*Meta-Psychology*, 2019, vol 3, MP.2018.1481,
https://doi.org/10.15626/MP.2018.1481
Article type: Original Article
Published under the CC-BY4.0 license

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: Yes

Edited by: Henrik Danielsson
Reviewed by: Stefan Kölsch and Carine Signoret
Analysis reproduced by: Martina Sladekova
All supplementary files can be accessed at the OSF project page: https://doi.org/10.17605/OSF.IO/EPD42

# Similar event-related potentials to music and language: A replication of Patel, Gibson, Ratner, Besson, & Holcomb (1998)

Joshua R. de Leeuw, Jan Andrews, Zariah Altman, Rebecca Andrews, Robert Appleby, James L. Bonanno, Isabella DeStefano, Eileen Doyle-Samay, Ayela Faruqui, Christina M. Griesmer, Jackie Hwang, Kate Lawson, Rena A. Lee, Yunfei Liang, John Mernacaj, Henry J. Molina, Hui Xin Ng, Steven Park, Thomas Possidente, Anne Shriver

Vassar College

We report a replication of Patel, Gibson, Ratner, Besson, and Holcomb (1998). The results of our replication are largely consistent with the conclusions of the original study. We found evidence of a P600 component of the event-related potential (ERP) in response to syntactic violations in language and harmonic inconsistencies in music. There were some minor differences in the spatial distribution of the P600 on the scalp between the replication and the original. The experiment was pre-registered at https://osf.io/g3b5j/. We conducted this experiment as part of an undergraduate cognitive science research methods class at Vassar College; we discuss the practice of integrating replication work into research methods courses.

*Keywords*: EEG, ERP, P600, language, music, replication.

Patel, Gibson, Ratner, Besson, and Holcomb (1998) found that violations of expected syntactic structure in language and violations of expected harmonic structure in music both elicit the P600 component of the event-related potential (ERP). The P600 is a positive ERP component that occurs approximately 600 ms after stimulus onset. While previous work had established a link between the P600 component and syntactic violations in language (Osterhout & Holcomb, 1992, 1993; Osterhout, Holcomb, & Swinney, 1994), Patel and colleagues were the first to report a direct comparison of the P600 for violations of musical and linguistic structure, finding that the amplitude and scalp distribution of the P600 was similar for linguistic and musical violations.

This result has been influential in theorizing about the relationship between music and language, with more than 700 citations twenty years after publication (Google Scholar search, September 2018). It has been used as evidence for the "shared syntactic integration resource hypothesis," a theory that posits that structural processing of music and language utilizes the same cognitive and neural resources (Patel, 2003). It has also been used to argue more broadly for the shared neurological basis of music and language (e.g., Abrams et al., 2011; Besson & Schön, 2001; Herdener et al., 2014; Merrill et al., 2012; Patel, 2010; Sammler et al., 2010, 2013), and for the existence of shared cognitive resources/constraints for processing music and language (e.g., Besson, Chobert, & Marie, 2011; Chobert, François,

*Acknowledgements*: We are extremely grateful to Debra Ratchford and Prashit Parikh for their assistance supervising EEG data collection, and Polyphony Bruna for assisting with data organization and stimulus measurement.

Velay, & Besson, 2014; Christiansen & Chater, 2008; Lima & Castro, 2011; Moreno et al., 2009; Thompson, Schellenberg, & Husain, 2004; Tillmann, 2012).

Though the work has been influential, we are not aware of any published direct replications of the main result. Several studies have found ERP correlates of structural violations in music (Besson & Faïta, 1995; Besson, Faïta, & Requin, 1994; Janata, 1995), though there is variation in the kinds of components that are found (Featherstone, Morrison, Waterman, & MacGregor, 2013; Featherstone, Waterman, & Morrison, 2012). Other studies have found that ERP markers of violations of linguistic structure are systematically affected by the presence or absence of simultaneous structural violations in music (Koelsch, Gunter, Wittfoth, & Sammler, 2005; Steinbeis & Koelsch, 2008). These findings, along with many other behavioral and non-ERP neural measures (see Koelsch, 2011 for a review), support the general conclusion of Patel et al. (1998) that there is overlap between the processing of structural violations in music and language. While this converging evidence should bolster our belief in the results, there is no substitute for a direct replication given the well-documented problem of publication bias in the literature (e.g., Ingre & Nilsonne, 2018; Rosenthal, 1979).

This experiment was part of an undergraduate research methods course in cognitive science, which 2 of us co-taught, 17 of us were enrolled in, and 1 of us was serving as a course intern. A major focus of this course was exposure to and training in practices that have developed in response to the replication crisis, including an increased emphasis on direct replications (Zwaan, Etz, Lucas, & Donnellan, 2017), pre-registration of experiments (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), and transparency through public sharing of materials, data, and analysis scripts (Nosek et al., 2015). To gain hands-on experience with these practices, the class conducted this replication study. We chose to replicate Patel et al. (1998) given its theoretical significance in the field, lack of prior direct replications, and practical considerations like the complexity of the data analysis and study design.

Our replication is what LeBel et al. (2017) would call a very close replication. While we were able to operationalize the independent and dependent variables in the same manner as Patel et al. and were able to use either the same exact (music) or close replicas (language) of the original stimuli, we did make some changes to their procedure. We removed two conditions (out of six) to shorten the overall length of the experiment, which was necessary to run the experiment in a classroom environment. We also focused the analysis on what we took to be the key findings of the original. We highlight these deviations from the original throughout the methods section below. Very close replications like this one are efforts to establish the "basic existence" of phenomena (LeBel et al., 2017), which is an essential step for creating a set of robust empirical facts for theory development.

