Reassessing Syntax-Related ERP Components Using Popular Music Chord Sequences: A Model-Based Approach

Andrew Goldman
Western University, London, Canada

Peter M. C. Harrison
Queen Mary University, London, United Kingdom

Tyreek Jackson
St. John's University

Marcus T. Pearce
Queen Mary University, London, United Kingdom

Electroencephalographic responses to unexpected musical events allow researchers to test listeners' internal models of syntax. One major challenge is dissociating cognitive syntactic violations—based on the abstract identity of a particular musical structure—from unexpected acoustic features. Despite careful controls in past studies, recent work by Bigand, Delbe, Poulin-Carronnat, Leman, and Tillmann (2014) has argued that ERP findings attributed to cognitive surprisal cannot be unequivocally separated from sensory surprisal. Here we report a novel EEG paradigm that uses three auditory short-term memory models and one cognitive model to predict surprisal as indexed by several ERP components (ERAN, N5, P600, and P3a), directly comparing sensory and cognitive contributions. Our paradigm parameterizes a large set of stimuli rather than using categorically "high" and "low" surprisal conditions, addressing issues with past work in which participants may learn where to expect violations and may be biased by local context. The cognitive model (Harrison & Pearce, 2018) predicted higher P3a amplitudes, as did Leman's (2000) model, indicating both sensory and cognitive contributions to expectation violation. However, no model predicted ERAN, N5, or P600 amplitudes, raising questions about whether traditional interpretations of these ERP components generalize to broader collections of stimuli or rather are limited to less naturalistic stimuli.

Received: September 24, 2020, accepted July 3, 2021.

Key words: EEG, ERP, music and language, cognitive neuroscience, statistical learning

Like natural language, music does not consist of a randomly ordered collection of sounds; rather, musical events and structures are combined into sequences according to constraints that characterize a musical style (Patel, 2003; Rohrmeier & Pearce, 2018). A set of such constraints may be described as the syntax of a musical style, similar to how the constraints determining word order in natural language constitute the syntax of the language. Music syntactic constraints operate at different hierarchical levels, applying to notes, chords, motifs, or other aspects of tonal structure and musical form.

Internal models of the syntactic structure of musical styles are thought to be acquired through experience and it has been argued that this process of enculturation reflects implicit statistical learning of sequential structure, subject to certain innately specified limits (Pearce, 2018). In contrast to language syntax (but akin to linguistic semantics), constraints on the structure of music tend not to be wholly categorical (grammatical versus ungrammatical), but rather probabilistic: most sequential contexts allow for a range of different continuations varying in likelihood. Internal syntax models are therefore thought to model the likelihood of various different continuations to a given context, allowing listeners to generate expectations about the likely (and unlikely) continuations at each point in a piece of music as it proceeds. Enculturated listeners possessing such an internal cognitive model are therefore able to identify the extent to which continuations of a sequential context confirm or disconfirm their expectations by the extent to which they conform to or violate the (internalized) probabilistic syntax of the style. Here, we primarily consider harmonic elements of syntax, though such violations can also occur for other symbolic musical parameters such as melody.

Measuring listeners' sensitivity to these violations has been the traditional way to test such theories of musical syntax (see Musical Syntax Related ERPs); however, investigating the underlying cognitive models of statistical regularities is difficult. In studies of language, the surprisal of a word in a given context due to its syntactic category or meaning can be effectively dissociated from the sound of the word (although, there are certainly

Music Perception, volume 39, issue 2, pp. 118–144, issn 0730-7829, electronic issn 1533-8312. © 2021 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's Reprints and Permissions web page, https://www.ucpress.edu/journals/reprints-permissions. DOI: https://doi.org/10.1525/mp.2021.39.2.118

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

cases where the sound of a word can affect syntax, such as by adding pluralization suffixes). By contrast, in music, expected chords—due either to explicit knowledge of music-theoretic rules or mere statistical exposure—tend to have acoustic features (such as harmonics) in common with their context, too. Indeed, part of what constitutes the identity of any given syntactic element in music *is* its sound. This presents a major hurdle to the study of musical syntax because acoustic changes may be confounding syntactic unexpectedness, leading to potential misattribution of purported sensitivity to syntactic responses to what is actually sensory surprisal.

Further, for a harmony to be unexpected, a piece of music must first establish a local harmonic context, which has its own *local* distribution of pitch classes. An unexpected harmony may thus not only differ from global probabilistic expectations acquired in the long-term, but also in features relative to this local context. Sensory surprisal may also arise relative to this local context. Several researchers have noted this issue and have controlled for it by constructing local contexts that control features of acoustic similarity independent of harmonic classification. Bigand, Poulin, Tillmann, Madurell, and D'Adamo (2003) controlled whether or not the target chord occurred in the priming context and considered the number of shared pitch classes between target chords and their priming sequence. Koelsch, Jentschke, Sammler, and Mietchen (2007) controlled for pitch commonality, pitch repetition, and roughness (using the IPEM toolbox, Leman, Lesaffre, & Tanghe, 2005) across syntactically regular and irregular chords. Regnault, Bigand, and Besson (2001) compared violations based on acoustic dissonance (by raising the fifth of a target chord by a half step) with syntactic differences (based on the harmonic function of the target chord). In all of these cases, unexpected harmonies were still found to evoke neurophysiological responses corresponding to long-term internalized models of musical syntax distinct from those evoked by sensory expectation violations.

Despite these careful controls, some issues have yet to be unequivocally resolved. Notably, Bigand, Delbe, Poulin-Charronnat, Leman, and Tillmann (2014) revisited many past music syntax studies and provided a sophisticated computational analysis of the stimuli that had been used, arguing that while some acoustic characteristics were controlled in the stimuli, others were not adequately considered. In particular, they argued that the auditory short-term memory (ASTM) model of Leman (2000) could explain findings from several previous experiments, suggesting that the reportedly syntactic expectation violations could not be unequivocally dissociated from sensory surprisal, despite carefully crafted experimental controls.

To be clear, by sensory surprisal, we mean simple and low-level aspects of perceptual processing that do not rely on cognitive representations of the statistical regularities of a musical style, but instead are driven exclusively by short-term acoustic properties of the stimulus on timescales that can be processed in sensory memory. By cognitive surprisal, we mean higher-level aspects of processing that, in the examples we consider here, do depend on prior learned familiarity with a given musical style, acquired over longer periods of previous exposure to a large number of stylistic examples and stored in long-term memory, which allows the generation of expectations based on comparison with longer extracts of the musical context stored in short-term memory (rather than sensory memory). Huron (2006) outlines how cognitive surprisal can result from four different kinds of expectation, each relying on different kinds of memory. *Schematic expectations* arise from learning the regularities making up the syntax of a musical style through extensive long-term exposure to a large number of pieces of music in the style. *Dynamic expectations* arise from short-term learning of regularities occurring within a piece of music (e.g., thematic and motivic structure). *Veridical expectations* arise from episodic memories of pieces that we are familiar with from previous listening. Finally, *conscious expectations* arise from explicit mental prediction of forthcoming musical structures. In the present research, we are concerned specifically with schematic and dynamic expectations.

Another related issue concerns the listening contexts characteristic of many syntactic surprisal paradigms. In many previous event related potential (ERP) studies of syntax violation (as reviewed below), experimenters contrasted categorical violation versus non-violation conditions that differed by degree of violation. Such designs require repeating stimuli within each condition, to accommodate ERP paradigms that require such repetition, and then averaging across trials. One consequence of this is that participants may listen differently at certain points in the harmonic progressions once they have learned—over the course of doing the experiment—where the violations occur in the stimuli, and what those violations might be (because there are limited types of chord progressions used as stimuli). This may bias how participants listen, and may differ from ecological music listening in which syntactic violation, as indexed by sensory or cognitive surprisal, may vary continuously while listening to an unfamiliar piece of music (in a familiar style). Thus, whether sensitivity to

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

syntactic violations—as indexed by various ERP components—generalizes to chord sequences extracted from actual music presented without repetition is an open question that deserves further attention. Demonstrating this generalizability would allow a more precise attribution of the physiological responses: if the same neurophysiological responses were to be evoked when participants listen to stimuli that do not have distinct points where strong violations occur, it would be less plausible that such effects could be attributed to artifacts of listening strategies present only in experimental settings, or characteristics of the particular stimuli used in these studies.

This manuscript aims to address these concerns and is organized as follows. First we will review previous electroencephalography (EEG) literature on event related potentials (ERPs) theorized to be evoked by syntax violations. Then we will review four computational models that have been used to simulate the sensory and syntactic surprisal of musical stimuli. We then present our own experimental study that measures ERPs evoked by a broad set of stimuli sampled from a popular music corpus that have been evaluated using both sensory and cognitive models of surprisal, investigating which models best predict the ERPs that have been associated with violations of harmonic expectation, and addressing limitations of previous ERP research that we have identified.

Musical Syntax Related ERP Components

ERAN/EAN

Perhaps the most studied response to violations of harmonic syntax in music is the early right anterior negativity (ERAN). Listeners are typically presented with chord progressions that either conform to or violate the syntax of a musical style (usually Western tonal music). The ERAN occurs 150–280 ms after the onset of the unexpected chord and is maximal over the right anterior portion of the scalp (Koelsch, Gunter, Friederici, & Schröger, 2000; Koelsch, Kilches, Steinbeis, & Schelinski, 2008; Pearce & Rohrmeier, 2018), although it is sometimes referred to as the EAN or E(R)AN because it has not always been found to be right lateralized (e.g., Loui, Grent-'t-Jong, Torpey, & Woldorff, 2005). Here we retain the original ERAN terminology. The amplitude is larger when the unexpected chord follows a larger number of expected chords that more firmly establish the context, and the amplitude is smaller for less unlikely substitutions in the context (Leino, Brattico, Tervaniemi, & Vuust, 2007). The amplitude of this component is insensitive to task relevance (Koelsch et al., 2000), though it is sensitive to attentional load (Loui et al., 2005). The ERAN has been observed for artificially introduced chords that are highly stylistically incongruent with the context (e.g., a Neapolitan sixth taking the place of a tonic triad in an authentic cadence) but also for stylistically unlikely but not entirely ungrammatical chords appearing in the chorales harmonized by J. S. Bach (Steinbeis, Koelsch, & Sloboda, 2006) as well as expressively performed classical and romantic piano sonatas (Koelsch et al., 2008).

The ERAN has been localized to Broca's area and its right hemisphere homologue, with a bias towards activity in the right hemisphere (Maess, Koelsch, Gunter, & Friederici, 2001), whereas the early left anterior negativity (ELAN), elicited by syntax violations in language, show similar source localization but with a left-hemisphere bias (Friederici, 2011). Carrus, Pearce, and Bhattacharya (2013) show that the ELAN (or, in their terminology, the *left anterior negativity*, or LAN) to syntactic violations in language (but not the N400 to semantic incongruity, see next section) is reduced in amplitude when presented simultaneously with a low-probability melodic note. Kunert, Willems, Casasanto, Patel, and Hagoort (2015) used fMRI to demonstrate that a task involving both music syntactical processing and language syntactical processing share neural resources in Broca's area. These results reinforce the association between the ERAN and syntax-related cognitive processes, although increasing evidence suggests that interactive effects of musical and linguistic structure are not specific to syntax and may reflect shared use of general-purpose cognitive control processes rather than overlapping processing of syntax per se (Slevc & Okada, 2015).

Compared with the mismatch negativity (MMN), which has a similar latency and topography but is not right lateralized (Näätänen, Paavilainen, Rinne, & Alho, 2007), it has been argued that the ERAN reflects sensitivity to the long-term acquired knowledge of statistical regularities characteristic of music-syntactic processing rather than sensitivity to local violations (Koelsch, 2009). The MMN, by contrast, has traditionally been thought to be related to sensory memory (although further research has shown that long-term memory and top-down processing play a role; Näätänen et al., 1997). Evidence comes from studies showing that the amplitude of the ERAN is related to the long-term transition probability of the chord (Kim, Kim, & Chung, 2011), is attenuated (though still present) in 5–6-year-old children compared to adults, and is accentuated in adult musicians, relative to adult nonmusicians (Koelsch, Schmidt, & Kansok, 2002). Furthermore, in an artificial

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

grammar learning study using unfamiliar stimuli, the amplitude of the ERAN was found to increase with greater learning of the grammar measured both by degree of exposure and performance in a grammaticality decision task (Loui, Wu, Wessel, & Knight, 2009). In addition, Miranda and Ullman (2007) describe a functional dissociation between an ERAN associated with out-of-key violations in both familiar and unfamiliar melodies, and a subsequent (220–380 ms) posterior negativity elicited by both in-key and out-of-key violations of familiar melodies only.

Finally, in a study of melodic expectation, Omigie, Pearce, Williamson, and Stewart (2013) used a computational model of auditory expectation based on statistical learning and probabilistic prediction (Pearce, 2005, 2018) and found that the amplitude of an early negative component—which usually occurs earlier at around 100–150 ms for violations of melodic expectation (Koelsch & Jentschke, 2010; Pearce, Ruiz, Kapasi, Wiggins, & Bhattacharya, 2010)—correlated positively with the unexpectedness of the note according to the model.

N5

In conjunction with the ERAN, violations of harmonic syntax also evoke a later bilateral negativity maximal around 500–570 ms (Koelsch et al., 2000; Loui et al., 2005), termed the N5 (or N500). This component is thought to reflect the subsequent integration of unexpected information into the ongoing processing of a chord sequence such that each successive expected chord requires less processing, and shows decreasing N5 amplitudes. Koelsch (2011) interprets the N5 as a process by which the interrelation between musical structures gives rise to meaning, thus implicating the N5 as sensitive to what can be thought of as musical semantics. Steinbeis and Koelsch (2008) showed that the N5 is reduced when presented alongside language semantic violations, suggesting that the N5 shares neural resources with processing of semantics in language. This is reinforced by Sun et al., (2018) who found that amusics—participants with impairment in discriminating pitch—showed attenuated ERAN and ELAN but intact N5 and P600 to syntactic violations of melody and language respectively. The N5 has an analogue in language, the N400, which is elicited by words that are syntactically correct but semantically implausible, e.g., the final word in "she drinks her coffee with sugar and pepper" (Kutas & Hillyard, 1980). N400s are elicited by every word in a sentence, but show reduced amplitude as context accumulates (see Swaab, Ledoux, Camblin, & Boudewyn, 2012, for a review).

