



PERCEPTION OF THEMATIC STRUCTURE IN MUSIC OVER SHORT AND LONG TIMESCALES

EDWARD T. R. HALL  & MARCUS T. PEARCE 
Queen Mary University of London, London, United
Kingdom

THEMATIC STRUCTURE IS VALUED HIGHLY IN MUSIC theory and in many models of music cognition. However, strong empirical evidence supporting its perception is scarce. To provide a basis for experimental research, Hall and Pearce (2021) developed a probabilistic computational model of the cognitive processes underlying perception of thematic structure. The model hypothesizes that thematic structures become perceptible due to the statistical regularities formed by repetition and variation of material. Two experiments were conducted to test this model. Experiment 1 focused on small timescales. Forty participants rated whether pairs of themes and repetitions ($N = 100$) came from the same composition, varying across model-based measures of variation and stylistic unpredictability. Significant effects on ratings were found for both measures, across pitch and rhythm representations. Experiment 2 focused on large-scale structures. Forty participants heard 40 two-minute-long melodies, varying in internal unpredictability, repetition, variation, and stylistic unpredictability. Participants identified whether moments in the stimuli were repetitions and rated each melody on its structural unity. Both tasks provided significant evidence for the effects of internal unpredictability on the perception of thematic structure, both for pitch and rhythm representations. These findings suggest perception of thematic structure depends on intrapopus unpredictability, as predicted by the model.

Received: May 23, 2022, accepted November 18, 2024.

Key words: thematic structure, repetition, theme, statistical learning, computational modeling

SENSORY INPUT FROM THE ENVIRONMENT IS rarely uniform but contains patterns that recur both exactly and approximately at a range of scales, allowing observers to learn and anticipate structural regularities. The same is true of human cultural domains such as language and music, where large-scale

temporal organization has an impact on the meaning conveyed by an utterance or piece of music. In language, the emphasis is usually on communicating specific referential semantic content, which places strict requirements on word order and compositional hierarchy prescribed by syntactic considerations. In music, where the semantic content is usually much less specifically referential and stylistic syntax less prescriptive, more emphasis is often placed on the (approximate) repetition of musical material throughout a piece of music, in general accordance with the schematic norms of a musical style. Examples of exact repetition are the literal reoccurrence of a chorus or re-entry of a principal theme, whereas approximate repetition often arises through progressive development and variation of a theme, reflecting an incremental process of change or developments upon successive reoccurrences.

Collectively, these phenomena can be referred to as *thematic structure*, which we distinguish from other kinds of large-scale musical structure, such as tonal relationships created by the introduction and modulation of tonal centres related to musical keys.¹ While some investigations of thematic structure have focused on the specific process of identifying themes themselves (e.g., Lartillot, 2005), here our primary goal is to understand perception of the arrangement of repeated thematic material across a piece of music. We make a distinction between large-scale thematic structure (repetition and variation of thematically related material across large timescales covering substantial portions of a composition) and small-scale thematic structure (reflecting individual instances of immediate thematic repetition or variation).

Traditionally, composers and music theorists have argued in favor of the existence and importance of thematic structures in music, and have gone to great lengths to create, label, and analyze them. For example, the late Peter Kivy (2017) argued from a philosophical perspective that thematic structure arising from

¹ Broadly speaking, large-scale thematic structure can be considered akin to the concept of musical *form*. Form, however, often implies additional relationships between the constituents of a composition reflecting tonal and stylistic conventions both of which, for example, are embodied along with thematic structure in sonata form.

repetition of musical material constitutes the most distinctive characteristic and an essential ingredient in the aesthetic experience of music. Thematic structure is also—even when not explicitly stated—of importance to several influential cognitive theories of music perception. This includes theories that involve structured representations of musical knowledge (e.g., Lerdahl & Jackendoff, 1983; Temperley, 2007) and others that focus more on psychological processes of expectation (e.g., Huron, 2006; Meyer, 1956; Narmour, 1990).