## Method

All stimuli, experiment scripts, data, and analysis scripts are available on the Open Science Framework at https://osf.io/zpm9t/. The study pre-registration is available at https://osf.io/g3b5j/. All participants provided informed consent and this study was approved by the Vassar College Institutional Review Board.

### Overview

In both the original experiment and our replication, participants listened to short sentences and musical excerpts and made judgments about whether the sentence/music was acceptable or unacceptable. ERPs in response to particular words or musical events were measured with EEG.

In the original experiment there were three critical kinds of sentences (grammatically simple, grammatically complex, and ungrammatical) and three critical kinds of musical excerpts (in key, nearby out of key, and distant out of key). The P600 is measured by comparing the amplitude of the ERP in the grammatically simple condition to the other two language conditions and the in-key condition to the other two music conditions (see Results, below).

Due to logistical constraints of lab availability, time, and class schedule, we opted to restrict the replication to two kinds of sentences and two kinds of musical excerpts. We used only the grammatically simple and ungrammatical sentences for the language stimuli (plus their associated control stimuli, see Stimuli below), and only the in-key and distant out-of-key musical excerpts. We believe that this choice is justifiable, as the theoretical claims of Patel et al. are most strongly based on the P600 that was found in the ungrammatical and distant out-of-key conditions, as these are the stronger contrasts (i.e., they are more "syntactically wrong"). The grammatical and in-key conditions serve as the baseline for these analyses, and so must also be included. The original also included unanalyzed filler sentences and musical excerpts to balance certain (possibly confounding) properties of the stimuli; by not including many of the original stimuli these properties were more balanced in the critical stimuli, and we were able to drop all of the fillers in the music condition and 20 of the fillers in the language condition. Altogether, the original experiment contained 150 sentences (3 x 30 plus 60 fillers) and 144 musical excerpts (3 x 36 plus 36 fillers), and our replication contained 100 sentences (2 x 30 plus 40 fillers) and 72 musical excerpts (2 x 36).

## Participants

44 Vassar College students, ages 18-22 (M = 19.8 years, SD = 1.2 years), participated in the study. Our pre-registered target was 40, which is slightly more than 2.5 times the original sample (N=15). We aimed for at least 2.5 times the original sample based on the heuristic provided by Simonsohn (2015). The goal of the heuristic is for replications to have sufficient power to detect effects that are smaller than the original but still plausibly detectable by the original study. While we ran more participants than the original target of 40, 5 participants did not complete the experiment due to technical difficulties such as recording problems with the EEG equipment. Thus, we ended up with 39 participants, one under our pre-registered target. We stopped data collection because we reached our pre-registered cutoff date of 2/24/18 prior to having 40 usable recordings. The cutoff date was necessary for the schedule of the class.

Participants in Patel et al. (1998) were musically trained but specifically did not have perfect pitch. Their participants had an average of 11 years of musical experience, had studied music theory, and played a musical instrument for an average of 6.2 hours per week. All of our participants had at least 4 years of prior musical experience (M = 9.7 years, SD = 3.3 years), which we defined as participation in music lessons, enrollment in music coursework, or experience with any musical instrument (including voice). We also required that participants not have perfect pitch (by self-report). We did not require that participants had studied music theory. Our participants played a musical instrument for an average of 5.8 hours per week (SD = 3.4 hours per week).

## Stimuli

Patel graciously provided the music stimuli used in the original study. The language stimuli were no longer available in audio form, but we were provided with a list of the text of the original stimuli. We refer the reader to Patel et al. (1998) for the full details of the stimuli. Here we describe a basic overview of the format to provide enough context for understanding the experiment, as well as our process for recording the audio stimuli.

The music stimuli were short sequences of chords synthesized using a generic piano MIDI instrument. They were about 6 seconds long. The chords initially established a harmonic key. The target chord — either the root chord of the established key (in-key condition) or the root chord of a distantly-related harmonic key (distant out-of-key condition) — occurred in the second half of the excerpt. An example in-key sequence can be heard at https://osf.io/z6vcu/. An example out-of-key sequence can be heard at https://osf.io/wde67/. To simplify condition labeling in what follows, the in-key (harmonically congruous) musical stimuli will be called grammatical and the distant out-of-key (harmonically incongruous) musical stimuli will be called ungrammatical, even though we recognize that the application of those terms to music is not necessarily as straightforward as it is for language.

The language stimuli were spoken sentences with a target noun phrase that was either grammatical or ungrammatical given the prior context. There were two primary types of sentences (grammatical and ungrammatical) as well as two kinds of filler sentences, designed to prevent listeners from using cues other than the target noun phrase in context to judge the acceptability of the sentence. The grammatical but unacceptable fillers make it so that not all instances of "had" are acceptable. The grammatical fillers make it so that not all instances of verb + "the" are unacceptable. Examples of each sentence type are below (the target noun phrase is italicized):

Grammatical: Some of the soldiers had discovered *a new strategy* for survival.
Ungrammatical: Some of the marines pursued the discovered *a new strategy* for survival.
Grammatical, unacceptable (filler): Some of the lieutenants had reimbursed *a new strategy* for survival.
Grammatical (filler): Some of the explorers pursued the idea of *a new strategy* for survival.

Sentences ranged from 2.9 to 4.8 seconds long, spoken by one of the female experimenters at a rate of approximately six syllables per second using a Blue Snowball iCE Condenser microphone, sampled at 44.1kHz. The audio files were later amplified in Audacity in order to be at a volume similar across sentences and approximately comparable to that of the music stimuli. For each file, the onset and duration of the target noun phrase was recorded (in milliseconds) to refer to in analysis when identifying the onset of ERP components (see https://osf.io/tr7mq/ for complete list).