P600

Another ERP component associated with syntactic violation in language is the P600, a posteriorly distributed positivity maximal around 600 ms after the stimulus onset (Osterhout & Holcomb, 1992, 1993). Patel, Gibson, Ratner, Besson, and Holcomb (1998) reported increased P600 amplitude in response to chords that violate Western tonal harmonic syntax, given the preceding harmonic context. The P600 indexes how easily a chord can be integrated into the ongoing harmonic context. Featherstone, Morrison, Waterman, and MacGregor (2013) observed a P600 in musicians and nonmusicians for unexpected chords but found that only musicians showed a P600 when expectations were not subsequently fulfilled by returning to the initial key. They suggested that the P600 reflects conscious analytical detection of a contextual incongruity. Compared with classical musicians, jazz musicians show a smaller P600, suggesting enhanced ability to integrate unexpected events into the ongoing context (Przysinda, Zeng, Maves, Arkin, & Loui, 2017).

P3

Early EEG studies of melodic expectation (Besson & Faita, 1995; Nittono, Bito, Hayashi, Sakata, & Hori, 2000; Paller, McCarthy, & Wood, 1992; Schön & Besson, 2005; Verleger, 1990) identified a P3 component (sometimes called a P300 or late positive component) peaking between 300–500 ms at central and posterior sites in response to stylistically unexpected notes in a melody. The amplitude and latency of the P3 are sensitive to musical expertise, the familiarity of the melody, the degree of expectancy violation (Besson & Faita, 1995), and also to the timing of the unexpected note (Nittono et al., 2000). The P3 is now understood to consist of two functionally distinct components: the P3a, an earlier and more frontally/centrally distributed response to novelty per se, including attentional orienting to novel stimuli (musical or otherwise); and the P3b, a later and more parietally distributed response to low-probability events that is dependent on the event in question being task relevant, e.g., requiring the participant to explicitly identify it (Comerchero & Polich, 1999; Polich, 2007; Walsh, Gunzelmann, & Anderson, 2017). Koelsch and Jentschke (2010) argue that the P3 observed in early EEG studies of melodic expectation reflected a task-related P3b rather than a response to violation of syntactic expectation per se. However, some research that has focused on the ERAN (see above) has also reported both P3a and P3b to violations of harmonic expectation (Koelsch et al., 2000; Steinbeis, Koelsch, & Sloboda, 2006), though the P3 is not significant in all cases

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

(e.g., Koelsch et al., 2008). Furthermore, in a magnetoencephalography (MEG) study, Vuust, Ostergaard, Pallesen, Bailey, and Roepstorff (2009) report a P3 am (the MEG analogue of the P3a in EEG recordings) using drum patterns containing violations of metrical expectation, which is interpreted to reflect the integration of a prediction error into a larger processing network. Trainor, McDonald, and Alain (2002) report a P3a elicited when the melodic contour or interval of a standard melody was changed.

Computational Models of Harmonic Expectation

The ERP study reported in this paper reconsiders the components described above in the context of four models of harmonic expectation, each formalizing different hypotheses about how listeners process chord sequences. Three of these models may be described as Auditory Short-Term Memory (ASTM) models, positing that harmonic expectation derives from the short-term retention and comparison of auditory images; these models represent sensory accounts of harmonic expectation. The fourth model treats harmonic expectation as the outcome of culturally informed statistical learning and probabilistic prediction, thereby representing a cognitive account of harmonic expectation. We used each model to calculate the degree of unexpectedness (be it sensory or syntactic) of chords we use as stimuli in the experiment reported below.

Milne, Sethares, Laney, and Sharp's (2011) Spectral Distance Model

Milne et al.'s (2011) spectral distance model is the least complex of the ASTM models evaluated here. The model estimates the perceptual dissimilarity of sets of pitches or pitch classes by expanding each pitch(-class) into its implied harmonic series, blurring the resulting spectrum by convolution with a Gaussian distribution to match the resolution of human pitch perception, and computing the cosine distance between the resulting spectra. The resulting model has proved effective in predicting various aspects of tonal perception (Milne & Holland, 2016; Milne, Laney, & Sharp, 2015, 2016). Because the model outputs a value describing the similarity of different spectra, it can be used as a predictive model: less similar spectra are less predictable in terms of acoustic features.

Leman's (2000) Periodicity Pitch Model

Leman's (2000) periodicity-pitch model is broadly similar to Milne et al.'s (2011) model, but includes three additional features. First, it is not constrained to inputs of pitch or pitch-class sets, but can take arbitrary audio input. Second, it includes a detailed simulation of how the audio signal is transformed by the peripheral auditory system, including acoustic filtering by the outer and middle ear, frequency analysis in the cochlea, conversion to neural rate-coding by inner hair cells, and periodicity analysis for pitch inference. Third, its memory extends past the most recent chord, accumulating every incoming sound in an echoic memory buffer (termed the *global image*) that decays exponentially over time. Leman's (2000) model may be used to predict the expectedness of an incoming chord by correlating the auditory image in the short-term memory buffer (the global image) with the auditory image evoked by the incoming chord (termed the local image). In this model, chords eliciting *high* correlations should be perceived as more *expected* (less surprising). Note that this is in contrast to the three other models described here, where *high* values correspond to more *surprising* events.

Collins, Tillmann, Barrett, Delbé, and Janata's (2014) Tonal Expectation Model

Collins et al.'s (2014) tonal expectation model represents a hybrid sensory-cognitive account of harmonic expectation. Like the previous two models, it is an auditory short-term memory model, predicting harmonic expectation by accumulating and comparing auditory images in short-term memory. However, it includes a wider variety of auditory representations, some of which are derived from cognitive processing.

The first auditory representation is a *periodicity-pitch* representation identical to that used in Leman's (2000) model. This representation describes the pitch content of the stimulus as a function of time. The second representation is a *chroma-vector* representation, derived from the periodicity-pitch representation by collapsing pitches to pitch classes, and thereby capturing the music-theoretic principle of octave invariance. The third representation is the *tonal-space* representation from Janata et al. (2002), intended to capture the listener's implicit knowledge of Western tonality. The tonal-space representation projects the periodicity-pitch representation onto the surface of a torus, a three-dimensional ring-like structure shown by Krumhansl and Kessler (1982) to capture perceptual distances between Western musical keys. The mapping between periodicity pitch and tonal space is defined by a self-organizing map (Kohonen, 1995), a type of artificial neural network that uses competitive learning to reduce the dimensionality of input data. This self-organizing map is pretrained on a melody designed to modulate through all 24 major and minor diatonic keys over the

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

course of eight minutes. At a given timepoint, each of these three representations defines a vector (or, in the case of the tonal-space representation, a two-dimensional array) of numbers corresponding to the activation pattern of that representation at that timepoint. For example, elements in the periodicity-pitch vector represent activations of different auditory nerve fibres, whereas elements in the tonal-space array represent activations of particular tonal regions, such as G major or B minor.

Analogous to Leman (2000), local and global images are created for each of these three representations, with the local image capturing the current sound and the global image capturing the tonal context as maintained in short-term memory. A 0.10 s half-life is used for accumulating the local images, and a 4.00s half-life for the global images.

For each time point, two types of features are computed from the local and global images. The first type of feature corresponds to the maximum value found within the global image, and is intended as a measure of tonal clarity. The second type of feature is based on correlations between local and global images, similar to Leman (2000). In particular, the model computes the difference in local-global correlations immediately before and just after the target chord; large increases in correlation imply that the new chord induces significant tonal resolution. Collins et al. (2014) define a large number of features through the factorial combination of these representations and feature types.

The model then predicts the expectedness of the target chord as a linear combination of these derived features—thus producing a single value—using coefficients from a stepwise regression analysis of seven empirical studies (Collins et al., 2014). These empirical studies were harmonic priming studies, quantifying tonal expectedness as the speed with which listeners perform simple perceptual judgements (e.g., timbre discrimination) on the target chord.

THE INFORMATION DYNAMICS OF MUSIC MODEL

The Information Dynamics of Music (IDyOM) model embodies a statistical-learning account of musical expectation, originally developed in the context of melody perception (Pearce, 2005; Pearce & Wiggins, 2012) and subsequently extended to the harmonic domain by Harrison and Pearce (2018). The model represents the hypothesis that listeners learn statistical properties of a musical style through passive exposure, and use these learned statistics to generate probabilistic predictions for successive musical events. We categorize the IDyOM model as "cognitive" rather than "sensory" because it simulates high-level cognitive processes involved in statistical learning and probabilistic prediction of musical structure, symbolically encoded at the level of musical notes rather than lower-level acoustic features, with both learning and prediction taking place over time scales greater than the capacity limits of sensory memory (up to about 4 s). This includes top-down expectations derived from musical knowledge acquired through listening to hundreds or thousands of pieces of music over periods of years as well as knowledge acquired within individual pieces of music over periods of seconds and minutes.

The underlying prediction engine of IDyOM is the Prediction by Partial Match (PPM) algorithm, a variable-order Markov model initially presented by Cleary and Witten (1984) and subsequently developed by Moffat (1990), Cleary and Teahan (1997), and Bunton (1997). PPM learns the statistical structure of a set of sequences composed from some *alphabet* of elements by recording the frequency with which each element in the alphabet follows each subsequence of the input, and then using these learned statistics to compute distributions that estimate the conditional probability of each symbol appearing at each point in every sequence. In the present use of PPM within IDyOM, the elements are musical chords that are organized into sequences in the pieces of music making up the corpus. The PPM algorithm works by blending together many n-gram models of different orders, where an n-gram model predicts the next symbol in a sequence by defining the *context* as the previous $n - 1$ symbols, and tabulating the different continuations to that context previously observed in the training data. These n-gram models are blended together using *interpolated smoothing*, with the maximum value of n depending on the amount of training data available (Bunton, 1997; Pearce, 2005).

The original PPM algorithm is domain-agnostic, being originally developed for general-purpose data compression. For example, applied to chord symbols, PPM cannot incorporate the intuition that G7 and G9 chords share the same root and are therefore likely to have similar statistical properties. IDyOM addresses this problem using the *multiple-viewpoint* technique of Conklin and Witten (1995). Multiple-viewpoint models operate over sequences of abstract symbols, such as chord symbols, and compute various derived features (termed *viewpoints*) from these symbols. A good feature equates symbols with similar function; for example, a chord root feature combines all chords with the same root into one symbol. Each derived feature is modeled using a separate PPM model. Predictions from these different PPM models are then combined using an

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

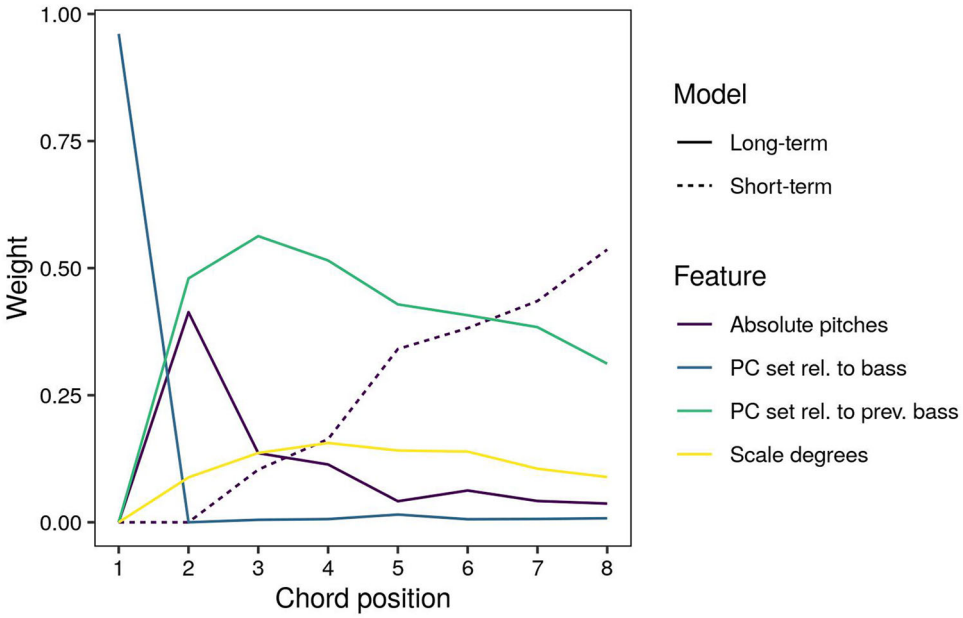


FIGURE 1. Viewpoint weights in Harrison and Pearce's (2018) IDyOM model as optimized on the present stimulus set. Weights are plotted as a function of chord position in the stimulus, model type (long-term versus short-term), and underlying feature. "PC" is an abbreviation of "pitch class"; "rel." is an abbreviation of "relative."

ensembling technique; after Harrison and Pearce (2018), we use a supervised arithmetic weighting scheme, where the weight of each viewpoint is optimized to maximize the likelihood of the stimulus set, and viewpoint weights are optimized separately—in the case of our experiment reported below, for each of the eight chord positions in each of our stimulus.

The selected viewpoints included four features (Figure 1). *Absolute pitches* is a representation of the chord's exact pitch content. *Pitch-class set relative to bass* expresses the chord as a pitch-class set where each pitch class is expressed relative to the bass note. *Pitch-class set relative to previous bass* expresses the chord as a pitch-class set relative to the bass note of the previous chord. *Scale degrees* expresses each pitch class in the chord relative to the local tonic, where the local tonic is estimated using Albrecht and Shanahan's (2013) key-finding algorithm with a sliding window of 16 quarter notes, and preserving inversion information (i.e., the knowledge of which pitch class was the bass note).