However, despite this prevalent position in theoretical research on music perception, there is limited experimental research and a resultant paucity of empirical evidence, to the extent that we cannot even be sure that large-scale thematic structure is perceived consistently by ordinary listeners without music training (as discussed further below). The present research constitutes an effort to rectify this situation through two experiments that experimentally test the predictions of a recently proposed computational model of the perception of large-scale thematic structure in music (Hall & Pearce, 2021), thereby aiming both to demonstrate perception of thematic structure and shed light on the underlying psychological mechanisms involved. The overall hypothesis embodied in this model is that perception of thematic structure is based on psychological mechanisms of statistical learning and probabilistic prediction of repeated structure in music. The two experiments focus on perception of small-scale and large-scale thematic structure, respectively. The first experiment investigates perception of individual thematic relationships between pairs of musical passages extracted from the same composition and presented in immediate succession, in isolation from their context within the composition. The second experiment investigates perception of relationships between musical passages that may be (approximately) repeated over timescales of up to two minutes within a composition as well as the effect of this repetition on perception of the overall structural unity or coherence of the composition. Our focus is on implicit perceptual mechanisms during naturalistic listening rather than analytical identification of thematic structures relying on extensive music training.

The paper is organized as follows. Past empirical work that investigates the effects of thematic structure on the perception of unity is reviewed, as is research on psychological mechanisms that may contribute to perception of thematic structure. An introduction is then given to a computational model of the perception of thematic structure (previously presented in Hall & Pearce, 2021) that provides a set of four quantitative measures, each describing different aspects of the thematic structure of musical compositions. Testing these

measures as predictors of perception of thematic structure in music, and therefore evaluation of the underlying theory embodied by the model, forms the basis for the experiments presented in this paper. An overview of the specific aims and hypotheses for each experiment is given, followed by the methods and results, and, finally, the findings of the experiments are discussed.

STRUCTURAL UNITY

The current lack of understanding of the role of large-scale structure in perception of music reflects, in part, a difficulty in concisely and concretely defining and measuring its effects; we cannot test whether it has been perceived directly, only indirectly through its secondary effects. When examined on the level of individual compositions, the effects of large-scale thematic structure are often analyzed in terms of the extent to which it creates a sense of unity or coherence, or by comparing preference for one large-scale structure over another. For example, experiments have sought to uncover differences in preference or perception of unity between complete compositions and modified versions, with the assumption that by modifying the order of material, or by removing it, thematic structure is disrupted.

Following this approach, a handful of studies have found significant evidence that such modifications can disrupt perception of thematic structure. Tan and Spackman (2005) found a significant increase in unity ratings for original versions of 15 classical compositions over patchwork versions that mixed sections between works and also over versions consisting of repetitions of single sections. Tan et al. (2006) replicated this effect for original versus patchwork versions but found that the effect diminished over repeated hearings. Lalitte and Bigand (2006) found significantly higher coherence ratings for original over reorganized versions of contemporary and popular music. Additionally, McAdams et al. (2004) found an influence of large-scale organization on listeners' continuous ratings of intra-opus familiarity between two differently structured versions of a piece of contemporary art music. In a recent study by McDonald and Wöllner (2022), ratings of coherence were found to differ significantly across reordered versions Bach's C minor Prelude from Book I of *The Well-Tempered Clavier* (original, "mildly scrambled," "highly scrambled," and "random"). Perceived coherence was found to decrease as versions were manipulated to be further from the original. A smaller effect was found when asking participants to explicitly rate predictability, with the original being judged as more predictable than the "random" version. No significant differences were

found between versions for ratings of enjoyment (pleasantness, interest, “desire to hear again”).

Further to these behavioral findings, Farbood et al. (2015) found evidence of differences in neural processing in musicians listening to an excerpt of a Brahms piano concerto in its original form and versions scrambled at measure, phrase, and section levels. Functional magnetic resonance imaging responses showed a hierarchy of auditory processing within the brain, with only the original version giving reliable responses at the top-most level.

However, the picture is complicated by other research reporting no significant differences between original thematic structures and modified versions. Karno and Konečni (1992) found no differences in preference between original and reordered versions of a movement of a Mozart symphony. Tillmann and Bigand (1996) found no differences in perceived expressiveness and coherence for music by Bach and Mozart with sections in original or reversed orders. However, a small difference was found in perceived expressiveness between versions for music by Schoenberg, providing parallels to the similar findings of Lalitte and Bigand (2006) and McAdams et al. (2004), both also focusing on contemporary music. Eitan and Granot (2008) found no significant differences in coherence or preference for two Mozart piano sonata movements in original forms or with corresponding sections swapped between works. Additionally, in popular music, Rolison and Edworthy (2012) found no significant preference for original versions of songs over versions with various levels of reorganization.