In addition to the music and language stimuli used in the original experiment, we created sample stimuli to provide a short pre-task tutorial for participants. These consisted of six new sentences and six new musical excerpts, designed to match the properties of the original stimuli. The music files were created in MuseScore (MuseScore Development Team, 2018).

## Procedure

Participants completed the experiment in a quiet room seated at a computer screen and keyboard. Audio files were played through a pair of speakers (the original study used headphones). The experiment was built using the jsPsych library (de Leeuw, 2015). Communication between jsPsych and the EEG recording equipment was managed through a Chrome extension that enables JavaScript-based control over a parallel port (Rivas, 2016).

Each trial began with the audio file playing while a fixation cross was present on the screen. Participants were asked to avoid blinking or moving their eyes while the fixation cross was present, to prevent eye movement artifacts in the EEG data. After the audio file concluded, participants saw a blank screen for 1450 ms. Finally, a text prompt appeared on the screen asking participants if the sentence or musical excerpt was acceptable or unacceptable. Participants pressed either the A (acceptable) or U (unacceptable) key in response. This procedure is nearly identical to the original, except for the use of a keyboard instead of a response box held in the participant's lap.

The experiment started with a short set of practice trials: 6 language trials followed by 6 music trials. Following the practice trials, the experimenter verified that the participant understood the instructions before the experiment proceeded.

The experiment consisted of 5 blocks: 3 language blocks containing 33, 33, and 34 trials, and 2 music blocks containing 36 trials each. The experiment always started with a language block and then alternated between language and music. Grammatical and ungrammatical trials were randomly intermixed within each block. At the conclusion of a block, participants were given the opportunity to take a break. Participants controlled the length of the break.

## ERP Recording

We recorded EEG activity using a 128-channel sensor net (Electrical Geodesics Inc.) at a sampling rate of 1000 samples/s referenced to Cz. The data were amplified using a Net Amps 400 Amplifier (Electrical Geodesics Inc.). We focused on the 13 scalp locations that were used in Patel et al. (1998). The locations and their corresponding electrode number on the EGI-128 system were Fz (11), Cz (129), and Pz (62) (midline sites), and F8 (122), ATR (115), TR (108), WR (93), O2 (83), F7 (33), ATL (39), TL (45), WL (42), and O1 (70) (lateral sites). Vertical eye movements and blinks were monitored by means of two electrodes located above and one located below each eye; horizontal eye movements were monitored by means of one electrode located to the outer side of each eye. Impedances for all of these electrodes were kept below 50 kΩ prior to data collection.

Netstation 5.4 waveform tools were used to process the EEG data offline, first by applying a high pass filter at 0.1 Hz and a low pass filter at 30 Hz. Data were segmented into 1100 ms segments starting 100 ms prior to and ending 1000 ms after target stimulus onset. Segments containing ocular artifacts were excluded from further analyses, as were any segments that had more than 20 bad channels. The

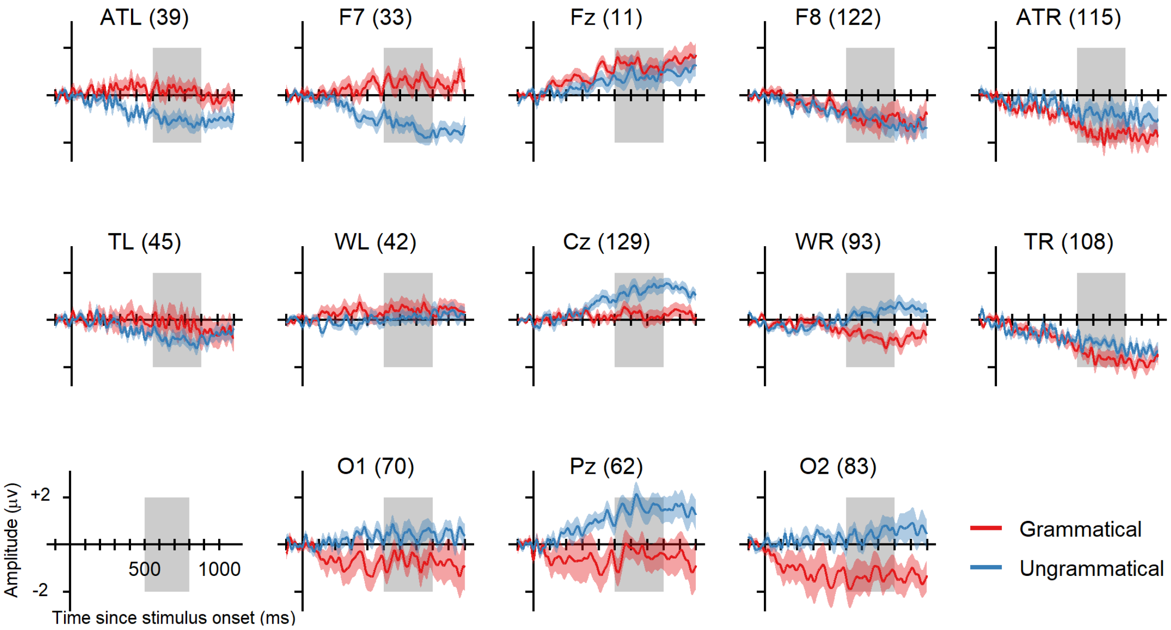

Figure 1. Grand average waveforms for language stimuli. The shaded box highlights the time window for analyzing P600 differences (500-800ms after stimulus onset) and the area surrounding each line represents ±1 SE. The plots are arranged to represent approximate scalp position of each electrode, with posterior electrodes at the bottom.