In Harrison and Pearce's (2018) IDyOM model, models can be trained in two different ways. *Long-term models* are pre-trained on a musical corpus intended to represent the listener's prior musical exposure; *short-term models* are trained from scratch on each stimulus (in our case, individual chord progressions), and are intended to capture local regularities in the musical extract. The predictions generated by the long- and short-term IDyOM models simulate respectively schematic and dynamic expectations, as outlined by Huron (2006). For each stimulus in our experiment reported below, the musical corpus for the long-term model is the 739 compositions in the Billboard popular music corpus, minus the particular composition from which the particular stimulus is excerpted; the short-term model is then trained incrementally on the eight chords in the stimulus (see Method). Here, the optimized model comprises a short-term model for the viewpoint *absolute pitches* and long-term models for all four selected viewpoints (*absolute pitches*, *pitch-class set relative to bass*, *pitch-class set relative to previous bass*, and *scale degrees*). The different viewpoints are assigned different weights over the course of the eight chords in each stimulus in such a way that maximizes the ability to predict the given chord (see Figure 1 and Method). For example, the *absolute pitch* model starts with zero weight, reflecting the fact that it is not yet possible to predict any pitch classes from a local distribution because none have occurred yet; as the stimulus progresses, the model accords *absolute pitch* successively more weight because it becomes possible to predict pitch classes on the basis of the local (short-term, stimulus specific) distribution. By contrast, pitch class set relative to bass contributes substantially to the model

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

at chord 1 because it accurately captures the prevalence of different chord types, but it contributes little after chord 1 because its transposition invariance makes it poorly suited to capturing phenomena such as local tonality and root progressions.

Thus, each chord in the stimulus is associated with a conditional probability distribution over the set of possible chords, representing the model's predictions at the previous timestep. We operationalize expectedness as the probability assigned to the chord that actually occurred; following Pearce (2005), we transform this probability by taking the negative logarithm (base two), terming the result *harmonic information content* (Harmonic IC). High information content means that a chord is considered to be unexpected. The IDyOM-derived information content for a given chord in a chord progression thus represents a weighted combination of the four viewpoints based on the local distribution of features (short term model) as well as based on a larger representative sample—i.e., the entire corpus of stimuli (long term model). Additional details about this calculation are reported below (see Method).

Finally, we note that IDyOM models a low-level constrained grammar (i.e., finite-state as opposed to less constrained grammars such as context-free). Thus, in using IDyOM to model psychological surprisal, we do not claim to be modeling all possible aspects of syntactic processing, and we do not claim that psychological syntactic processes are purely finite state.

SUMMARY OF HARMONIC EXPECTATION MODELS

The four models of harmonic expectation—Milne et al. (2011), Leman (2000), Collins et al. (2014), and IDyOM—thus cover a range of different possible contributions to expectation violation. The Milne et al. model covers information about spectral content and is thus a purely acoustic model. The Leman model filters spectral information according to ear physiology, representing a psychoacoustic model. The Collins et al. model incorporates elements of both sensory and cognitive features. The IDyOM model is a purely cognitive model. There are other models in the literature that would cover additional elements of acoustic processing, e.g., Thompson and Parncutt (1997) model masking effects. However, the models we have selected cover a range from sensory to cognitive effects of expectation violation and have been widely used in existing literature; thus, they are suitable for our purposes.

MELODIC FEATURES

In addition to the four main models described above, here we also considered model predictions based on melodic features as an additional exploratory analysis. To be clear, our main aim is to test the harmonic models, but there are two reasons to investigate melodic models as well. The first is simply a matter of convenience: with our experimental paradigm, we are able to test melodic models just as easily as harmonic ones. Second, and more importantly, melodic models act as an experimental control. Even in studies that use chords as stimuli, listeners may be surprised by features associated with melody (Hansen & Pearce, 2014; Hansen, Vuust, & Pearce, 2016) as well as contrapuntal features of the voice leading between chords. We thus considered four measures of melodic surprisal. *Melodic entropy* measures the predictive uncertainty of the melody notes only, calculated using the IDyOM model; higher values mean more predictive uncertainty in the melody notes. *Melodic information content* is analogous to the Harmonic IC used in the main analysis but uses only the melody notes (the top notes in each chord) and is also calculated using IDyOM, after Pearce (2005); higher values mean more surprisal in that melody note. We also included metrics of *melodic distance* and *voice-leading distance* (see Method). The inclusion of these metrics in part acts as a control so that we can properly attribute any observed effects to our primary variables of interest: models of harmonic expectation. Of course, models of syntax for other musical parameters deserve their own studies, but our focus here is primarily on harmonic features.

The Present Study

The aim in the present experiment is to examine whether the four models reviewed above predict ERP components associated with harmonic expectation. We investigate whether the three ASTM models can account for the four ERP components reviewed above and compare these to a cognitive model (IDyOM) designed to control for sensory confounds in its evaluation of the surprisal of harmony. We hypothesize that more unexpected chords according to each model should elicit larger ERAN, N5, P600, and P3a components, though the previous literature does not permit confident predictions about which model will best predict these amplitudes, nor how the sensory models will compare with the cognitive model. Given the nature of our linear statistical modeling in the experiment reported below—i.e., comparing across a range of stimuli rather than measuring difference waves between two categories (see Method)—ERAN and N5 component amplitude effects could manifest as larger negativities or smaller positivities, and for the P600 and P3a,

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

component amplitude effects could manifest as larger positivities or smaller negativities.

We measure component amplitude based on predetermined latency ranges in the ERP waveform. We note that changes in the ERP waveform are not equivalent to changes in ERP component amplitude (because a given latency range may contain a number of independently varying components). However, past ERP studies have also measured component amplitudes by examining ERP waveforms without necessarily using statistical techniques to dissociate components (such as independent components analysis). By using a priori predefined latency ranges, based on the literature, we make our findings comparable with previous results.

We presented listeners with chord progressions randomly sampled from the McGill Billboard corpus of commercially successful Western pop music (Burgoyne, Wild, & Fujinaga, 2011). Each of the four models produced a surprisal value for each chord in each progression. As explained in Method, we also included four measures of melodic expectation as an additional exploratory analysis. We note that using the Billboard corpus as the basis of a study on prediction—particularly with regard to IDyOM's cognitive modeling—presumes that listeners are enculturated in popular music to a sufficient extent to have internalized the structural regularities inherent in the corpus.

This experiment expands on previous research in three ways. First, we directly compare sensory and cognitive models as explanations for neural signatures of harmonic expectations. Second, by using a broad range of chord progressions from real-world pop songs, we examine the extent to which ERP components previously reported for a limited range of stimuli, often artificially manipulated, generalize to commercially successful popular music. In doing so, we assess a wide range of stimuli that parameterize expectedness across a broader and more nuanced range (rather than having only two or three categorical conditions of expected vs. unexpected), which is more characteristic of—albeit not equivalent to—ecological listening, and allows for the theories to be generalized across a wider range of real-world stimulus contexts (see Discussion for a consideration of the so-called *fixed-effect fallacy* in psychological research). Third, because we use a large number of stimuli that are not repeated, we avoid the possibility (present in previous ERP studies of syntax violation) that participants may learn where to expect violations in the stimuli and what violations might occur.

This approach may lead to less pronounced ERP component amplitude differences because we are not exclusively sampling the extremes of the expectation range (i.e., the most likely versus the least likely chord continuations). Also, estimates of ERP amplitudes for individual stimuli will have low reliability due to the large amount of variation in EEG signals. However, by including many chord progressions (see Method) each with its own set of sensory and cognitive model predictions, we gain the ability to draw statistical inferences across a wide range of responses to the different stimuli. If the theory is that the ERP components are really responses to syntax processing, it is important to show that these effects can be evoked by a range of ecologically valid stimuli. Our approach provides for this.

Method

PARTICIPANTS

Thirty participants, recruited in the greater New York City metropolitan area took part in the study (mean age = 29.93 years, $SD = 9.02$, 13 male, 16 female, 1 other). The inclusion criteria required participants to be aged 18–64 years, right handed (note that one participant scored as ambitextrous on the handedness questionnaire), and to have no history of neurological disease. Participants gave informed consent to participate in the study in accordance with the ethical policies of Columbia University, and were compensated $20/hr for participating.

Five participants' data were eliminated from the final analysis (see Data Analysis below). The 25 participants included in the final analysis ranged in age from 18–43 years old ($M = 28.24$, $SD = 6.18$). All participants scored as right-handed on the Edinburgh Handedness Inventory ($M = 87.65$, $SD = 14.94$). Participants' scores on the Goldsmith Musical Sophistication Index musical training subscale ranged from 2–47, $M = 28.00$, SD = 14.69. Participants were not selected on the basis of their musical experience and this information is provided only for purposes of replicability; further, we are not confident that we would have enough statistical power to detect effects of the MSI scores. We do not conduct any analyses of putative effects of Goldsmith MSI scores on ERPs.

MATERIALS

Questionnaires

Participants completed two questionnaires before undertaking the experimental task. The first was the Musical Training subscale from the Goldsmiths Musical Sophistication Index self-report questionnaire (Müllensiefen, Gingras, Musil, & Stewart, 2014). The second was the Edinburgh Handedness Inventory (Oldfield, 1971).

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

Corpus

We derived our corpus from the McGill Billboard corpus (Burgoyne et al., 2011), which (at the time of writing) comprised transcriptions of 739 unique compositions sampled from the Billboard "Hot 100" charts between 1958 and 1991. The original corpus represents chords using textual labels such as "Db: maj7"; we translated these chords to a *pitch-class chord* representation, defined as a tuple of the chord's bass pitch class and the set of non-bass pitch classes, using the chord dictionary from the *hrep* R package (Harrison & Pearce, 2020). Under this representation, the textual label "Db: maj7" would be mapped to the tuple '(1, {0, 5, 8})'. As part of this preprocessing we collapsed consecutive chord repetitions into single chords and removed section repeats. The resulting dataset is available in the *hcorp* R package (https://github.com/pmcharrison/hcorp).

Stimuli

Three hundred 8-chord sequences were randomly sampled from the corpus described above. Each chord lasted one second each and was synthesized with a piano timbre using the software Timidity (v. 2.14.0; http://timidity.sourceforge.net).[1] On 50% of trials, the final chord of the progression was panned to the left (25% of trials) or the right (25% of trials); panning was implemented by increasing the sound pressure level at one ear by 0.83 dB, and decreasing it at the other ear by the same amount. For the other 50% of trials, the final chord had equal sound pressure level in both ears. These manipulations were performed using the audio software SoX (http://sox.sourceforge.net/). Each stimulus was randomly assigned to one or the other of these conditions individually for each participant. The task described below (see Procedure) involved detecting these differences in order to ensure participants remained alert during the course of the experiment.

Model Predictions

For each of the eight chords in each of the 300 progressions, we calculated the surprisal prediction from the three ASTM models and the one cognitive model (see Background). From here on, we term these "Leman," "Milne," "Collins," and "Harmonic IC" (harmonic information content derived from the IDyOM model). We thus had three ASTM sensory predictors and one cognitive predictor whereby we could examine their relationship to the ERP components of interest.

Milne. This model simulates harmonic expectation by computing the spectral dissimilarity between each chord and its immediate predecessor (thus the first chord in each progression does not have a defined value). High model predictions correspond to less expected events.

Note that this model can be applied either to pitch or pitch-class representations of chords. Pitch-class representations, common in music theory, embody the well-established principle of *octave invariance*, which holds that a given note's musical function is largely preserved under octave transpositions (i.e., frequency manipulations of the form $f \rightarrow 2^n f$, where n is an integer). We used pitch-class representations, consistent with prior harmonic applications of the model (Milne & Holland, 2016).

We implemented the model in R (R Core Team, 2018), verifying a selection of outputs against the original author's MATLAB implementation (http://www.dynamictonality.com/probe_tone_files/). We parametrized the model according to the optimized configuration reported in Milne and Holland (2016): each pitch was represented as 12 harmonics (including the fundamental frequency), with the ith harmonic $(1 \leq i \leq 12)$ having a magnitude of $i^{-0.75}$, and blurred by a Gaussian distribution of standard deviation 6.83.

Leman Model. We used the model implementation provided in the IPEM toolbox (http://www.ipem.ugent.be/Toolbox; Leman, Lesaffre, & Tanghe, 2001), which is parametrized by two decay constants determining the half-lives of the local and global images. In Leman (2000) these decay constants were optimized to 0.10 s and 1.50 s respectively, but Bigand et al. (2014) have subsequently argued for exploring different parameter combinations. Here we evaluated the model for six parameter combinations (0.1/1.5; 0.1/2.5; 0.1/4.0; 0.5/1.5; 0.5/2.5; 0.5/4.0) and found that model predictions were affected little by parameter choice (mutual correlations > .95), so we retained the original parameter set of 0.10 s and 1.50 s. Note that for this model, lower values correspond to more unexpected events (in contrast to the other three models used in this experiment).

Collins. To calculate these model predictions, we used the model implementation provided in the Janata lab music toolbox (http://atonal.ucdavis.edu/resources/software/jlmt/), an extension to the IPEM toolbox (Leman et al., 2001). We left all parameters at their default values.

Harmonic IDyOM Model. We pretrained the IDyOM model on the popular music corpus described above to

[1] We note that previous studies (such as Koelsch et al., 2007) used shorter chord durations. (In their case, for 5-chord-long progressions, 600 ms / chord for chords 1–4, and 1,200 ms for the final chord.) We used longer inter-chord time intervals because we are investigating ERP components extending beyond 600 ms, and are measuring the ERPs evoked by each chord.

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

represent the prior musical exposure of the average Western listener. Though individual listeners will differ in their musical backgrounds, and will typically be exposed to additional styles outside Western popular music, we reasoned that the average Western listener will have received significant exposure to Western popular music over their lifetimes, whether through incidental exposure (e.g., radio, TV) or active listening (e.g., CDs, concerts), and will use this internalized knowledge when listening to our stimuli. Naively generating IDyOM predictions for compositions within the training set would be problematic, however, because the model would memorize the compositions perfectly and generate predictions reflecting veridical knowledge of the specific compositions rather than schematic knowledge of the musical style. We therefore used a leave-one-out cross-validation scheme for generating predictions from the long-term models. This means that, for each stimulus, the long-term models are pretrained on all chord progressions in the corpus except the composition from which the stimulus was originally extracted, thereby avoiding overlap between the models' training and test datasets. Having generated predictions for the individual viewpoint models using leave-one-out cross validation, we then optimized the viewpoint weights to maximize the likelihood of the stimulus set (i.e., the ability to predict each chord conditioned on the previous chords), after Harrison and Pearce (2018). Eight sets of viewpoint weights were estimated in total, corresponding to the eight chord positions in the stimuli. The resulting viewpoint weights are plotted in Figure 1. We then used these weights to generate final model outputs for the stimulus set, computing model surprisal (Harmonic IC) for each chord in each stimulus conditioned on the preceding chords in that stimulus.