While all of these studies investigate some aspect of large-scale structure, a broad observation can be made between the findings of tasks in which participants rated the perceived unity or coherence of stimuli, and tasks in which they rated preference or enjoyment. Studies reporting significant differences in ratings largely did so only for ratings of unity/coherence (Lalitte & Bigand, 2006; McDonald & Wöllner, 2022; Tan & Spackman, 2005; Tan et al., 2006; with the exception of the effect of familiarity found by McAdams et al., 2004). Little evidence was found for an effect of structure on ratings of preference (Eitan & Granot, 2008; Karno & Konečni, 1992; McDonald & Wöllner, 2022; Rolison & Edworthy, 2012; Tillmann & Bigand, 1996). While this distinction is by no means decisive—other studies considering coherence found no such effect (Eitan & Granot, 2008; Tillmann & Bigand, 1996)—these combined findings provide a small suggestion that listeners have a greater sensitivity to the coherence or unity of a composition’s structure than a preference for a particular

version over another. This difference may, at least in part, be due to the more explicit link between judging unity and directly considering structure.

These experiments all share the assumption that the original version of a composition should elicit the greatest preference or sense of unity. The experiments of Granot and Jacoby (2011, 2012) avoid this assumption. Instead of comparing versions, participants were tasked with placing the disordered sections of a Mozart (in the former experiment) or Haydn (in the latter) piano sonata movement in the order that maximized its coherence. Analysis of the resulting section orders revealed a sensitivity to elements of thematic structure, such as the respective positioning of opening, developmental and closing material, and the more general directional-ity of material.

When manipulating large-scale structure, it is important to consider that perceived unity may also be influenced by tonal closure; manipulated compositions could create a stronger sense of unity if global tonal closure is strong, or strong closure within individual segments could lessen the need for a larger coherent structure. In some studies, these effects are mitigated by the construction of stimuli that maintain consistent tonal closure (Eitan & Granot, 2008; McDonald & Wöllner, 2022) or by the use of atonal music (Lalitte & Bigand, 2006; McAdams et al., 2004; Tillmann & Bigand, 1996). In other studies, such as the puzzle experiments of Granot and Jacoby (2011, 2012), it is more difficult to distinguish between unity generated by thematic or by tonal structures. However, no significant evidence was found to suggest that participants were sensitive to large-scale tonal structures, with participants frequently failing to identify harmonic resolutions between segments (Granot & Jacoby, 2012). A similar effect of greater sensitivity to local, rather than global, tonal structure was found in the puzzle tasks of Tillmann et al. (1998); when pairing halves of short minuets, ordering errors occurred based on the position of strong cadences rather than wider considerations of key.

STRUCTURAL SIMILARITY AND REPETITION

Aside from composition-wide measures, the effects of structure can be examined within the composition itself, in the form of internal relations between the parts of a musical composition that underlie the perception of thematic structure. Two such relations have received attention: perception of similarity and direct repetition of musical material. Both repetition and similarity are widely thought to underlie perception of thematic structure in music and contribute to properties such as unity

in both hierarchical and linear theories of musical structure (Deliège, 2007; Kivy, 1993; Meyer, 1973; Schoenberg, 1967; Smyth, 1993).

Meyer (1973) provides a useful formalization of the relationship between repetition and similarity within musical compositions. Meyer proposes that the extent to which a repetition is recognizable over a substantial time period, or with substantial amounts of intervening material, is proportional to: (1) its similarity to its original instance, (2) its individuality within the composition, and (3) its schematic regularity. To this end, repetition that takes place over smaller timespans can undergo greater variation while maintaining recognition as repetition. Following this formalization, Meyer makes the distinction between two structural roles of repetition: first, as a motif returning over a large time period in which there is no related material and so emphasising a structural unit, and, second, as a more frequently recurring repetition that allows for a gradual succession of modifications.

To investigate similarity as an indicator of thematic structure, empirical research has attempted to uncover the importance of different types of musical feature when making similarity judgements, using the following experimental logic. If the evidence suggests that similarity judgements are made purely on the basis of surface properties of the music—such as its general dynamics or texture—compared with deeper comparisons of its thematic material, the case for listeners' perception of large-scale thematic structure would be weakened. The results of experiments following this approach are ambiguous and require further research. For example, Lamont and Dibben (2001) found no significant evidence that similarity ratings between pairs of themes and extracts taken from piano works by Beethoven and Schoenberg were based on anything other than surface features. Using the same stimuli, Ziv and Eitan (2007) instead examined participants' ratings of excerpts as belonging to the principal themes of either work. Themes were presented to participants multiple times after which excerpts were presented in turn in their order of appearance in the original work. Concurrence between participants' responses and the similarity ratings of Lamont and Dibben (2001) was significantly above chance for the Beethoven extract but not for the Schoenberg, raising the possibility of an effect of stylistic familiarity.

