies in artificial grammar learning. *Topics in Cognitive Science*, *4*(4), 554–567.
Loui, P., Grent, T., Torpey, D., & Woldorff, M. (2005). Effects of attention on the neural processing of harmonic syntax in Western music. *Cognitive Brain Research*, *25*(3), 678–687.
Loui, P., & Wessel, D. L. (2008). Learning and liking an artificial musical system: Effects of set size and repeated exposure. *Musicae Scientiae: The Journal of the European Society for the Cognitive Sciences of Music*, *12*(2), 207.
Loui, P., Wessel, D. L., & Kam, C. L. H. (2010). Humans rapidly learn grammatical structure in a new musical scale. *Music Perception: An Interdisciplinary Journal*, *27*(5), 377–388.
Luft, C. D. B., Baker, R., Goldstone, A., Zhang, Y., & Kourtzi, Z. (2016). Learning temporal statistics for sensory predictions in aging. *Journal of Cognitive Neuroscience*, *28*(3), 418–432.

Luft, C. D. B., Meeson, A., Welchman, A., & Kourtzi, Z. (2015). Decoding the future from past experience: Learning shapes predictions in early visual cortex. *Journal of Neurophysiology*, *113*(9), 3159–3171. doi:10.1152/jn.00753.2014

Maidhof, C., Vavatzanidis, N., Prinz, W., Rieger, M., & Koelsch, S. (2010). Processing expectancy violations during music performance and perception: an ERP study. *Journal of Cognitive Neuroscience*, *22*(10), 2401–2413. doi:10.1162/jocn.2009.21332

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 609.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Meyer, L. (1956). *Emotion and meaning in music*. University of chicago Press.

Michalski, A. (2000). Expectation of an important event affects responses to irrelevant stimuli of different modalities. *Acta Neurobiologiae Experimentalis*, *60*(4), 467–478.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*(1), 138–153.

Moldwin, T., Schwartz, O., & Sussman, E. S. (2017). Statistical learning of melodic patterns influences the brain's response to wrong notes. *Journal of Cognitive Neuroscience*, *29*(12), 2114–2122. doi:10.1162/jocn_a_01181

Molholm, S., Ritter, W., Javitt, D., & Foxe, J. (2004). Multisensory visual–auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex*, *14*(4), 452–465.

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*, Vol. 9.

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*(4), 313–329.

Näätänen, R., & Picton, T. W. (1986). N2 and automatic versus controlled processes. *Electroencephalogr Clin Neurophysiol Suppl*, *38*, 169–186.

Nichols, E., & Grahn, J. (2016). Neural correlates of audiovisual integration in music reading. *Neuropsychologia*, *91*, 199–210.

Oliveira, F. T. P., McDonald, J. J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, *19*(12), 1994–2004. doi:10.1162/jocn.2007.19.12.1994

Omigie, D., Pearce, M. T., Williamson, V. J., & Stewart, L. (2013). Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia*, *51*(9), 1749–1762. doi:10.1016/j.neuropsychologia.2013.05.010

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1.

Pantev, C., Paraskevopoulos, E., Kuchenbuch, A., Lu, Y., & Herholz, S. C. (2015). Musical expertise is related to neuroplastic changes of multisensory nature within the auditory cortex. *European Journal of Neuroscience*, *41*(5), 709–717. doi:10.1111/ejn.12788

Paraskevopoulos, E., Chalas, N., Kartsidis, P., Wollbrink, A., & Bamidis, P. (2018). Statistical learning of multisensory regularities is enhanced in musicians: An MEG study. *NeuroImage*, *175*, 150–160.

Paraskevopoulos, E., Kraneburg, A., Herholz, S., Bamidis, P., & Pantev, C. (2015). Musical expertise is related to altered functional connectivity during audiovisual integration.