NetStation bad channel replacement tool was applied to the EEG data which were re-referenced using an average reference and baseline corrected to the 100 ms prior to stimulus onset. These processing steps are similar to those used by Patel et al. (1998; see pgs. 729-730), with some minor differences due to the use of a different EEG system. Information about all tool settings is available at https://osf.io/96bjn/.

## Results

We conducted our analyses in R v3.4.2 (R Core Team, 2017) using several packages (Henry & Wickham, 2017; Lawrence, 2016; Morey & Rouder, 2015; Wickham, 2016; Wickham, Francois, Henry, & Müller, 2017; Wickham & Henry, 2018; Wickham, Hester, & Francois, 2017; Wilke, 2017). The complete annotated analysis script is available as an R Notebook at https://osf.io/m9kej/.

### Data Exclusions

39 participants had a complete data set. We pre-registered a plan to exclude trials that contained artifacts, but we did not pre-register a decision rule for how many good ERP segments a participant would need in each condition to be included in the analysis. To avoid making a biased decision, we tabulated the number of artifact-free segments for each of the four conditions for each participant and chose a cutoff as the very first step in our analysis, prior to any examination of the waveforms. Based on this ad-hoc inspection of the data (see https://osf.io/w7hrm/), we decided to exclude 4 participants who had at least one condition with fewer than 19 good segments. We chose 19 as the cutoff because the data had some natural clustering; the 4 participants who did not meet that cutoff had 15 or fewer good segments in at least one condition. This left us with data from 35 participants. All subsequent analyses are based only on these 35 participants. The mean number of usable trials across participants was 27.7 for language-grammatical and

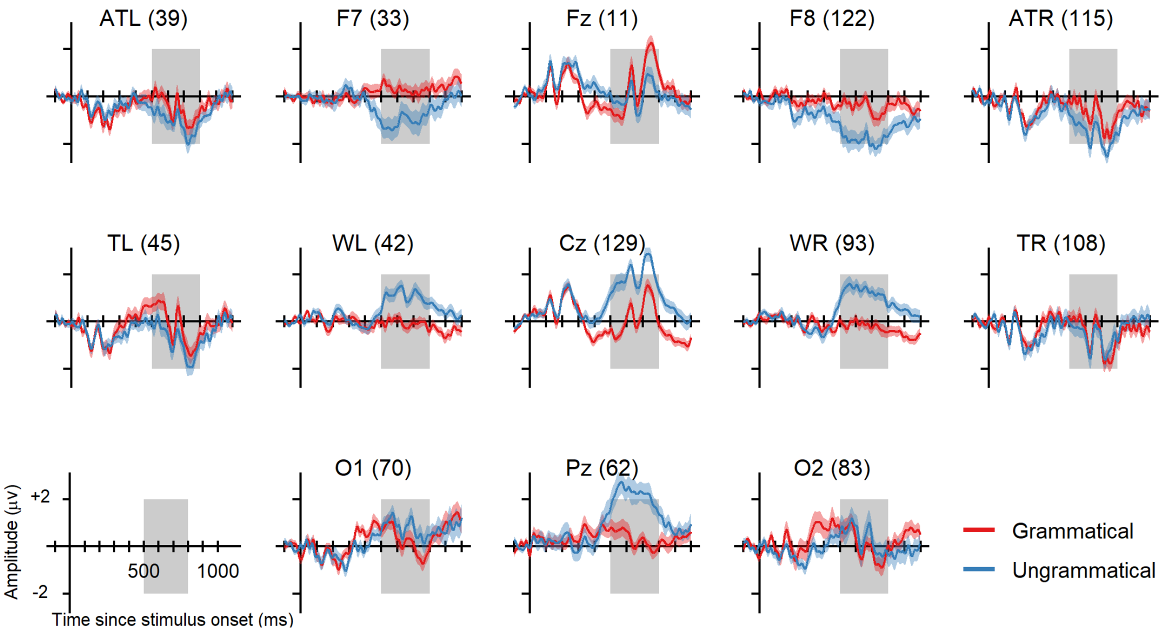

Figure 2. Grand average waveforms for music stimuli. The shaded box highlights the time window for analyzing P600 differences (500-800ms after stimulus onset) and the area surrounding each line represents ± 1 SE. The plots are arranged to represent approximate scalp position of each electrode, with posterior electrodes at the bottom.

language-ungrammatical, 33.3 for music-grammatical, and 33.0 for music-ungrammatical.

Table 1.
*Behavioral data. Accuracy in participants' judgements of whether stimuli were acceptable or unacceptable in our study compared to the Patel et al. (1998) study. Patel et al. did not report SDs.*

| Condition | Patel et al. (1998) | Replication |
|---|---|---|
| Language, Grammatical | M = 95% | M = 93.3%, SD = 5.3% |
| Language, Ungrammatical | M = 96% | M = 88.2%, SD = 18.0% |
| Music, Grammatical | M = 80% | M = 84.5%, SD = 14.5% |
| Music, Ungrammatical | M = 72% | M = 69.1%, SD = 15.5% |

## Behavioral Data

We calculated the accuracy of the acceptable/unacceptable judgments that participants made, and we compare these data with the data from Patel et al. in Table 1. Overall, the accuracy of our participants seems consistent with the accuracies reported in Patel et al. with the largest difference in the ungrammatical language condition.

## EEG Data

In the original experiment, Patel et al. analyzed the EEG data in two primary ways. We repeat and extend these analyses below.