Melodic IDyOM Model. We also analyzed the stimuli using the melodic IDyOM model of Pearce (2005). In particular, we supposed that listeners hear a melodic line joining the uppermost pitches of each chord in the stimulus, and develop probabilistic expectations for this melodic line that reflect their prior experience with Western tonal melodies. Correspondingly, we pretrained the melodic IDyOM model using a dataset of 1,000 European folksongs encoded by Damien Sagrillo and publicly available on KernScores (https://kern.humdrum.org/), intended to reflect general principles of melody construction in Western music. Note that IDyOM has successfully predicted listeners' melodic expectations in a variety of styles, even when trained on folksongs, indicating that folksongs provide a good general approximation of melodic expectation in Western tonal music (Egermann, Pearce, Wiggins, & McAdams, 2013; Omigie et al., 2019). We then used this model to generate predictions for the implied melody lines of each chord sequence in the stimulus set, configuring the model with the two viewpoints *chromatic pitch* (i.e., the MIDI note number) and *chromatic pitch interval* (i.e., the signed interval from the previous pitch in semitones), otherwise retaining the default settings from Pearce (2005). As before, we summarized model predictions as the model surprisal (Melodic IC) for each melody note conditioned on the previous melody notes in the stimulus. After Pearce (2005), we also computed the model's uncertainty for predicting each note in the melody (Melodic Entropy), operationalized as predictive entropy (i.e., the expected information content, conditioned on the prior melody notes, in advance of observing the next melody note).

Other Melody Features. We also calculated two additional predictors relating to melody: *melodic distance* measures the unsigned distance in semitones from one melody note to the next, and *voice-leading distance* (VL Distance) calculates the minimal voice-leading distance between the chord and the next chord, as estimated by Tymoczko's (2006) algorithm using a taxicab pitch-distance metric. For both of these models, high values mean a greater distance. These metrics allowed us to assess whether melodic features including voice-leading characteristics—as an alternative to our main theoretical interest in harmonic features—might also predict ERP components.

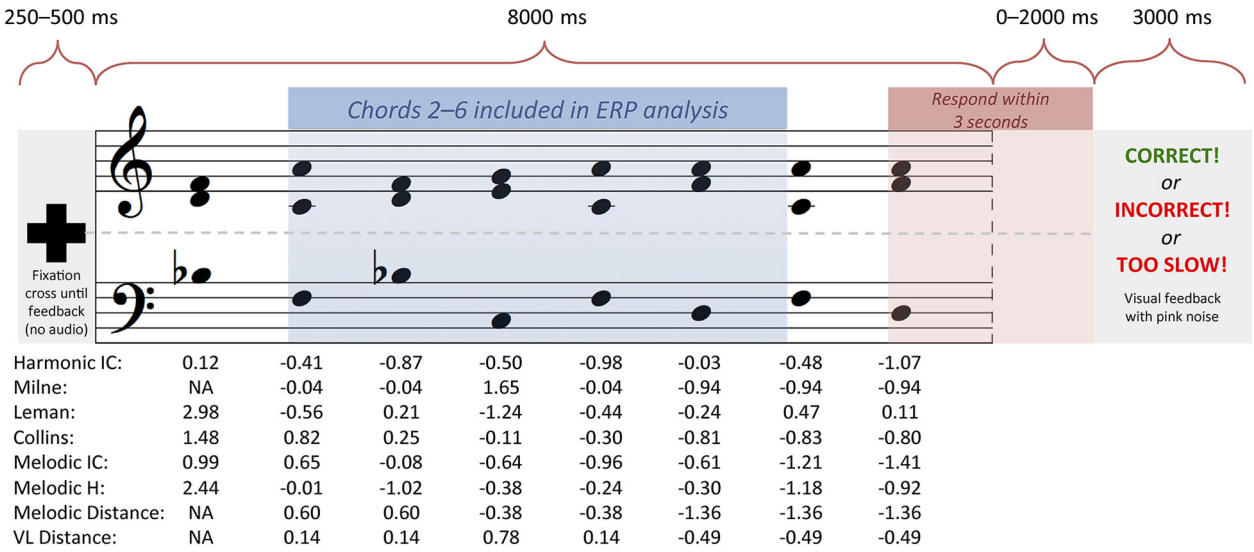
EEG Apparatus

Participants' EEG was acquired while listening to the chord progressions through 64 active scalp electrodes arranged according to the 10-20 system, sampled at 2,048 Hz using a BioSemi ActiveTwo AD Box ADC-12 amplifier (BioSemi, The Netherlands) in an electrostatically shielded room (ETS-Lindgren, Glendale Heights, IL, USA). EEG data were recorded using Actiview (v. 8.6.1).

Stimulus Presentation Apparatus

During the experiment, participants were seated at a table. They listened to stimuli over Bose in-ear noise canceling headphones (Model QC-20). Volume was adjusted to a comfortable level for each participant. Stimuli were presented using Psychtoolbox v. 3.0.12 (Brainard, 1997) in MATLAB. Behavioral responses were collected in MATLAB via a wireless computer keyboard. Instructions and feedback were displayed on a monitor placed at a comfortable viewing distance from the participants.

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021



Harmonic IC:	0.12	-0.41	-0.87	-0.50	-0.98	-0.03	-0.48	-1.07
Milne:	NA	-0.04	-0.04	1.65	-0.04	-0.94	-0.94	-0.94
Leman:	2.98	-0.56	0.21	-1.24	-0.44	-0.24	0.47	0.11
Collins:	1.48	0.82	0.25	-0.11	-0.30	-0.81	-0.83	-0.80
Melodic IC:	0.99	0.65	-0.08	-0.64	-0.96	-0.61	-1.21	-1.41
Melodic H:	2.44	-0.01	-1.02	-0.38	-0.24	-0.30	-1.18	-0.92
Melodic Distance:	NA	0.60	0.60	-0.38	-0.38	-1.36	-1.36	-1.36
VL Distance:	NA	0.14	0.14	0.78	0.14	-0.49	-0.49	-0.49

FIGURE 2. Timeline of a trial, showing 1 out of the 300 chord progressions used in the experiment. Participants responded whether the final chord was louder in one ear, or equally loud in both ears. Feedback did not appear until at least 1,000 ms after the onset of the final chord (even if response times were < 1,000 ms). Each chord has associated model predictions. The model ratings shown here are after z-scoring across the range of values for all chords in the full stimulus set. The inter-stimulus interval was sampled from uniform random distribution over the interval 250–500 ms.

PROCEDURE

After giving informed consent, the participants completed the questionnaires and the experimenters outfitted participants with the EEG cap. Participants were then seated in the shielded room. Experimental instructions were displayed on the screen and read by an experimenter. Participants were instructed to look at a fixation cross on the screen while listening to the chord progressions, and to press a key as quickly and accurately as possible if the final chord of each progression was louder in one ear, and a different key if the final chord was equally loud in both ears. The keys were "K" and "L" on the computer keyboard, and participants were instructed to use their right index finger and right middle finger respectively for the keys. The key and associated finger that corresponded to each kind of stimulus (distractors vs. targets) were counterbalanced across participants. Following their response, feedback was displayed on the screen as either "correct!" in green font, "incorrect!" in red font, or "too slow!" (also in red) if participants did not respond within three seconds. If a participant responded in less than one second, i.e., before the audio file finished playing, the audio file finished playing before displaying the feedback. Participants heard pink noise during the feedback period, generated with SoX (http://sox.sourceforge.net/), to avoid sensory memory affecting the following trial, followed by a silent interstimulus interval of 250–500 ms, uniform-randomly jittered, during which a fixation cross was displayed. The next trial then began. See Figure 2 for a graphical depiction of a single trial. Note that Figure 2 diaplys a trial with only three-note chords, whereas some stimuli had more than three notes (for such stimuli, pitch classes were never doubled).

Before beginning the experiment, participants had three practice trials in which they were told the correct answer in advance (to familiarize them with the timing of the trial, and to alert them to what the difference in loudness would sound like). Then, they completed three additional practice trials without being told the correct answer. If participants required further practice following these six trials, they were allowed to practice again.

The 300 chord progressions were divided into six blocks of 50 trials each (randomly for each participant). After each block, a screen instructed participants to take a break, and press a key when they were comfortable and ready to continue with the experiment. Including setup, participation, and a short debriefing session in which we asked participants for comments on the task and explained the purpose of the study, the experimental session took approximately two hours.

DATA ANALYSIS

Behavioral

We collected data on participants' reaction times and accuracy in identifying the presence of interaural lateralization of the final chord in each progression. While it could be possible to analyze behavioral data with respect

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

to our hypotheses (e.g., whether the degree of surprisal according to different models predicts response times), we note that the behavioral task was primarily intended to make sure participants remained alert. We do not analyze or report additional behavioral analyses in the present manuscript. However, we do report descriptive statistics about task performance below.

EEG Preprocessing

EEG data were processed using the MATLAB toolbox EEGLAB v.13.5.4b (Delorme & Makeig, 2004). We did not include data from five participants: one participant did not complete the experiment, another closed their eyes during the experiment, and three had poor recording quality. Data were low-pass filtered using a finite impulse response (FIR) filter with passband edge 50.00 Hz (6dB/octave, cutoff frequency = 56.23 Hz), downsampled from 2,048 Hz to 128 Hz, high-pass filtered with an FIR filter with passband edge 0.05 Hz (-6dB/octave, cutoff frequency = 0.025 Hz), and epoched from 250 ms before the onset of the first chord of the progression to 8,000 ms after that onset (which is 1000 ms after the onset of the eighth chord of the progression). Bad epochs were rejected by eye under the criterion of severe, obvious non-cognitive artifacts (excluding eye movements). Bad channels were identified by eye by examining their power spectral density to find very noisy electrodes, as well as using EEGLAB's automatic detection algorithm using the kurtosis measure (> 5 standard deviations); the electrodes selected for rejection were interpolated spherically. We then used independent components analysis (ICA), computed using the Infomax algorithm (EEGLAB's runica function), to further identify artifacts, and rejected components corresponding to eye movements and other muscular activity by eye, only when it was possible to cleanly separate them from cognitive activity (i.e., if the power spectral density of the independent component did not resemble the smooth decrease characteristic of artifacts, we left it in). The number of rejected independent components ranged from 0 to 4 ($M = 1.16$, $SD = 1.14$). After rejecting these artifacts, we used EEGLAB's automatic epoch rejection algorithm (using probability as a measure, > 5 *SD*s within channel and across all channels). We then re-referenced the data to the common average reference. Out of 7,500 possible trials to include (25 included participants x 300 progressions), we rejected 1,323 of them (17.64%).

Each long epoch was then further divided into eight 1,000 ms epochs, one for each chord within each progression. Each 1,000 ms epoch had a 50 ms baseline correction. Each participant's EEG data was normalized using *z*-scores over the entire distribution of all samples from all epochs. Then, the EEG data for each specific chord in the experiment were averaged over all participants (excluding rejected trials), yielding 2,400 ERPs (300 trials x eight chords/trial). Because some chord progression epochs had been rejected, this means that a given individual chord's ERP would be averaged over a maximum of 25 participants. Finally—keeping in mind that each progression had eight chords—we did not include chords in positions 1, 7, or 8 in our final analysis. The reason for this is because we had reason to believe these chord positions would systematically differ from the others in ways that did not pertain to our hypotheses: chord 1 would be preceded by a longer period of silence potentially leading to heightened sensory ERP components and reduced cognitive components due to the lack of a musical context; chord 7 could contain contingent negative variation due to participants preparing to respond to the following chord; and chord 8 had altered auditory characteristics as described above and would contain ERP components pertaining to target detection (such as a P3b), as well as related to the movement of the response itself. Thus, 1,500 of the 2,400 ERPs were included in the final analysis.

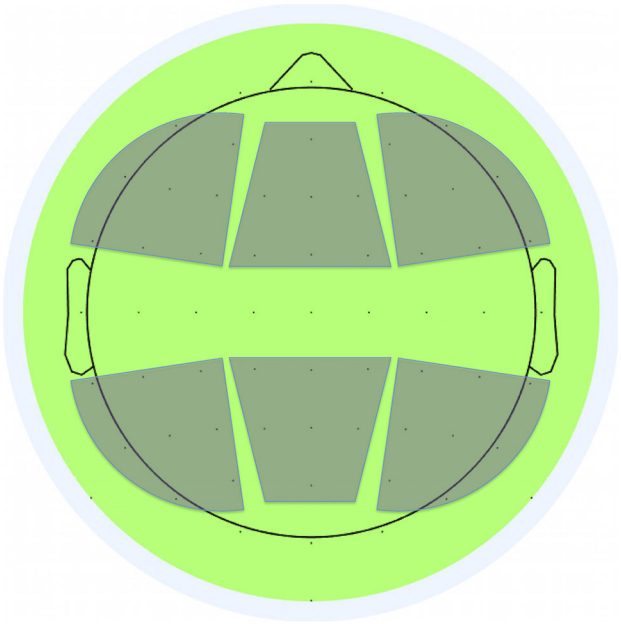


FIGURE 3. Scalp ROIs. We divided the scalp into six spatial ROIs (anterior left, anterior middle, anterior right, posterior left, posterior middle, and posterior right).

ERP Component Calculation

As illustrated in Figure 3, we divided the scalp into six spatial regions of interest (ROIs): anterior left

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

(electrodes AF3, AF7, F3, F5, F7, FC3, FC5, and FC7), anterior middle (electrodes AFz, Fz, F1, F2, FCz, FC1, and FC2), anterior right (electrodes AF4, AF8, F4, F6, F8, FC4, FC6, and FC8), posterior left (electrodes CP3, CP5, CP7, P3, P5, P7, PO3, and PO7), posterior middle (electrodes CPz, CP1, CP2, Pz, P1, P2, and POz), and posterior right (electrodes CP4, CP6, CP8, P4, P6, P8, PO4, and PO8).