Proceedings of the National Academy of Sciences, *112*(40), 12522–12527.
Paraskevopoulos, E., Kuchenbuch, A., Herholz, S. C., & Pantev, C. (2014). Multisensory integration during short-term music reading training enhances both uni-and multisensory cortical processing. *Journal of Cognitive Neuroscience*, *26*(10), 2224–2238. doi:10.1162/jocn_a_00620
Paraskevopoulos, E., Kuchenbuch, A., Herholz, S., Foroglou, N., Bamidis, P., & Pantev, C. (2014). Tones and numbers: A combined EEG-MEG study on the effects of musical expertise in magnitude comparisons of audiovisual stimuli. *Human Brain Mapping*, *35*(11), 5389–5400. doi:10.1002/hbm.22558
Paraskevopoulos, E., Kuchenbuch, A., Herholz, S., & Pantev, C. (2012a). Musical expertise induces audiovisual integration of abstract congruency rules. *Journal of Neuroscience*, *32*(50), 18196–18203.
Paraskevopoulos, E., Kuchenbuch, A., Herholz, S., & Pantev, C. (2012b). Statistical learning effects in musicians and non-musicians: An MEG study. *Neuropsychologia*, *50*(2), 341–349.
Pearce, M., & Rohrmeier, M. (2018). *Musical syntax II: Empirical perspectives*. Berlin, Heidelberg: Springer.
Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition (Doctoral dissertation, City University London)*.
Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*(1), 378–395. doi:10.1111/nyas.13654
Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, *39*(10), 1365–1389.
Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*(1), 302–313.
Perfetti, C. A., & Wang, M. (2006). Visual analysis and lexical access of Chinese characters by Chinese as second language readers. *Language and Linguistics*, *7*(3), 637–657.
Pisoni, D. B., & Cleary, M. (2004). *Learning, Memory, and Cognitive Processes in Deaf Children Following Cochlear Implantation*. Springer, New York, NY.
Polich, J., Ellerson, P., & Cohen, J. (1996). P300, stimulus intensity, modality, and probability. *International Journal of Psychophysiology*, *23*(1–2), 55–62.
Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin*, *133*(2), 227–244.
Proverbio, A. M., Leoni, G., & Zani, A. (2004). Language switching mechanisms in simultaneous interpreters: An ERP study. *Neuropsychologia*, *42*(12), 1636–1656.
Rapp, B., & Hendel, S. (2003). Principles of cross-modal competition: Evidence from deficits of attention. *Psychonomic Bulletin & Review*, *10*(1), 210–219. doi:10.3758/BF03196487
Reber, A. S. (1993). *Implicit learning and knowledge: An essay on the cognitive unconscious*. New York, NY: Oxford University Press.
Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*(2), 123.
Reinke, K., He, Y., Wang, C., & Alain, C. (2003). Perceptual learning modulates sensory evoked response during vowel segregation. *Cognitive Brain Research*, *17*(3), 781–791.
Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, 1850–1855.

Robinson, C., Chandra, M., & Sinnett, S. (2016). Existence of competing modality dominances. *Attention, Perception, & Psychophysics*, *78*(4), 1104–1114. doi:10.3758/s13414-016-1061-3

Robinson, C., & Sloutsky, V. (2013). When audition dominates vision. *Experimental Psychology*. doi:10.1027/1618-3169/a000177

Rohrmeier, M. A., & Cross, I. (2014). Modelling unsupervised online-learning of artificial grammars: Linking implicit and statistical learning. *Consciousness and Cognition*, *27*, 155–167.

Rohrmeier, M. A., Rebuschat, P., & Cross, I. (2011). Incidental and online learning of melodic structure. *Consciousness and Cognition*, *20*(2), 214–222.

Rohrmeier, M., & Rebuschat, P. (2012). Implicit learning and acquisition of music. *Topics in Cognitive Science*, *4*(4), 525–553.

Rugg, M., & Coles, M. (1995). *Electrophysiology of mind: Event-related brain potentials and cognition.* Oxford University Press.