First, they calculated mean amplitude of the waveforms in all conditions (they had six total conditions, but we have four) and then used ANOVAs to model the effects of grammaticality and electrode site on the amplitude of the ERP. They used separate ANOVA models for the language and music conditions and did not treat this as a factor in this part of the analysis. They analyzed three time windows, 300-500 ms, 500-800 ms, and 800-1100 ms, replicating the ANOVAs separately in each time window. Finally, they repeated this analysis separately for

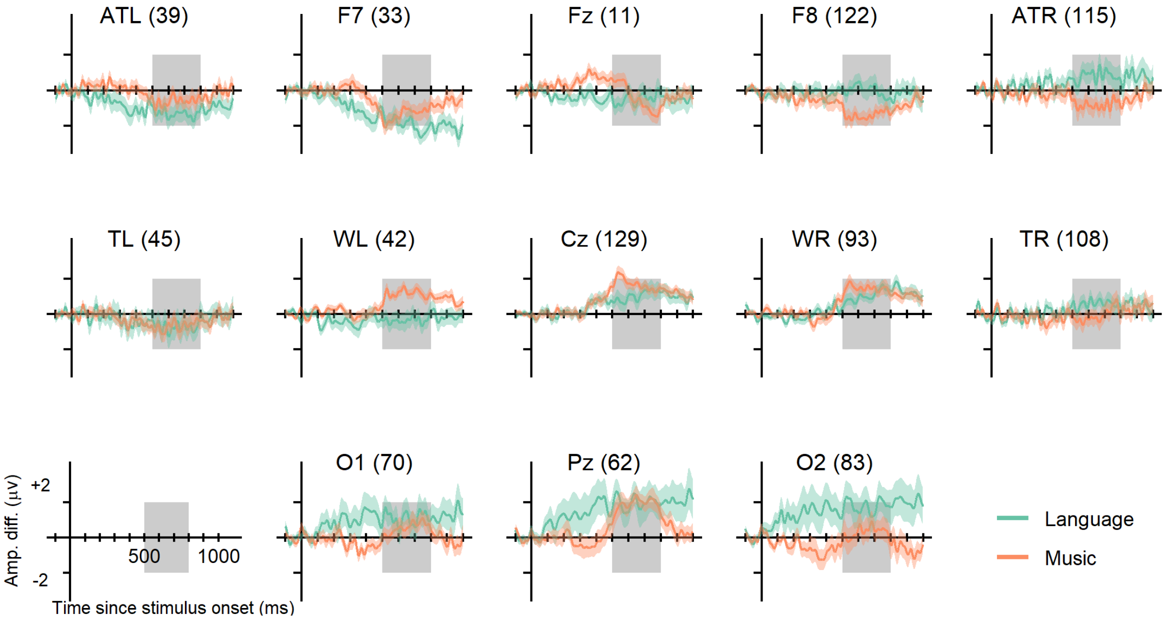

Figure 3. Grand average difference waves (ungrammatical minus grammatical) for language and music. The shaded box highlights the time window for analyzing P600 differences (500-800ms after stimulus onset) and the area surrounding each line represents ± 1 SE.

midline electrodes and lateral electrodes. This was a total of 12 ANOVAs. Given that the P600 should be strongest in the 500-800 ms window, we pre-registered a decision to restrict our analysis to the 500-800 ms window only, reducing the number of ANOVAs to 4. We view this as the strongest test of the original conclusion.

The results of these four ANOVAs are reported in Table 2. While we cannot make direct comparisons with the ANOVA results reported by Patel et al. because we dropped one of the levels of the grammar factor from the procedure, we can look at whether the results align at a high level. For both music and language stimuli, Patel et al. report a significant effect of grammaticality at both midline and lateral electrode sites, as well as a significant interaction between electrode location and grammaticality at both midline and lateral electrode sites. We found most of these effects; the exceptions were that we found no main effect of grammaticality for lateral electrodes and language stimuli, and no main effect of grammaticality for lateral electrodes and music stimuli. However, we did consistently find an interaction between electrode site and grammaticality for all conditions, which makes the differences in main effects somewhat difficult to interpret. For language stimuli, the interaction between electrode site and grammaticality was due to a stronger effect of grammaticality at posterior electrode sites. This is also what Patel et al. found. For music stimuli, the effect of grammaticality was also stronger at posterior sites, with the exception of the two most posterior sites (O1 and O2), where there was no clear effect of grammaticality. This is a difference from the original study, as Patel et al. did observe the music-based P600 effect at these sites.

The second analysis that Patel et al. ran was to calculate difference waves — subtracting the grammatical ERP from the ungrammatical ERP — to, in theory, isolate the P600 and then directly compare the amplitude of the difference waves for language and music stimuli. For an unexplained reason, they shifted the time window of analysis to 450-750 ms. We pre-registered a decision to analyze the difference waves in the 500-800 ms range, to remain consistent with the prior analysis.

Patel et al. found no significant difference in the amplitude of the difference waves and concluded “in the latency range of the P600, the positivities to

Table 2.
ANOVA *results for grammaticality x electrode models. "Electrode" refers to the specific electrode sites within the midline and lateral site groups.*

| Stimulus | Electrode Set | Factor | | |
|---|---|---|---|---|
| | | Grammaticality | Electrode | Grammaticality * Electrode |
| Language | Midline | $F(1, 34) = 6.41$,<br>$p = 0.016$ | $F(2, 68) = 1.00$,<br>$p = 0.372$ | $F(2, 68) = 5.11$,<br>$p = 0.009$ |
| Language | Lateral | $F(1, 34) = 0.44$,<br>$p = 0.512$ | $F(9, 306) = 3.68$,<br>$p = 0.0002$ | $F(9, 306) = 4.99$,<br>$p = 0.000003$ |
| Music | Midline | $F(1, 34) = 23.94$,<br>$p = 0.00002$ | $F(2, 68) = 7.00$,<br>$p = 0.002$ | $F(2, 68) = 12.43$,<br>$p = 0.00003$ |
| Music | Lateral | $F(1, 34) = 1.12$,<br>$p = 0.298$ | $F(9, 306) = 11.10$,<br>$p < 0.000001$ | $F(9,306) = 6.47$,<br>$p < 0.000001$ |

structurally incongruous elements in language and music do not appear to be distinguishable" (pg. 726). We note that a failure to find a statistically significant difference is not necessarily indicative of equivalence (Gallistel, 2009; Lakens, 2017). We repeat this analysis for the sake of comparison, but we also include an analysis using Bayes factors to examine how the relative probabilities of models that do and do not include the factor of stimulus type (language v. music) are affected by these data.