We also predesignated four latency ranges: 150–300 ms (ERAN), 250–450 ms (P3a), 450–600 ms (N5), and 600–900 ms (P600). We are aware that our latency ranges for the ERAN and P3a overlap. Predefining latency ranges allows us to make our analysis more confirmatory than exploratory. We note that variations in an experimental task can shift ERP component latencies; while our task differs from previous EEG studies of musical syntax, we argue that it is similar enough to investigate similar latency ranges. We also note that restricting the ERAN range to 150–250 ms (as opposed to the 150–300 ms range that we report below) does not change which effects bear out as statistically significant. For each latency range, we calculated the mean amplitude for each scalp ROI (yielding 4 x 6 = 24 values per epoch). These values served as the dependent variables in our analysis. Each of these values had an associated laterality level (left, middle, or right), and an associated anterior-posterior level (anterior or posterior); these location factors were used as predictors in our statistical analysis. Each epoch (i.e., each individual chord presented over the course of the experiment) also had a set of harmonic and melodic model predictions: Leman, Milne, Collins, Harmonic IC, Melodic entropy, Melodic IC, Melodic distance, and VL Distance. We normalized each predictor's ratings across its own distribution of values. In addition to the categorical topographic location factors, these model ratings served as continuous predictor variables in our analysis.

Note that ERP amplitudes are often calculated from a difference wave; in the case of the ERAN, for example, the difference between a syntactically irregular stimulus and a control stimulus. In our study, there was no single control condition, but rather a set of stimuli with a range of values from our different model predictions. Thus, instead of looking for categorical differences (and conducting a repeated measures ANOVA with categorical variables), we looked for a linear relationship between the component amplitudes (as defined by their scalp ROIs and latency ranges) and the model predictions (see Statistical Models).

STATISTICAL MODELS

We constructed linear mixed-effects models for each component, and for each set of model predictions (melodic and harmonic). Note that we prespecified the anteriority: i.e., for the early and middle latency ranges (corresponding to hypotheses about the ERAN, N5, and P3a components), we only examined frontal regions, and for the late latency range (corresponding to hypotheses about the P600), we only examined posterior regions. For the analysis of the harmonic models, the fixed effects were laterality (left, middle, or right) and the four models (Harmonic IC, Leman, Collins, and Milne). The random effect was stimulus number, a unique identifier for each of the 1,500 chords used in the analysis. This random effect was necessary because there were multiple observations in our models corresponding to the different scalp ROIs (three levels of laterality for each of anterior and posterior). The statistical models for the melodic model predictions were similar but used the melodic model predictions instead of the harmonic ones. In Wilkinson notation, the model with the harmonic predictors for each of the three specified latency ranges was as follows:

$$EEGamplitude \sim (HarmonicIC + Leman + Collins + Milne) * laterality + (1|stimNum) \quad (1)$$

The model with the melodic predictors was as follows:

$$EEGamplitude \sim (MelodicIC + Melodicentropy + Melodicdistance + VLdistance) * laterality + (1|stimNum) \quad (2)$$

We constructed a harmonic model and a melodic model for each of the four ERP time ranges (ERAN, N5, P600, and P3a). For example, for the ERAN component model with the harmonic predictors, "EEG amplitude" corresponded to the average amplitude in the 150–300 ms range over anterior regions of the scalp. As another example, the melodic model for the P600 component used the melodic predictors to predict the average EEG amplitude between 600–900 ms over posterior regions.

We considered including the ordinal position of the chord in the progression (2–6 after rejecting chords 1, 7, and 8; see ERP Component Calculation above) as a fixed effect in our model. On the one hand, there could be differences in the ERPs corresponding to the chord position, independent of the model predictions, so chord position could be a potential confound. On the other hand, this context was precisely what we wished to measure, and in fact, some models were significantly correlated (using Spearman tests) with chord position: Harmonic IC ($r_s = -.29$, $p < .001$), Collins ($r_s = -.72$,

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

$p < .001$), Melodic IC ($r_s = -.15$, $p < .001$), and Melodic entropy ($r_s = -.60$, $p < .001$). We wanted to know if any of the models predict ERP amplitudes independent of *mere* chord position, despite being partially determined by it. Thus, we did not include chord position in the initial models, but where effects were present, we checked whether they were explained by chord position in a post hoc analysis that also included chord position as a fixed effect. This way, we could check both whether the models predicted the ERP amplitudes, and whether or not those effects could be explained by an effect of chord position.

Further, where there were interactions between the various models and laterality, we conducted post hoc tests to assess whether there were main effects in the relevant individual laterality region. For example, if there was an interaction between a particular model rating and the middle region, we conducted a post hoc test on the middle region only, and the left region only (because interactions were calculated by comparing against the left region in all cases). For these models, we also included chord position as a factor. These models focus on single regions of interest, obviating the need for a random effect of stimulus number in the model (because each stimulus corresponds to only one dependent variable measurement). Thus, these models constituted multiple linear regression models on specific ROIs. In these post hoc tests, we also included any predictor models that had significant main effects in the multiple linear regression models, as this allowed us to assess whether the effect of chord position mediated the predictive power of the model ratings. Note that for the post hoc multiple linear regression models, we report the Bonferroni corrected p value, multiplying by the number of post hoc tests performed. For the sake of brevity and clarity, if a statistical model returned no significant effects of the model predictions, we note this in the results, but do not report the full regression table. The full data and analysis tools are available on the Open Science Framework (OSF) page associated with this manuscript (see Appendix) should the reader wish to review these full tables.

INTERQUARTILE COMPARISONS

Our main statistical analysis might appear to lack a control condition: there are no categorically "expected" stimuli against which to compare "unexpected" stimuli. However, in place of a categorical control condition, we compare across a range of levels of unexpectedness allowing us to test multiple points along this range (for each model) rather than only the extremes. This approach might appear to lack statistical sensitivity: with interstimulus variation, the lack of a significant regression coefficient might be the result of too much variance across measurements of ERP components for individual stimuli. But, in such a situation, regression analyses are at least as powerful (if not more so) in detecting an effect than comparing the extreme ends of the range. In other words, our regression approach is at least as powerful as categorically comparing the stimuli for the top and bottom parts of the model ranges with categorical statistical tests. We demonstrate this point with a simulation script included on this project's OSF page.[2]

Still, as an additional complementary analysis, we include here an analysis of interquartile comparisons for each of the predictor metrics. We grouped the top and bottom quartile of stimuli according to each predictor metric (i.e., the stimuli in the top versus bottom 25% of values for each model), and contrasted the ERAN, N5, P600, and P3a amplitudes evoked by the stimuli in these groups using independent samples t-tests. This effectively considers all top quartile stimuli for a given metric as the same condition and increases the reliability because we can reduce the within-subject variability by collapsing together multiple trials. In other words, we can compare stimuli with generally "high" and "low" values for each of the metrics. In total, this meant we conducted three t-tests for each of the eight metrics (four harmonic models plus four melodic models), one for each of the laterality levels (the three anterior regions for the ERAN, N5, and P3a, and the three posterior regions for the P600). We Bonferroni-corrected the p values by a factor of 24 (8 metrics x 3 ROIs). This is conservative, but the risk of type-I errors is high here given how many comparisons we are making. We report the corrected p values only for significantly different comparisons.

VISUALIZATION ANALYSES

For visualization purposes, we constructed topography regression models. For these, we used the average values in each of the latency ranges for each individual electrode, giving us 1,500 1-by-64 vectors (one for each chord) per participant (again, omitting rejected trials). We constructed linear mixed-effects models for each electrode using the same model predictions as in our main analysis. This allowed us to visualize the contribution of each predictor rating to each electrode's amplitude and plot the beta coefficients topographically. We did this for both the harmonic and the melodic predictors. We thresholded these plots to include only

[2] See the script entitled "public_Regression_v_ttest.m" at https://osf.io/xw9v5/

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

TABLE 1. *ERAN: Early Latency Range (150 ms–300 ms) Model for Anterior ROIs (Linear Mixed Effects Model)*

Predictor Name	β	SE	*t-stat*	*df*	*p*	Sig
(Intercept)	0.04	0.00	12.70	4485	< .001	***
laterality_right	0.02	< 0.01	4.67	4485	< .001	***
laterality_middle	0.07	< 0.01	17.39	4485	< .001	***
Harmonic_IC	< 0.01	< 0.01	0.65	4485	.515	
Milne	0.01	< 0.01	1.80	4485	.072	
Leman	0.01	0.01	1.22	4485	.222	
Collins	< 0.01	< 0.01	−0.59	4485	.556	
laterality_right: Harmonic_IC	< 0.01	< 0.01	0.79	4485	.429	
laterality_middle: Harmonic_IC	< 0.01	< 0.01	−0.07	4485	.941	
laterality_right: Milne	< 0.01	< 0.01	−0.14	4485	.887	
laterality_middle: Milne	< 0.01	< 0.01	−1.25	4485	.210	
laterality_right: Leman	−0.01	0.01	−1.66	4485	.098	
laterality_middle: Leman	−0.02	0.01	−2.77	4485	.006	**
laterality_right: Collins	0.01	< 0.01	1.57	4485	.117	
laterality_middle: Collins	0.01	< 0.01	1.73	4485	.084	

ERAN: Early Latency Range (150 ms–300 ms)
Post hoc Model for Anterior-Middle ROI (Multiple Linear Regression Model)

Predictor Name	β	SE	*t-stat*	$p^{\dagger}$	Sig
(Intercept)	0.12	0.01	19.84	<.001	***
Leman	−0.01	0.01	−2.81	.010	*
chordNum_3	< 0.01	0.01	−0.23	1.632†	
chordNum_4	−0.02	0.01	−1.98	.095	
chordNum_5	−0.01	0.01	−1.78	.150	
chordNum_6	−0.01	0.01	−1.24	.428	

$F(5, 1494) = 2.75$, $p = .018$, *Adjusted* $R^2 = .01$

Note. A post hoc model on the anterior-left ROI showed no significant main effects of the Leman predictor ($p = .850$ uncorrected).
* $p < .05$, ** $p < .01$, *** $p < .001$.
† p values for the post hoc model are Bonferroni-corrected, making some p values > 1.

the beta coefficients with p values less than .01, with the rest of the coefficients displayed as 0.

Finally, as another visualization of the data, we calculated the average ERP for the anterior 3 scalp ROIs and the posterior 3 scalp ROIs for the top and bottom quartile of each metric. This gave us an ERP corresponding to relatively high and low values of each metric so that the difference in the ERPs can be more familiarly visualized (i.e., it contrasts "high" and "low" expectedness conditions).

Results

BEHAVIORAL

All participants responded to the stimuli throughout the task, indicating that they were awake and alert throughout the experiment. The task was relatively difficult (range: 49.00%–94.00% correct, $M = 70.43\%$, $SD = 15.99\%$). Four participants were below 50% accuracy (150 correct out of 300) and eight failed to exceed chance performance past a 95% probability threshold (165 correct out of 300) calculated from a binomial distribution ($n = 300$, $p = .50$). Mean reaction times varied across participants (range: 524.43-1819.71 ms, $M = 1110.57$ ms, $SD = 373.62$ ms).

ERAN LATENCY RANGE RESULTS

Harmonic Models

Table 1 reports the full results of the linear mixed effects model predicting early latency range amplitudes (150 ms–300 ms) corresponding to the ERAN amplitude in the anterior scalp ROIs from the various harmonic models. There was a significant interaction between the Leman rating and laterality-middle, $\beta = -0.02 \pm 0.01$, $p = .006$. None of the other models predicted amplitude differences in this latency range. A post hoc linear model was constructed with the Leman rating and chord position as predictors for both the anterior-left and anterior-middle regions. There were no effects of the model for the anterior-left region. The

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

TABLE 2. *N5: Middle Latency Range (450 ms–600 ms) Model for Anterior ROIs*

Predictor Name	β	SE	*t-stat*	*df*	*p*	Sig
(Intercept)	< 0.01	< 0.01	0.24	4485	.814	
laterality_right	−0.01	0.01	−1.47	4485	.143	
laterality_middle	−0.05	0.01	−10.54	4485	< .001	***
Harmonic_IC	0.01	< 0.01	1.85	4485	.064	
Milne	< 0.01	< 0.01	0.93	4485	.354	
Leman	0.01	0.01	1.21	4485	.225	
Collins	< 0.01	< 0.01	−1.17	4485	.243	
laterality_right: Harmonic_IC	< 0.01	< 0.01	−0.80	4485	.424	
laterality_middle: Harmonic_IC	< 0.01	< 0.01	−0.04	4485	.965	
laterality_right: Milne	< 0.01	0.01	−0.02	4485	.981	
laterality_middle: Milne	−0.01	0.01	−1.34	4485	.179	
laterality_right: Leman	−0.01	0.01	−1.31	4485	.189	
laterality_middle: Leman	−0.02	0.01	−2.00	4485	.046	*
laterality_right: Collins	0.01	0.01	2.53	4485	.012	*
laterality_middle: Collins	0.02	0.01	2.98	4485	.003	**

Note. Post hoc models on the anterior-right, anterior-middle, and anterior-left ROIs showed no significant main effects of the Leman predictor ($p = .250$, $p = .315$, and $p = .367$ uncorrected, respectively) nor the Collins predictor ($p = .853$, $p = .395$, and $p = .362$ uncorrected, respectively).
* $p < .05$, ** $p < .01$, *** $p < .001$.

anterior-middle region showed a significant effect of the Leman rating, $\beta = 0.01 \pm 0.01$, $p = .010$ (Bonferroni corrected). Higher Leman ratings (less surprising chords) showed greater negativity in this latency range, i.e., more surprising chords predicted relatively greater positivity.

Melodic Models

There were no significant main effects of or interactions with the melodic predictor variables. None of the melodic predictors predicted amplitude differences in this latency range.