However, the concept of similarity tested in these experiments differs in important ways from the operation of similarity in large-scale musical structure. First, thematic relatedness suggests a sense of belonging to a coherent, unified compositional whole that goes

beyond the more general concept of similarity. Second, there is an empirical focus on similarity at a small-scale without any time separation; passages to be compared are presented adjacently (this is particularly the case for the pairwise comparisons of Lamont & Dibben, 2001), whereas in actual music, thematic structures might be separated by significant time intervals and intervening musical material. Third, the perception of similarity is likely to be highly sensitive to the context in which it occurs—for example, perception of similarity between the present and other parts of the music are likely to be made in the context of the themes and other material previously occurring before that point in the composition. In other words, thematic relationships can be developed continuously across many occurrences of related material spanning substantial portions of a composition (the second of Meyer's types of structural repetition). While both the experiments of Lamont and Dibben (2001) and Ziv and Eitan (2007) included the preparatory playing of stimulus compositions (with Ziv & Eitan, 2007, also including an additional playing in which participants were encouraged to follow the themes and their development), the nature of the tasks used likely results in much of the original context being lost.

Some of these issues are resolved by the investigation of repetition perception, which takes place over longer timescales and can be situated in much longer and more complex musical contexts. Dowling et al. (2001) provide some empirical insight into the second of these issues. In a series of experiments exploring listeners' memory for repetition, after hearing excerpts of classical minuetts, listeners were asked to discriminate between exact target repetitions, similar lures or different lures of the minuet's opening phrases. Targets, or lures, were presented removed from the original by 5 to 30 seconds, with the intervening time covered by the continuation of the minuet (as composed, but controlled to avoid any reinstatement of the target repetition), silence, or a content-less repeated beat pattern. Throughout the experiments, results showed that, over time, discrimination of exact targets from similar lures improved, but only when preceded by the original minuet continuation. These findings were corroborated when using expressive performances (Tillmann et al., 2013) and a wider range of musical styles (Dowling et al., 2016). The improved discrimination between exact and similar repetitions provides some support for Meyer's (1973) theory—that larger time spans favor increasingly similar repetitions. Additionally, this apparent ability of an intervening musically relevant context to aid the perception of repetition presents interesting

implications for the study of repetition that highlight the need for its investigation in the context of real-world musical compositions.

While some of the research discussed investigates the perception of repetition using real-world musical contexts (Dowling et al., 2001, 2016; Tillmann et al., 2013), there is little empirical research using stimuli that are longer and without manipulation—the most prominent experiments being those of Margulis (2012; with further discussion in Margulis, 2014). Listeners were asked to detect—by the press of a button—exact repetitions of material within excerpts of between one and two minutes long taken from four classical recordings. Responses indicated that participants found it difficult to actively identify repetitions on a first listening, with the likelihood that participants would correctly identify a given repetition only marginally above chance for the best-performing excerpt. However, not all types of repetition were equally salient; shorter repeating units were more easily identified.² Additionally, within-phrase repetition was identified more often than repetition of material across phrases, while complete phrases were more easily identified than fragments.

However, despite the apparent difficulty of explicitly identifying repetitions in real time, there is evidence that repetition can implicitly influence the ways in which listeners perceive music. Margulis (2013) demonstrated a significant implicit effect of repetition on listeners' aesthetic responses to contemporary music—increasing enjoyment, interest, and judgements of artistic ability. Though not addressed directly in that study, repetition may also influence perception of thematic structure in a similarly implicit manner. Indeed, it is possible that the enhanced aesthetic experience of the music with increasing repetition actually depended on perception of increased thematic unity or coherence.

Outside of the purely musical context, research on learning of repetition in auditory stimuli provides strong evidence that listeners are both highly sensitive to auditory repetition and that repetitions are implicitly learned (Agus & Pressnitzer, 2013; Bianco et al., 2020). In a series of experiments, Bianco et al. (2020) tested the speeds with which participants could detect emerging regular patterns (a repeated cycle of 20 tones, 50 ms each) in sequences of otherwise random tones. While some patterns were novel, others were repeated across trials. Reaction times significantly decreased after each reoccurrence of a pattern, compared to novel patterns.

² This study also repeated this task over multiple exposures to the excerpts, finding that longer units became the easiest to detect as the number of exposures increased.