The results of the 2 ANOVAs are shown in Table 3. Like Patel et al., we found no main effect of stimulus type (language v. music) in either lateral or midline electrodes. However, we did find a significant interaction between stimulus type and electrode site for lateral electrodes, though we note that the p-value is relatively high (p = 0.038) with no correction for multiple comparisons.

We conducted the Bayes factor analysis using the BayesFactor R package (Morey & Rouder, 2015). Briefly, the analysis evaluates the relative support for five different models of the data. All models contain a random effect of participant; models 2-5 also contain one or more fixed effects. Model 2 contains the fixed effect of electrode; model 3 contains the fixed effect of stimulus type; model 4 contains both fixed effects; and model 5 contains both fixed effects

Table 3.
ANOVA *results for the difference waves.*

| Electrode Set | Factor | | |
|---|---|---|---|
| | Stimulus | Electrode | Stimulus * Electrode |
| Midline | $F(1, 34) = 0.226$,<br>$p = 0.637$ | $F(2, 68) = 16.289$,<br>$p = 0.000002$ | $F(2, 68) = 0.315$,<br>$p = 0.731$ |
| Lateral | $F(1, 34) = 1.784$,<br>$p = 0.190$ | $F(9, 306) = 12.181$,<br>$p < 0.000001$ | $F(9, 306) = 2.009$,<br>$p = 0.038$ |

*Note.* "Stimulus" refers to language v. music.

Table 4.
*Bayes Factors for models of the effect of electrode site and stimulus type (language v. music) at midline and lateral electrodes.*

| Model | Bayes factor relative to Participant-only model | |
|---|---|---|
| | Midline Electrodes | Lateral Electrodes |
| Electrode + Participant | 18,160 ±1.26% | 1.81x$10^9$ ±0.35% |
| Stimulus + Participant | 0.170 ±2.03% | 0.157 ±2.03% |
| Electrode + Stimulus + Participant | 3,015 ±1.32% | 2.93x$10^8$ ±1.22% |
| Electrode + Stimulus + Electrode*Stimulus + Participant | 367 ±2.02% | 1.24x$10^9$ ±1.30% |

*Note:* Bayes factors indicate the change in posterior odds for the model relative to the model that contains only the random effect of participant. Bayes factors larger than 1 therefore indicate relative support for the model, with larger Bayes factors representing more support. Bayes factors less than 1 indicate relative support for the participant-only model, with numbers closer to 0 indicating more support.

plus their interaction. In each model, the scaling factor for fixed-effect priors is 0.5, and the scaling factor for random-effect priors is 1.0. See Rouder et al. (2012) for model details.

The Bayes factors for all models are reported in Table 4. For the midline electrodes, the model with the greatest positive change in posterior probability relative to just the random effect of participant was the model that added only the fixed effect of electrode. The Bayes factor in favor of this model relative to the next best model, which added the fixed effect of stimulus type, was 6.02 (ratio of 18,160 to 3,015). Thus, these data should shift our belief in the model that does not contain the stimulus type relative to the model that does by about 6x.

For the lateral electrodes, the model with only a fixed effect of electrode and random effect of participant was also the winning model. However, the evidence against an effect of stimulus type is not as strong here. The Bayes factor in favor of the electrode-only model relative to the full model with both main effects and their interaction is only 1.46. The full model is also favored over the main-effects only model by a Bayes factor of 4.22. These suggest that our relative belief in these models is not shifted much by the data.

## Discussion

Patel et al. (1998) concluded that "... the late positivities elicited by syntactically incongruous words in language and harmonically incongruous chords in music were statistically indistinguishable in amplitude and scalp distribution in the P600 latency range. ... This strongly suggests that whatever process gives rise to the P600 is unlikely to be language-specific" (pg. 726). The results of our replication mostly support this conclusion. We found that the amplitude of the ERP 500-800 ms after stimulus onset was more positive for ungrammatical words and chords than for grammatical words and chords. We also found that the effect of grammaticality is stronger in posterior electrodes, though we do find some minor differences from Patel et al. in the consistency of this effect for lateral electrodes. The data are somewhat inconclusive as to whether there is an effect of stimulus type on the amplitude of the ERP in the P600 window, with (at best) moderate evidence to support the conclusion that there is no difference in mean amplitude. This is despite a sample size (N = 35) that is more than twice the original (N = 15).

One aspect of the data that is visually striking is the clear differences in the shape of the waveforms for music and language stimuli (Figures 1 and 2). Patel et al. (1998) also noted this difference and attributed it to theoretically-irrelevant differences between the musical and linguistic stimuli. The musical excerpts are rhythmic with short gaps of silence, while the sentences are more variable and continuous. Patel et al. argued that this could explain the difference. This seems plausible, but the statistical models they (and therefore we) used are limited to making comparisons on the mean amplitude in a particular time window, which is a substantial reduction in the information content of the waveforms. An advantage of making the full data set available is that other researchers can choose to analyze the data with other kinds of models. Another difference between the language and music waveforms reported by Patel et al. was a right anterior temporal negativity (RATN) in the 300-400 ms range (N350) only for the music condition. This was reported as an interesting, unexpected effect but not one that was important theoretically for the main result of similar processing of language and music structural violations. The RATN pattern was not evident in our music waveform data and the relevant statistical analysis did not replicate this element of Patel et al.'s findings (see Appendix for further details).