N5 LATENCY RANGE RESULTS

Harmonic Models

Table 2 reports the full results of the linear mixed effects model predicting middle latency range amplitudes (450 ms–600 ms) in the anterior scalp ROIs from the various harmonic models. There was a significant interaction between the Leman rating and laterality-middle, $\beta = -0.02 \pm 0.01$, $p = .046$, as well as between the Collins rating and laterality-middle, $\beta = 0.02 \pm 0.01$, $p = .003$. There was also a significant interaction between laterality-right and Collins, $\beta = 0.01 \pm 0.01$, $p = .012$. Post hoc linear models were constructed with the Leman and Collins ratings, also including chord position, for the anterior-left, anterior-middle, and anterior-right ROIs. None of these post hoc models showed significant main effects for any of the predictor ratings.

Melodic Models

There were no significant main effects of or interactions with the melodic predictor variables. None of the melodic predictors predicted amplitude differences in this latency range.

P600 LATENCY RANGE RESULTS

Harmonic Models

Table 3 reports the full results of the linear mixed effects model predicting late latency range amplitudes (600 ms–900 ms) in the posterior scalp ROIs from the various harmonic models. There was a significant main effect of the Collins rating, $\beta = -0.01 \pm 0.004$, $p = .011$, and a significant interaction between the Leman rating and laterality-middle (compared to left), $\beta = -0.02 \pm 0.01$, $p = .047$. Post hoc linear models examining the posterior-middle and posterior-left ROIs separately with the Collins and Leman predictors were constructed along with chord position as a predictor. The posterior-left model did not show any significant main effects of the predictors (though the Leman rating was marginally significant after Bonferroni correction, $p = .068$). For the posterior-middle model, there were no significant effects for the predictor ratings. Thus, the Collins model predicted P600 amplitude (more surprising chords had higher P600 amplitudes), but this effect is explained by chord position.

Melodic Models

There were no significant main effects of or interactions with the melodic predictor variables. None of the melodic predictors predicted amplitude differences in this latency range.

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

TABLE 3. *P600: Late Latency Range (600 ms–900 ms) Model for Posterior ROIs*

Predictor Name	β	SE	*t-stat*	*df*	*p*	Sig
(Intercept)	0.02	< 0.01	5.41	4485	< .001	***
laterality_right	−0.01	< 0.01	−2.88	4485	.004	**
laterality_middle	−0.04	< 0.01	−8.00	4485	< .001	***
Harmonic_IC	< 0.01	< 0.01	−0.66	4485	.508	
Milne	< 0.01	< 0.01	−0.49	4485	.626	
Leman	0.01	0.01	1.34	4485	.179	
Collins	−0.01	< 0.01	−2.55	4485	.011	*
laterality_right: Harmonic_IC	< 0.01	< 0.01	0.72	4485	.473	
laterality_middle: Harmonic_IC	< 0.01	< 0.01	0.86	4485	.391	
laterality_right: Milne	< 0.01	< 0.01	0.30	4485	.765	
laterality_middle: Milne	< 0.01	< 0.01	−0.23	4485	.821	
laterality_right: Leman	−0.01	0.01	−1.33	4485	.183	
laterality_middle: Leman	−0.02	0.01	−1.99	4485	.047	*
laterality_right: Collins	< 0.01	< 0.01	0.48	4485	.632	
laterality_middle: Collins	< 0.01	< 0.01	1.00	4485	.319	

Note. Post hoc models on the posterior-middle and posterior-left ROIs showed no significant main effects of the Leman predictor ($p = .601$ and $p = .034$ uncorrected, respectively) nor the Collins predictor ($p = .744$ and $p = .753$ uncorrected, respectively). The Leman predictor for the posterior-left ROI was only marginally significant after Bonferroni correction ($p = .068$).
* $p < .05$, ** $p < .01$, *** $p < .001$.

P3A LATENCY RANGE RESULTS

Harmonic Models

Table 4 reports the full results of the linear mixed effects model predicting the P3a latency range amplitudes (250 ms–450 ms) in the anterior scalp ROIs from the various harmonic models. There was a main effect of Harmonic IC, $\beta = 0.01 \pm 0.003$, $p = .042$. Higher Harmonic IC (more surprising chords) predicted greater positive amplitude in this latency range. There was a significant interaction between the Leman rating and laterality-middle, $\beta = 0.03 \pm 0.01$, $p = .003$, as well as a significant interaction between Collins and laterality-middle, $\beta = 0.01 \pm 0.004$, $p = .006$. There was also a significant interaction between Collins and laterality-right, $\beta = 0.01 \pm 0.004$, $p = .015$.

Post hoc linear models were constructed for the anterior-right, anterior-middle, and anterior-left ROIs with the predictors Harmonic IC, Leman, Collins, and chord position to test these effects at the individual ROIs. The anterior-left and the anterior-right models returned no significant main effects for any of the three predictor ratings. For the anterior-middle model, there was a significant main effect of Harmonic IC, $\beta = 0.01 \pm 0.003$, $p = .003$ after Bonferroni correction. There was also a significant main effect of the Leman predictor, $\beta = -0.017 \pm 0.01$, $p = .009$ after Bonferroni correction. Higher Harmonic IC values (more surprising chords) and lower Leman values (corresponding to more surprising chords for this model) predicted more positivity in the anterior-middle scalp ROI during this latency range. In other words, according to both of these metrics, more surprising chords predicted greater positivity.

Melodic Models

There were no significant main effects of or interactions with the melodic predictor variables. None of the melodic predictors predicted amplitude differences in this latency range.

QUARTILE COMPARISONS

After Bonferroni correction, the only significant quartile comparison was for the P3a component. The P3a amplitudes corresponding to the top quartile of the Harmonic IC predictions had significantly higher amplitudes than the bottom quartile, $t(721) = 4.34$, $p < .001$ (Bonferroni corrected) for the anterior middle ROI. More unpredictable chords according to this model elicited greater ERP amplitudes in the latency range of the P3a. None of the other models predicted any significant difference between the top and bottom quartiles for any of the ERP components, over any of the ROIs.

MODEL VISUALIZATIONS

Topographic Representations

Figure 4 shows topographic representations of the relationship between all of the predictor models (harmonic and melodic) and the amplitudes averaged over each of the four latency ranges. The colors represent beta values from linear mixed effects models run on individual electrodes (as opposed to averaging over scalp ROIs for our

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

TABLE 4. *P3a Latency Range (250 ms–450 ms) Model for Anterior ROIs*

Predictor Name	β	SE	*t-stat*	*df*	*p*	Sig
(Intercept)	0.02	< 0.01	6.01	4485	< .001	***
laterality_right	0.01	< 0.01	2.49	4485	.013	*
laterality_middle	< 0.01	< 0.01	1.01	4485	.313	
Harmonic_IC	0.01	< 0.01	2.03	4485	.042	*
Milne	0.01	< 0.01	1.38	4485	.169	
Leman	0.01	0.01	0.91	4485	.364	
Collins	< 0.01	< 0.01	−1.31	4485	.190	
laterality_right: Harmonic_IC	< 0.01	< 0.01	0.06	4485	.952	
laterality_middle: Harmonic_IC	< 0.01	< 0.01	1.15	4485	.251	
laterality_right: Milne	< 0.01	< 0.01	−0.01	4485	.991	
laterality_middle: Milne	−0.01	< 0.01	−1.68	4485	.092	
laterality_right: Leman	−0.01	0.01	−1.15	4485	.251	
laterality_middle: Leman	−0.03	0.01	−2.96	4485	.003	**
laterality_right: Collins	0.01	< 0.01	2.44	4485	.015	*
laterality_middle: Collins	0.01	< 0.01	2.77	4485	.006	**

P3a Latency Range (250 ms–450 ms)
Post hoc Model for Anterior-Middle ROI

Predictor Name	β	SE	*t-stat*	$p^{\dagger}$	Sig
(Intercept)	0.04	0.01	4.83	< .001	***
Harmonic_IC	0.01	< 0.01	3.33	.003	**
Leman	−0.02	0.01	−2.98	.009	**
Collins	< 0.01	< 0.01	0.12	2.711†	
chordNum_3	< 0.01	0.01	0.33	2.231†	
chordNum_4	−0.01	0.01	−1.39	.495	
chordNum_5	−0.01	0.01	−1.34	.538	
chordNum_6	−0.02	0.01	−1.72	.254	

$F(7,1492) = 4.69$, $p < .001$, *Adjusted* $R^2 = .02$

Note. Post hoc models on the anterior-right and anterior-left ROIs showed no significant main effects of the Harmonic IC model ($p = .154$ and $p = .070$ uncorrected, respectively) the Leman model ($p = .112$ and $p = .920$ uncorrected, respectively) nor the Collins model ($p = .844$ and $p = .334$ uncorrected, respectively).
* $p < .05$, ** $p < .01$, *** $p < .001$.
†*p values for the post hoc model are Bonferroni-corrected.*

main analysis reported above). The beta values are thresholded at $p = .01$. If a beta value did not reach this significance threshold, it is plotted as 0 (corresponding to the green color on the plot). These models include chord position as a predictor.

Quartile ERP Representations

Figure 5 shows the average ERPs for the top and bottom quantile of each of the predictor ratings (harmonic and melodic) along with the resultant difference wave. This plot shows more traditional visualizations of the ERPs, allowing the effects of the predictor ratings to be visualized by contrasting the ERPs from stimuli with high versus low values for each.

Discussion

Surprisingly, our findings indicate that none of our harmonic predictor ratings, cognitive or sensory, predict an ERAN, nor are they predictive of N5 or P600 amplitudes (though the Collins model is predictive of the P600 before controlling for chord position). The Leman model did predict *positivity* in the ERAN time range for more surprising chords, contrary to the hypothesis that more surprising chords should predict *negativity*. However, it is likely this positivity is actually part of the P3a response, considering that the latency ranges for these components overlapped in our analysis, and that there was also a main effect of the Leman model for the P3a component. Harmonic information content, derived from the cognitive model, also significantly predicted P3a activity along with the Leman rating: more surprising chords evoked more anterior positivity in this time range. Finally, it is notable that neither VL-distance or melodic-distance were predictive of any components which shows that features of the melody and voice leading are not predictive of these effects.

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

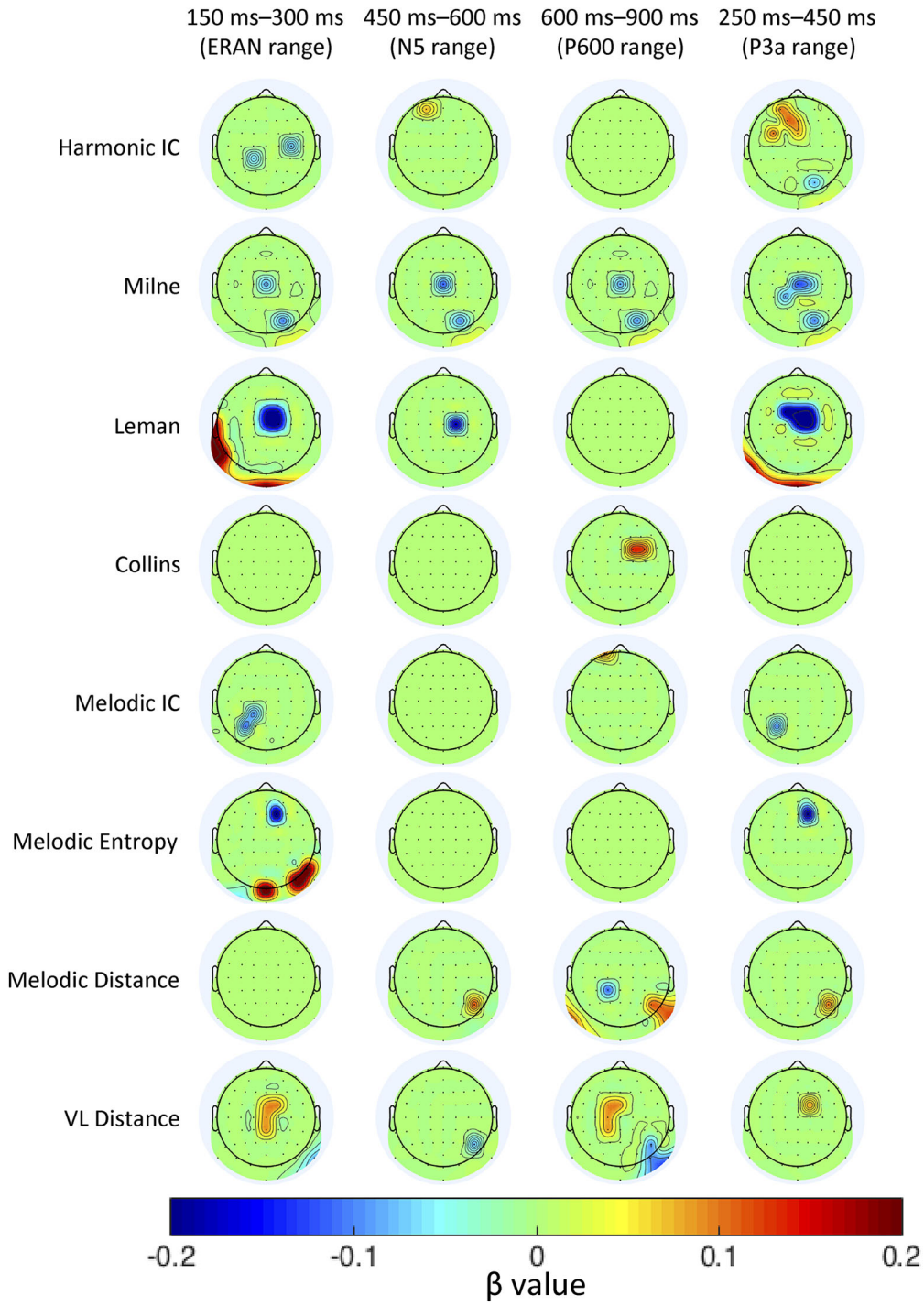


FIGURE 4. Model predictions for individual electrode activity in the four latency ranges. Note that these figures have been thresholded so if beta coefficients have associated $p < .01$, they are displayed as having a value of 0. Note that few electrodes are predictive of amplitude in these latency ranges. Compare against the linear model tables (e.g., Harmonic IC is predictive of P3a activity).