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

Saffran, J. R., Reeck, K., Niebuhr, A., & Wilson, D. (2005). Changing the tune: The structure of the input affects infants' use of absolute and relative pitch. In *Developmental Science* (Vol. 8).

Saffran, J R, Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

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

Schmidt, R., & Bjork, R. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–218. doi:10.1111/j.1467-9280.1992.tb00029.x

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology*, *16*(14), 1422–1427. doi:10.1016/J.CUB.2006.05.048

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417. doi:10.1016/J.TICS.2008.07.006

Simpson, M. (2015). U.S. Patent Application No. 14/341.

Sinnett, S., Soto-Faraco, S., & Spence, C. (2008). The co-occurrence of multisensory competition and facilitation. *Acta Psychologica*, *128*(1), 153–161.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: The Colavita effect revisited. *Perception & Psychophysics*, *69*(5), 673–686.

Soderstrom, N., & Bjork, R. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199. doi:10.1177/1745691615569000

Spence, C. (2009). Explaining the Colavita visual dominance effect. *Progress in Brain Research*, *176*, 245–258.

St Clair-Thompson, H., Stevens, R., Hunt, A., & Bolder, E. (2010). Improving children's working memory and classroom performance. *Educational Psychology*, *30*(2), 203–219. doi:10.1080/01443410903509259

Stanford, T., & Stein, B. (2007). Superadditivity in multisensory integration: putting the computation in context. *Neuroreport*, *18*(8), 787–792.

Steinbeis, N., & Koelsch, S. (2008). Shared neural resources between music and language indicate semantic processing of musical tension-resolution patterns. *Cerebral Cortex*, *18*(5), 1169–1178.

Steinbeis, N., Koelsch, S., & Sloboda, J. (2006). The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses.

Journal of Cognitive Neuroscience, *18*(8), 1380–1393. doi:10.1162/jocn.2006.18.8.1380
Stuellein, N., Radach, R. R., Jacobs, A. M., & Hofmann, M. J. (2016). No one way ticket from orthography to semantics in recognition memory: N400 and P200 effects of associations. *Brain Research*, *1639*, 88–98. doi:10.1016/J.BRAINRES.2016.02.029
Tierney, A., Bergeson-Dana, T., & Pisoni, D. (2008). Effects of early musical experience on auditory sequence memory. *Empirical Musicology Review: EMR*, *3*(4), 178.
Tillmann, B., & McAdams, S. (2004). Implicit learning of musical timbre sequences: statistical regularities confronted with acoustical (dis) similarities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1131–1142.
Tremblay, K, Kraus, N., McGee, T., Ponton, C., & Otis, B. (2001). Central auditory plasticity: changes in the N1-P2 complex after speech-sound training. *Ear and Hearing*, *22*(2), 79–90.
Tremblay, KL, & Kraus, N. (2002). Auditory training induces asymmetrical changes in cortical neural activity. *Journal of Speech, Language, and Hearing Research*, *45*(3), 564–572.
Tsogli, V., Jentschke, S., Daikoku, T., & Koelsch, S. (2019). When the statistical MMN meets the physical MMN. *Scientific Reports*, *9*.
Tunney, R., & Altmann, C. (1999). The transfer effect in artificial grammar learning: Reappraising the evidence on the transfer of sequential dependencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1322.
Villarreal, E. A. G., Brattico, E., Leino, S., & Østergaard, L. (2011). Distinct neural responses to chord violations: a multiple source analysis study. *Brain Research*, *1389*, 103–114.
Wechsler, D. (1955). *Manual for the Wechsler adult intelligence scale.* Oxford, England: Psychological Corp.
Zioga, I., Harrison, P. M., Pearce, M. T., Bhattacharya, J., & Luft, C. D. B. (2020). From learning to creativity: Identifying the behavioural and neural correlates of learning to predict human judgements of musical creativity. *NeuroImage*, *206*, 116311.