Of course, some concerns can only be addressed through changes to the experimental design, such as creating stimuli that have different properties designed to control for additional factors. Featherstone, Waterman, and Morrison (2012) point out potential confounding factors in the stimuli used by Patel et al. (1998) and other similar, subsequent studies. For example, the musical violations are both improbable in context and violate various rules of Western musical harmony. Direct replications, while crucial for establishing the reliability of a particular finding, necessarily also contain any methodological weakness of the original study. While we contend that this replication supports the empirical conclusions of the original study, we are mindful of the need to also examine support for the theoretical conclusion with a variety of methodological approaches. The relative increase in mean amplitude in the 500-800 ms window after structural violations in music and language might reflect shared processing resources, but it's also possible that there are two distinct processes that both generate this kind of EEG signal. As we described in the introduction, there is already a literature with numerous studies that examine the behavioral and neurological overlap between music and language, a literature in which debates about the best theoretical interpretation of the empirical findings are unfolding.

Finally, we note that there has been a growing interest in conducting serious replication studies in undergraduate and graduate research methods classes (Frank & Saxe, 2012; Grahe, Brandt, IJzerman, & Cohoon, 2014; Hawkins et al., 2018; Wagge, Baciu, Banas, Nadler, & Schwarz, 2019; Wagge, Brandt, et al., 2019). The hypothesized benefits are numerous: students act as real scientists with tangible outcomes, motivating careful and engaged work on the part of the students and benefiting the scientific community with the generation of new evidence; students learn about the mechanics and process of conducting scientific research with well-defined research questions and procedures, providing a stronger foundation for generating novel research in the future; reading papers with the goal of replication teaches students to critically evaluate the methods and rationales in order to be able to replicate the work (Frank & Saxe, 2012). Exposing the next generation of researchers to methodological innovations that improve replicability and reproducibility spreads those practices, hopefully producing a more reliable corpus of knowledge in the future.

Our experience with this project anecdotally supports these hypotheses. Students were engaged and produced high-quality work. Moreover, the replication project provided a strong foundation for novel experimental work. The class was structured so that smaller teams of students conducted original studies following the whole-class replication effort. Students were able to apply a variety of methodological skills learned from the replication project — pre-registration, data analysis techniques, use of the Open Science Framework, and, more abstractly, an understanding of what the complete research process entails — to this second round of projects.

Given our experiences, we endorse similar initiatives that involve students in replication work as part of their methodological training.

## Open Science Practices

This article earned the Preregistration Plus, Open Data and the Open Materials badge for preregistering the hypothesis and full analysis plan before data collection, and for making the data and materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

## Author Note

Correspondence regarding this article should be sent to Joshua de Leeuw: jdeleeuw@vassar.edu

## Author Contributions

de Leeuw and Andrews were the leaders of the project and are the first and second authors of this article. The remaining authors contributed equally and are listed in alphabetical order.

## Conflict of Interest

The authors have no conflict of interest to declare.

## References

Abrams, D. A., Bhatara, A., Ryali, S., Balaban, E., Levitin, D. J., & Menon, V. (2011). Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. *Cerebral Cortex*, *21*(7), 1507–1518.

Besson, M., Chobert, J., & Marie, C. (2011). Transfer of Training between Music and Speech: Common Processing, Attention, and Memory. *Frontiers in Psychology*, 2, 94.

Besson, M., & Faïta, F. (1995). An event-related potential (ERP) study of musical expectancy: Comparison of musicians with nonmusicians. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1278–1296.

Besson, M., Faïta, F., & Requin, J. (1994). Brain waves associated with musical incongruities differ for musicians and non-musicians. *Neuroscience Letters*, *168*(1-2), 101–105.

Besson, M., & Schön, D. (2001). Comparison between language and music. *Annals of the New York Academy of Sciences*, *930*, 232–258.

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

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *The Behavioral and Brain Sciences*, *31*(5), 489–508; discussion 509–558.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Featherstone, C. R., Morrison, C. M., Waterman, M. G., & MacGregor, L. J. (2013). Semantics, syntax or neither? A case for resolution in the interpretation of N500 and P600 responses to harmonic incongruities. *PLoS ONE*, *8*(11), e76600.

Featherstone, C. R., Waterman, M. G., & Morrison, C. M. (2012). Norming the odd: creation, norming, and validation of a stimulus set for the study of incongruities across music and language. *Behavior Research Methods*, *44*(1), 81–94.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600–604.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453.

Grahe, J., Brandt, M., IJzerman, H., & Cohoon, J. (2014). Replication education. APS *Observer*, 27(3).

Hawkins, R. X. D., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., … Frank, M. C. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science*, 1(1), 7–18.

Henry, L., & Wickham, H. (2017). purrr: Functional Programming Tools. Retrieved from https://CRAN.R-project.org/package=purrr

Herdener, M., Humbel, T., Esposito, F., Habermeyer, B., Cattapan-Ludewig, K., & Seifritz, E. (2014). Jazz drummers recruit language-specific areas for the processing of rhythmic structure. *Cerebral Cortex*, 24(3), 836–843.

Ingre, M., & Nilsonne, G. (2018). Estimating statistical power, posterior probability and publication bias of psychological research using the observed replication rate. *Open Science*, 5(9), 181190.