We identified a reliable P3a effect as explained by the IDyOM and Leman models. Although the effect sizes for these models are low ($R^2 = .02$), these results nevertheless suggest a meaningful connection between the P3a and the psychological processes captured by the computational models. Our regression approach means

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

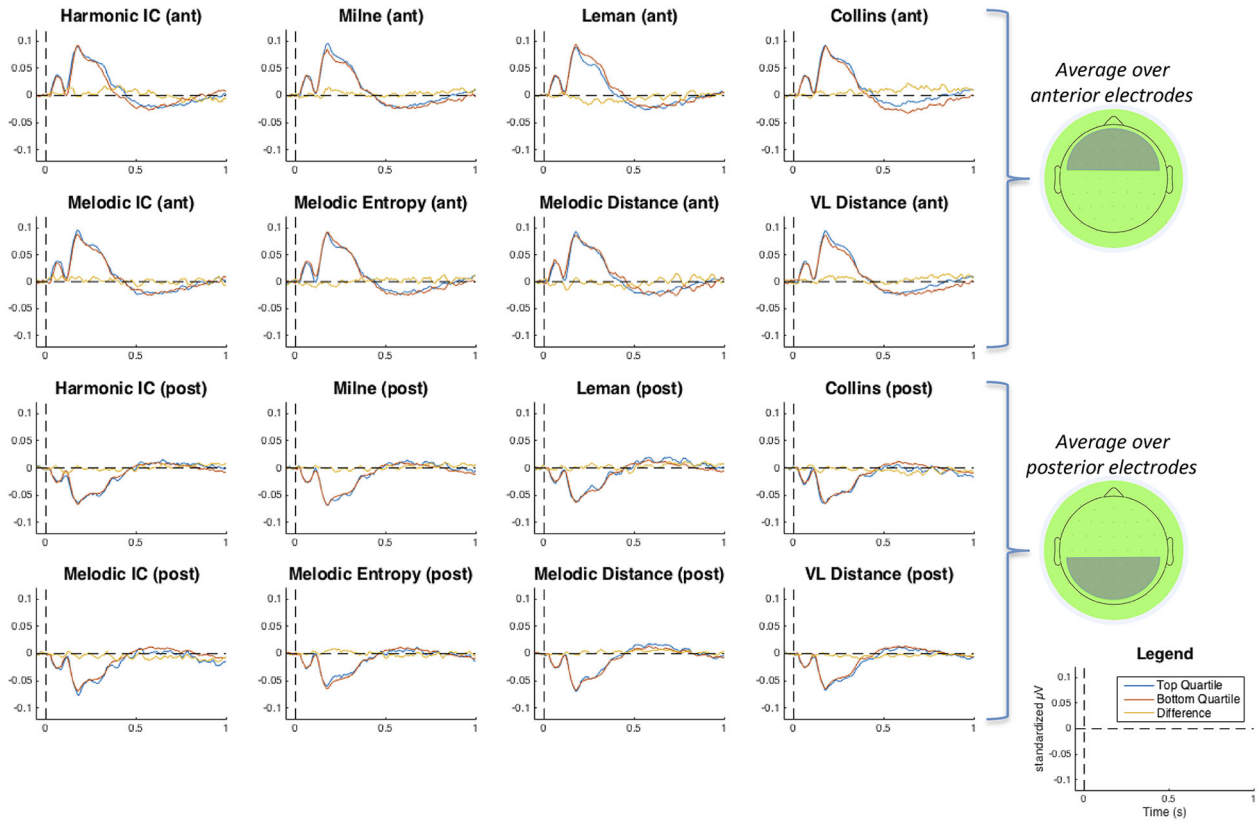


FIGURE 5. Average ERPs for top and bottom quantile of each rating and difference between them. The top half of the plots show all anterior electrodes averaged together, and the bottom half show the posterior electrode average. The difference waves are subtle but show differences matching the findings from the main analysis: the Leman and Harmonic IC predictors show positivity for more surprising chords in the P3a latency range (lower Leman values and higher Harmonic IC values corresopnd to more surprising chords).

that individual estimates for particular stimuli have low reliability, but because we have many stimuli, we are still able to observe statistically significant effects over each model's range. This explains the low R^2 values, but also affirms why a significant effect still points to a connection between neurophysiological ERP responses, and the predictable sensory and cognitive features captured by the models.

In our experimental design, we traded reliability of ERP amplitudes of individual stimuli for generalizability: each of our stimuli was only heard once by each participant, but we had 1,500 different stimuli (300 chord progressions with 5 chords each included in the final analysis). This is an important design feature of our study motivated by a desire to avoid what has been termed the "fixed effect fallacy" (Clark, 1973; Yarkoni, 2019), wherein psychological studies fail to demonstrate that their observed effects would generalize outside the small stimulus set used in their research. In the music expectation ERP literature, this problem manifests as only testing a small number of violations (e.g., using the Neapolitan chord in place of a tonic harmony) and making claims about harmonic expectation violation that apply to other harmonies more generally. In our study, the reliability of our estimates of the ERP amplitudes for individual stimuli may be low because we only repeated each stimulus once, but we are instead able to examine claims about the generalizability over a wide range of stimuli. This allows us to test whether these ERPs are reasonably claimed to be related to harmonic syntax processing *in general* as opposed to being an artifact of some kind of previous experimental designs that treat the stimuli as fixed effects. We were able to detect an effect for the P3a component with this approach, and one that aligns well with the theoretical understanding of this component. P3a components are evoked in response to improbable and unexpected stimuli and can be used as an index of people's statistical knowledge of the world, with unlikely events eliciting a larger P3a (Polich, 2007). The IDyOM-derived

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

Harmonic IC predictor, which is based on long-term knowledge of statistical structure of chord sequences, predicted the P3a, suggesting that participants do indeed have knowledge of statistical transition probabilities as calculated by this model, and that these effects can be evoked across a wide range of harmonic sequences extracted from real-world pop music. In other words, the P3a does appear to generalize over a wide range of stimuli, and the surprisal component of the harmonic IDyOM model predicts its amplitude. The Leman model, which is a sensory model, also predicted P3a amplitude independently, suggesting that surprisal is *also* sensitive to short-term sensory influences at least as captured by this particular model. Thus, with regard to our research question concerning cognitive versus sensory contributions to syntax violations, we conclude that both types of violation contribute to the neural response (specifically, the P3a), and that listeners are sensitive to both sensory and cognitive features of unlikely musical events.

IDyOM predicted P3a amplitudes, but this result does not demonstrate that IDyOM (as configured here) is the best possible predictive model, nor is it shown that human cognitive processes necessarily use the same features (or weighting of features) that IDyOM uses. However, this result does support the conclusion that the computational modeling is at least capable of simulating human neural processing, and the P3a physiological response we observed may provide a potential bridge between information-theoretic measures of musical structure and cognitive information processing of music.

The Collins model did not predict the P3a even though it has a cognitive component as well. The models work in different ways, so although Collins has both sensory and cognitive components, these components are not necessarily identical to other sensory or cognitive models. For example, IDyOM acquires a finite-context model of musical structure through statistical learning of the music to which it is exposed whereas the "cognitive" component in the Collins model reflects tonal stability learned through training on a single artificially constructed melody that cycles through different keys. One explanation for why IDyOM predicts the P3a is that it makes its cognitive predictions in a way more closely aligned with how humans make predictions.

Is the P3a component a neural correlate of syntactic processing? In our study, we focus on the probabilistic aspects of musical syntax that govern the local transitions between harmonies. We argue that cognitive processes that detect unlikely harmonic continuations are engaging in syntactic processing (again, we are limited to discussing finite-state aspects of syntactic processing rather than other less constrained syntactic structures and cognition). We do not explain the mechanistic role played by the neural generators of the P3a, or indeed answer whether the P3a is merely a correlate of another process. Regardless, we can say that our participants were sensitive to harmonic surprisal in certain contexts, and that our cognitive model (IDyOM) predicts those contexts.

The fact that none of the harmonic expectancy models predicted an ERAN, N5, or P600 (after controlling for chord position), urges reflection on the sensory-cognitive debate recently articulated by Bigand et al. (2014). By including statistical surprisal based on long-term learning—a cognitive construct—alongside several sensory models, we were able to examine the relative ability of each of these models to predict ERP components traditionally associated with syntax violations. In our effort to dissociate them, the findings showed that neither type of model predicted the ERAN, N5, or P600 amplitudes. Importantly, we do not conclude that there is no relationship between syntactic violation and these ERP components; rather, that any effect was too small to detect in our study. What this means is that when listening to unpredictable chords in a more ecologically valid listening context, and treating stimulus type as a random effect, ERANs, N5s, and P600s are not robustly predicted by the cognitive or sensory models we included in our analysis, and effects reported in previous studies may not generalize over a wider range of stimuli.

We note that, in contrast to previous studies in the literature, there was variability in the stimuli preceding each time-locked ERP that we analyzed. In other words, the chord preceding the ERP for chord 3 was variable, the chord preceding the ERP for chord 4 was variable, etc. In many similar past studies, the stimuli preceding the time-locked ERP were controlled. We argue that, despite this variability, the stimuli are similar in most acoustic characteristics (e.g., timbre and amplitude envelope). What differs is the expectations that the different chords induce, and this variability is captured by our sensory and cognitive models.

Another potential critique of our analysis is how we define ERP components in our analysis, and how we predefine our latency ranges. We predefined our latency ranges based on the existing literature, and quantified ERP component amplitudes by averaging EEG samples within those ranges. This has advantages: it makes our analysis more confirmatory than exploratory and facilitates comparison with previous findings. However, a latency range, of course, is not the same as an ERP

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

component—EEG amplitude samples within a latency range will be affected by multiple ERP components (i.e., multiple neuroelectrical events propogating to the scalp). So, to quantify an ERP component based on an average within a prespecified latency range potentially mixes together different components. The traditional way to avoid this issue is to compute difference waves, which are intended to cancel out variance from ERP components unrelated to the task. By analogy, in our regression approach, we argue that such variance in other ERP components should not vary systematically with the predictors in our regression analysis. True, such ERP components unrelated to our task may increase the noise in our data (the same noise that difference waves attempt to cancel), lowering the R^2 values, but we are concerned primarily with the question of whether the models can predict significant variance in the ERP data. The results show significant proportions of amplitude variance in the P3a time window are predicted by the IDyOM and Leman models.

As for the ROI latency ranges we chose, one specific concern may be that the ERAN range overlaps with the P3a range. We re-ran our analysis with non-overlapping ranges (ERAN: 150 ms–250 ms, P3a: 250 ms–400 ms, with the N5 and P600 ranges remaining the same). The results returned the same significant effects. Readers can use our online materials on our OSF page to construct models with latency ranges of their own choosing (see Appendix).

In an alternative interquartile comparison taking the top and bottom 75 stimuli, we still did not detect an effect for the ERAN, N5, or P600. These ERP components might thus be more dependent on the pecularities of laboratory listening environments than previously thought, urging a re-evaluation of the traditional paradigm of syntax violation in music cognition research. It is possible that the specific stimuli used in previous experiments evoke these ERPs for some other reason than harmonic predictability per se because if it is indeed the degree of unpredictability that is truly driving these effects, they should generalize over other stimuli with varying predictability. Instead, it is possible that in traditional paradigms that have limited types of harmonic expectation violation at limited locations within a chord progression stimulus, some additional pecularity of this listening context interacts with harmonic expectation violation to produce these effects. This could include knowing when to expect a possible violation and paying more attention (after all, the ERAN amplitude is sensitive to attentional load, Loui et al., 2005). By contrast, in our study, there is no information that could lead a listener to expect violations to appear at any given chord position. It could also be due to some other specific feature of the chords typically used to evoke surprise. This latter interpretation seems unlikely—could there really be something special about Neapolitan chords, or V/V chords?—but should not be ruled out *a priori*. For example, substituting a Neapolitan chord for a tonic chord is extremely unlikely; the stimuli in the present study did not include such extreme cases that are nigh impossible in real music. Therefore, it is possible that traditional ERP effects of harmonic expectation violation may only be evoked by extreme ecologically unlikely stimulus conditions, which casts doubt on the ability to make general claims about the association of such neural signals and harmonic syntax *generally*.

By contrast, the present results suggest that more subtle variations in harmonic predictability present within real music reflect combined influence of cognitive and sensory influences on expectation and can be detected reliably via the P3a component. Future research should try to understand in greater depth the conditions under which different neural responses to expectation violation emerge—including expanding the range of types of stimuli and listening contexts—and the cognitive processes to which these neural responses correspond.

Author Note

AG was supported by the Presidential Scholars in Society and Neuroscience Program and the Music, Cognition, and the Brain Initiative, and was partially supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-15-2-0074. AG is now at Indiana University. PH was supported by a doctoral studentship from the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1). PH is now at the University of Cambridge. The authors declare no conflicts of interest. The authors wish to acknowledge Mari Tervaniemi and the three anonymous reviewers for their thorough and insightful consideration of our work, which helped us strengthen our arguments and their presentation.

Correspondence concerning this article should be addressed to Andrew Goldman, Simon Center, Room 225B, 1201 E 3rd Street, Bloomington, IN 47405. E-mail: andrewjgoldman@gmail.com

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

References

Albrecht, J., & Shanahan, D. (2013). The use of large corpora to train a new type of key-finding algorithm: An improved treatment of the minor mode. *Music Perception*, *31*(1), 59–67. https://doi.org/10.1525/mp.2013.31.1.59

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. https://doi.org/10.1037/0096-1523.21.6.1278

Bigand, E., Delbe, C., Poulin-Charronnat, B., Leman, M., & Tillmann, B. (2014). Empirical evidence for musical syntax processing? Computer simulations reveal the contribution of auditory short-term memory. *Frontiers in Systems Neuroscience, 8*, 94. https://doi.org/10.3389/fnsys.2014.00094

Bigand, E., Poulin, B., Tillmann, B., Madurell, F., & D'Adamo, D. A. (2003). Sensory versus cognitive components in harmonic priming. *Journal of Experimental Psychology, 29*(1), 159–171. https://doi.org/10.1037/0096-1523.29.1.159

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Bunton, S. (1997). Semantically motivated improvements for PPM variants. *The Computer Journal*, *40*(2/3), 76–93. https://doi.org/10.1093/comjnl/40.2_and_3.76

Burgoyne, J. A., Wild, J., & Fujinaga, I. (2011). An expert ground-truth set for audio chord recognition and music analysis. In A. Klapuri & C. Leider (Eds.), *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 633–638). University of Miami.

Carrus, E., Pearce, M. T., & Bhattacharya, J. (2013). Melodic pitch expectation interacts with neural responses to syntactic but not semantic violations. *Cortex*, *49*, 2186–2200. https://doi.org/10.1016/j.cortex.2012.08.024.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Cleary, J. G., & Teahan, W. J. (1997). Unbounded length contexts for PPM. *The Computer Journal*, *40*(2/3), 67–75. https://doi.org/10.1093/comjnl/40.2_and_3.67

Cleary, J., & Witten, I. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, *32*(4), 396–402. https://doi.org/10.1109/TCOM.1984.1096090

Collins, T., Tillmann, B., Barrett, F. S., Delbé, C., & Janata, P. (2014). A combined model of sensory and cognitive representations underlying tonal expectations in music: From audio signals to behavior. *Psychological Review, 121*(1), 33–65. https://doi.org/10.1037/a0034695

Comerchero, M. D., & Polich, J. (1999). P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology, 110*(1), 24–30. https://doi.org/10.1016/S0168-5597(98)00033-1

Conklin, D., & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, *24*(1), 51–73. https://doi.org/10.1080/09298219508570672

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Egermann, H., Pearce, M. T., Wiggins, G., & McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective and Behavioural Neuroscience*, *13*, 533–553. https://doi.org/10.3758/s13415-013-0161-y

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. https://doi.org/10.1371/journal.pone.0076600

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91, 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Hansen, N. C., & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology*, *5*, 1052. https://doi.org/10.3389/fpsyg.2014.01052

Hansen, N. C., Vuust, P., & Pearce, M. (2016). " If you have to ask, you'll never know": Effects of specialised stylistic expertise on predictive processing of music. *PLoS One*, *11*(10), e0163584. https://doi.org/10.1371/journal.pone.0163584

Harrison, P. M. C., & Pearce, M. T. (2018). Dissociating sensory and cognitive theories of harmony perception through computational modeling. In R. Parncutt & S. Sattmann (Eds.), *Proceedings of ICMPC15/ESCOM10* (pp. 194–199). University of Graz. https://doi.org/10.31234/osf.io/wgjyv

Harrison, P. M. C., & Pearce, M. T. (2020). Representing harmony in computational music cognition. *PsyArXiv*. https://doi.org/10.31234/osf.io/xswp4

Huron, D. B. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT Press.

Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B., & Bharucha, J. J. (2002). The cortical topography of tonal structures underlying Western music. *Science*, *298*(5601), 2167–2170. https://doi.org/10.1126/science.1076262

Kim, S. G., Kim, J. S., & Chung, C. K. (2011). The effect of conditional probability of chord progression on brain response: An MEG study. *PloS one*, *6*(2), e17337. https://doi.org/10.1371/journal.pone.0017337

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

Koelsch, S. (2009). Music-syntactic processing and auditory memory: Similarities and differences between ERAN and MMN. *Psychophysiology, 46*, 179–190. https://doi.org/10.1111/j.1469-8986.2008.00752.x

Koelsch, S. (2011). Towards a neural basis of processing musical semantics. *Physics of Life Reviews, 8*, 89–105. https://doi.org/10.1016/j.plrev.2011.04.004

Koelsch, S., Gunter, T., Friederici, A. D., & Schröger, E. (2000). Brain indices of music processing: "Nonmusicians" are musical. *Journal of Cognitive Neuroscience, 12*(3), 520–541. https://doi.org/10.1162/089892900562183

Koelsch, S., & Jentschke, S. (2010). Differences in electric brain responses to melodies and chords. *Journal of Cognitive Neuroscience*, *22*(10), 2251–2262. https://doi.org/10.1162/jocn.2009.21338

Koelsch, S., Jentschke, S., Sammler, D., & Mietchen, D. (2007). Untangling syntactic and sensory processing: An ERP study of music perception. *Psychophysiology, 44*, 476–490. https://doi.org/10.1111/j.1469-8986.2007.00517.x

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), e2631. https://doi.org/10.1371/journal.pone.0002631

Koelsch, S., Schmidt, B. H., & Kansok, J. (2002). Effects of musical expertise on the early right anterior negativity: An event-related brain potential study. *Psychophysiology*, *39*(5), 657–663. https://doi.org/10.1017/S0048577202010508

Kohonen, T. (1995). *Self-organizing maps*. Springer.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334–368. https://doi.org/10.1037/0033-295X.89.4.334

Kunert, R., Willems, R. M., Casasanto, D., Patel, A. D., & Hagoort, P. (2015). Music and language syntax interact in Broca's area: An fMRI Study. *PLoS ONE, 10*(11), e0141069. https://doi.org/10.1371/journal.pone.0141069

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205. https://doi.org/10.1126/science.7350657

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. https://doi.org/10.1016/j.brainres.2007.01.049

Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception, 17*(4), 481–509. https://doi.org/10.2307/40285830

Leman, M., Lesaffre, M., & Tanghe, K. (2001). Introduction to the IPEM toolbox for perception-based music analysis. *Mikropolyphonie-The Online Contemporary Music Journal, 7*.

Leman, M., Lesaffre, M., & Tanghe, K. (2005). IPEM toolbox for perception-based music analysis (Version 1.02). Available at: http://www.ipem.ugent.be/Toolbox/index.html.

Loui, P., Grent-'t-Jong, T., Torpey, D., & Woldorff, M. (2005). Effects of attention on the neural processing of harmonic syntax in Western music. *Cognitive Brain Research, 25*, 678–687. https://doi.org/10.1016/j.cogbrainres.2005.08.019

Loui, P., Wu, E. H., Wessel, D. L., & Knight, R. T. (2009). A generalized mechanism for perception of pitch patterns. *Journal of Neuroscience*, *29*(2), 454–459. https://doi.org/10.1523/JNEUROSCI.4503-08.2009

Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D. (2001). Musical syntax is processed in Broca's area: An MEG study. *Nature Neuroscience, 4*(5), 540–545. https://doi.org/10.1038/87502

Milne, A. J., & Holland, S. (2016). Empirically testing Tonnetz, voice-leading, and spectral models of perceived triadic distance. *Journal of Mathematics and Music*, *10*(1), 59–85. https://doi.org/10.1080/17459737.2016.1152517

Milne, A. J., Laney, R., & Sharp, D. B. (2015). A spectral pitch class model of the probe tone data and scalic tonality. *Music Perception*, *32*(4), 364–393. https://doi.org/10.1525/mp.2015.32.4.364

Milne, A. J., Laney, R., & Sharp, D. B. (2016). Testing a spectral model of tonal affinity with microtonal melodies and inharmonic spectra. *Musicae Scientiae*, *20*(4), 465–494. https://doi.org/10.1177/1029864915622682

Milne, A. J., Sethares, W. A., Laney, R., & Sharp, D. B. (2011). Modelling the similarity of pitch collections with expectation tensors. *Journal of Mathematics and Music, 5*(1), 1–20. https://doi.org/10.1080/17459737.2011.573678

Miranda, R. A., & Ullman, M. T. (2007). Double dissociation between rules and memory in music: An event-related potential study. *Neuroimage*, *38*(2), 331–345. https://doi.org/10.1016/j.neuroimage.2007.07.034

Moffat, A. (1990). Implementing the PPM data compression scheme. *IEEE Transactions on communications*, *38*(11), 1917–1921. https://doi.org/10.1109/26.61469

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE, 9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*(6615), 432–434. https://doi.org/10.1038/385432a0

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*, 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

Nittono, H., Bito, T., Hayashi, M., Sakata, S., & Hori, T. (2000). Event-related potentials elicited by wrong terminal notes: Effects of temporal disruption. *Biological Psychology, 52*(1), 1–16. https://doi.org/10.1016/S0301-0511(99)00042-3

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113.

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

Omigie, D., Pearce, M. T., Lehongre, K., Hasboun, D., Navarro, V., Adam, C., & Samson, S. (2019). Intracranial recordings and computational modeling of music reveal the time course of prediction error signaling in frontal and temporal cortices. *Journal of Cognitive Neuroscience*, *31*, 855–873. https://doi.org/10.1162/jocn_a_01388

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596X(92)90039-Z

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. https://doi.org/10.1080/01690969308407584

Paller, K. A., McCarthy, G., & Wood, C. C. (1992). Event-related potentials elicited by deviant endings to melodies. *Psychophysiology*, *29*(2), 202–206. https://doi.org/10.1111/j.1469-8986.1992.tb01686.x

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience, 6*(7), 674–681. https://doi.org/10.1038/nn1082

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. https://doi.org/10.1162/089892998563121

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. https://doi.org/10.1111/nyas.13654

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. https://doi.org/10.1016/j.neuroimage.2009.12.019

Pearce, M., & Rohrmeier, M. (2018). Musical syntax II: Empirical perspectives. In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 487–505). Springer.

Pearce, M. T., & Wiggins, G. A. (2012). Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science*, *4*(4), 625–652. https://doi.org/10.1111/j.1756-8765.2012.01214.x

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Przysinda, E., Zeng, T., Maves, K., Arkin, C., & Loui, P. (2017). Jazz musicians reveal role of expectancy in human creativity. *Brain and Cognition, 119*, 45–53. https://doi.org/10.1016/j.bandc.2017.09.008

R Core Team (2018). *R (v. 3.5.2): A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Regnault, P., Bigand, E., & Besson, M. (2001). Different brain mechanisms mediate sensitivity to sensory consonance and harmonic context: Evidence from auditory event-related brain potentials. *Journal of Cognitive Neuroscience, 13*(2), 241-255. https://doi.org/10.1162/089892901564298

Rohrmeier, M., & Pearce, M. (2018). Musical syntax I: Theoretical perspectives. In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 473–486). Springer.

Schön, D., & Besson, M. (2005). Visually induced auditory expectancy in music reading: A behavioral and electrophysiological study. *Journal of Cognitive neuroscience*, *17*(4), 694–705. https://doi.org/10.1162/0898929053467532

Slevc, L. R., & Okada, B. M. (2015). Processing structure in language and music: A case for shared reliance on cognitive control. *Psychonomic Bulletin and Review*, *22*, 637–652. https://doi.org/10.3758/s13423-014-0712-4

Steinbeis, N., & Koelsch, S. (2008). Shared neural resources between music and language indicate semantic processing of musical tension-resolution patterns. *Cerebral Cortex, 18*, 1169–1178. https://doi.org/10.1093/cercor/bhm149

Steinbeis, N., Koelsch, S., & Sloboda, J. A. (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. https://doi.org/10.1162/jocn.2006.18.8.1380

Sun, Y., Lu, X., Ho, H. T., Johnson, B. W., Sammler, D., & Thompson, W. F. (2018). Syntactic processing in music and language: Parallel abnormalities observed in congenital amusia. *NeuroImage: Clinical*, *19*, 640–651. https://doi.org/10.1016/j.nicl.2018.05.032

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-related ERP components. *Oxford Handbook of Event-Related Potential Components*, 397–440.

Thompson, W. F., & Parncutt, R. (1997). Perceptual judgments of triads and dyads: Assessment of a psychoacoustic model. *Music Perception*, *14*(3), 263–280. https://doi.org/10.2307/40285721

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021

Trainor, L. J., McDonald, K. L., & Alain, C. (2002). Automatic and controlled processing of melodic contour and interval information measured by electrical brain activity. *Journal of Cognitive Neuroscience*, *14*(3), 430–442. https://doi.org/10.1162/089892902317361949

Tymoczko, D. (2006). The geometry of musical chords. *Science*, *313*(5783), 72–74. https://doi.org/10.1126/science.1126287

Verleger, R. (1990). P3-evoking wrong notes: Unexpected, awaited, or arousing? *International Journal of Neuroscience*, *55*(2-4), 171–179. https://doi.org/10.3109/00207459008985972

Vuust, P., Ostergaard, L., Pallesen, K. J., Bailey, C., & Roepstorff, A. (2009). Predictive coding of music-brain responses to rhythmic incongruity. *Cortex*, *45*(1), 80–92. https://doi.org/10.1016/j.cortex.2008.05.014

Walsh, M. M., Gunzelmann, G., & Anderson, J. R. (2017). Relationship of P3b single-trial latencies and response times in one, two, and three-stimulus oddball tasks. *Biological Psychology*, *123*, 47–61. https://doi.org/10.1016/j.biopsycho.2016.11.011

Yarkoni, T. (2019, November 22). *The generalizability crisis*. https://doi.org/10.31234/osf.io/jqw35

Appendix

Description of OSF Contents

This project has an associated Open Science Framework (OSF) website. On the website, we share de-identified data, information about our stimuli, and analysis scripts in MATLAB to help readers understand our analysis and make their own should they wish. The OSF page can be accessed at this URL: https://osf.io/xw9v5/. Below are the contents of the OSF page:

1. **biosemi64.sph** This file is for use with EEGLAB and contains information about the electrode locations on the scalp.
2. **electrode_num.bmp** This file is an image showing which electrode indexes correspond to which scalp locations. If readers wish to select electrodes in our associated MATLAB scripts, they will need to do so by these indexes.
3. **features-csv-readme.txt** This file describes the contents of the features.csv file in detail.
4. **features.csv** This file is a spreadsheet showing the stimuli we used and their associated metrics.
5. **publicAnalysisScript_R1.m** This script is for use with MATLAB and can be run and altered by readers to better understand our analyses and to create their own.
6. **publicData_part1.mat – publicData_part4.mat** These four files are the de-identified processed data we used for our experiment. The publicAnalysisScript_R1.m file contains instructions on how to analyze the data. MATLAB is required.
7. **PublicQuestionnaireData.mat** This file is a MATLAB table that contains the data from our questionnaires.
8. **publicRegressionTable.mat** This file is used in the publicAnalysisScript_R1.m file. It is a MATLAB table containing the data we used for our regression models.
9. **public_Regression_v_ttest.m** This file compares the statistical power of a linear regression model compared to a t-test comparing top and bottom quartiles.

Downloaded from http://online.ucpress.edu/mp/article-pdf/39/2/118/485012/mp.2021.39.2.118.pdf by Indiana University Bloomington user on 15 December 2021