Janata, P. (1995). ERP measures assay the degree of expectancy violation of harmonic contexts in music. *Journal of Cognitive Neuroscience*, 7(2), 153–164.

Koelsch, S. (2011). Toward a neural basis of music perception - a review and updated model. *Frontiers in Psychology*, 2, 110.

Koelsch, S., Gunter, T. C., Wittfoth, M., & Sammler, D. (2005). Interaction between syntax processing in language and in music: an ERP Study. *Journal of Cognitive Neuroscience*, 17(10), 1565–1577.

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments. Retrieved from https://CRAN.R-project.org/package=ez

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, 113(2), 254–261.

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, 11(5), 1021–1031.

Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., & Friederici, A. D. (2012). Perception of words and pitch patterns in song and speech. *Frontiers in Psychology*, 3, 76.

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., & Besson, M. (2009). Musical training influences linguistic abilities in 8-year-old children: more evidence for brain plasticity. *Cerebral Cortex*, 19(3), 712–723.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs. Retrieved from https://CRAN.R-project.org/package=BayesFactor

MuseScore Development Team. (2018). MuseScore (Version 2.3.2). Retrieved from https://musescore.org/

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785–806.

Osterhout, L., & Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. *Language and Cognitive Processes*, 8(4), 413–437.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), 786–803.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, 6(7), 674–681.

Patel, A. D. (2010). *Music, Language, and the Brain*. Oxford University Press, USA.

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: an event-related potential study. *Journal of Cognitive Neuroscience*, 10(6), 717–733.

R Core Team. (2017). R: A Language and Environment for Statistical Computing (Version 3.4.2). Vienna, Austria: R Foundation for Statistical

Computing. Retrieved from https://www.R-project.org/

Rivas, D. (2016). jsPsych Hardware (Version v0.2-alpha). Retrieved from https://github.com/rivasd/jsPsychHardware

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5), 356–374.

Sammler, D., Baird, A., Valabrègue, R., Clément, S., Dupont, S., Belin, P., & Samson, S. (2010). The relationship of lyrics and tunes in the processing of unfamiliar songs: a functional magnetic resonance adaptation study. *Journal of Neuroscience*, 30(10), 3572–3578.

Sammler, D., Koelsch, S., Ball, T., Brandt, A., Grigutsch, M., Huppertz, H.-J., … Schulze-Bonhage, A. (2013). Co-localizing linguistic and musical syntax with intracranial EEG. *NeuroImage*, 64, 134–146.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5), 559–569.

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.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: do music lessons help? *Emotion*, 4(1), 46–64.

Tillmann, B. (2012). Music and language perception: expectations, structural integration, and cognitive sequencing. *Topics in Cognitive Science*, 4(4), 568–584.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638.

Wagge, J. R., Baciu, C., Banas, K., Nadler, J. T., & Schwarz, S. (2019). A demonstration of the Collaborative Replication and Education Project: replication attempts of the red-romance effect. *Collabra: Psychology*, 5(1). https://doi.org/10.1525/collabra.177

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing Research With Undergraduate Students via Replication Work: The Collaborative Replications and Education Project. *Frontiers in Psychology*, 10, 247.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). dplyr: A Grammar of Data Manipulation. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2018). tidyr: Easily Tidy Data with "spread()" and "gather()" Functions. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Hester, J., & Francois, R. (2017). readr: Read Rectangular Text Data. Retrieved from https://CRAN.R-project.org/package=readr

Wilke, C. O. (2017). cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2." Retrieved from https://CRAN.R-project.org/package=cowplot

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral and Brain Sciences*, 41, 1–50.

## Appendix

An unexpected discovery reported by Patel et al. was a right anterior temporal negativity (RATN) in the 300-400 ms window only for the music condition. Patel et al. also referred to this peak as an N350 and noted its potential relation to the left anterior negativity (LAN) reported for linguistic grammatical processing (but not observed by Patel et al. for their linguistic stimuli). Patel et al. note that these hemispheric effects of opposite laterality for language and music suggest distinct but possibly analogous cognitive processes and propose that they should receive additional investigation but do not discuss them further.

We did not pre-register any analyses of this effect because we did not consider it relevant to the theoretical claim of syntactic processing similarity between music and language shown by the P600 effect. However, in response to a reviewer's request we investigated whether our data supported Patel et al.'s finding of an RATN/N350 for the music condition.

The key statistical result reported by Patel et al. was a significant three-way interaction between condition (in-key chord vs. distant-key chord), hemisphere, and electrode site for the 300-400 ms window. The corresponding result for a grammaticality x hemisphere x electrode site ANOVA performed on our data was not significant ($F(4, 136) = .366$, $p = .832$), an outcome that fits with the appearance of the waveforms for the music condition of our experiment which show no sign of the RATN that appeared in Patel et al.'s Figure 5 (compare to our Figure 2).

In order to further address the strength of evidence provided by our data with respect to this three-way interaction, we conducted a Bayes factor analysis using the BayesFactor R package (Morey & Rouder, 2015) to evaluate the relative support for models containing fixed effects of electrode, hemisphere, and grammaticality (and their interactions) relative to the null model containing only a random effect of participant. The data are 431,034 times less likely under the full model that adds in all three main effects, the three two-way interactions, and the three-way interaction than under the null model. To isolate the contribution of the three-way interaction, we can compare the full model containing the three-way interaction to the model containing all terms except the three-way interaction. The data are 39 times less likely under the model with the three-way interaction. Thus, we clearly did not replicate the RATN for music reported by Patel et al.

The complete set of results for this analysis and the analysis scripts are available on the Open Science Framework at https://osf.io/zpm9t/.