Of particular importance to the perception of music are the timescales over which these effects were observed; patterns that reoccurred across trials were sparsely spaced within stimuli, occurring at intervals of approximately three minutes. The same reoccurring patterns were used again in subsequent trials one day and then seven weeks later, with the effect of faster reaction times compared to novel patterns persisting. In the context of music, this spacing indicates that even a little repetition should be able to influence perception of structures at the level of a musical composition. Similar to the findings of Margulis (2013), Bianco et al. (2020) found these effects of repetition to be implicit, with participants being largely unaware of having any memory for the repeated stimuli.

STATISTICAL LEARNING OF THEMATIC STRUCTURE

Hall and Pearce (2021) proposed that statistical learning provides a plausible underlying mechanism for perception of thematic structure. According to this proposal, large-scale thematic structures are perceived through implicit recognition of statistical regularities learned through both exact and inexact repetition and variation of material. Based on this proposal, a probabilistic computational model of the perception of thematic structure was implemented and its behavior analyzed—when applied to a corpus of 623 Western tonal melodies ranging in period from the Baroque to the 20th century. In the following, we first introduce IDyOM, a computational model of auditory expectation that provided a platform for the model of Hall and Pearce (2021), then summarize the main components of the model itself, and finally present the four model-based measures of thematic structure underpinning the present empirical research.

INFORMATION DYNAMICS OF MUSIC

Information Dynamics of Music (IDyOM; Pearce, 2005, 2018) is a computational model that simulates the cognitive process of statistical learning for symbolic representations of auditory sequences, generating conditional probabilities for sequential events. IDyOM, a variable-order Markov model, tallies the occurrences of subsequences of varying length within a given sequence and computes conditional probability distributions for the estimated likelihoods of occurrence of musical events given the preceding context. These distributions are computed using the Prediction by Partial Match (PPM*) algorithm, smoothing between predictions of different order (Bunton, 1997; Cleary & Teahan, 1997). IDyOM may be configured as a *short-term model* (STM) that learns incrementally from an initially empty

state within a given musical sequence—such as the pitches of notes in a melody—representing a listener's short-term acquisition of statistical knowledge about repeated structure within an individual piece of music. It may also be configured as a *long-term model* (LTM), in which case it is trained on a separate set of musical sequences, representing long-term learning of the statistical structure of a musical style, before being applied to predicting the notes of a musical composition.

To date, IDyOM has been used to model perceptual expectation and uncertainty (Egermann et al., 2013; Hansen & Pearce, 2014; Hansen et al., 2016; Omigie et al., 2012, 2013; Pearce, 2005; Pearce, Ruiz, et al., 2010; Sauvé et al., 2018), boundary perception (Pearce, Müllensiefen, & Wiggins, 2010), metre induction (van der Weij et al., 2017), similarity (Pearce & Müllensiefen, 2017), memory (Agres et al., 2018), emotional response (Egermann et al., 2013; Gingras et al., 2016) and aesthetic experience (Cheung et al., 2019; Gold et al., 2019).

A Model of the Perception of Thematic Structure

The Hall and Pearce (2021) model of the perception of thematic structure employed the probabilistic modeling of IDyOM to calculate a range of measures quantifying different properties of a composition's thematic structure. These measures represent competing hypotheses about the psychological mechanisms underlying perception of thematic structure. The present experiments provide an empirical comparison of four of these measures—*internal unpredictability*, *thematic repetition*, *thematic variation*, and *stylistic unpredictability*.

For a given composition, three configurations of IDyOM were used, differing in the music used for training: the STM and LTM configurations introduced above, and *theme-trained models* (TTM), trained on thematic candidates identified in a composition (one TTM per candidate), the identification of which is described below, and then used to predict the remainder of that composition. When combined within a single composition, the TTMs provide information on the predictability of note-events only with regards to that composition's preceding identified themes (this distinguishes it from an STM for the same composition that provides information on event predictability given all of the preceding material). All IDyOM models, however they are configured for training, compute a conditional probability estimate for each note-event in the composition when trained, respectively, on the preceding portion of the composition (dynamically), on a corpus of melodies, or on the possible themes within the composition. Probability estimates were converted into *information content* values, $h = -\log_2 p$, giving

a representation of the unpredictability of the event, given the event's context and the respective model training.

Internal Unpredictability

In the present work, the measure of *internal unpredictability* specifically refers to the STM information content of a composition—most commonly averaged across a composition or part of a composition.

Identification of Thematic Material

While *internal unpredictability* is based on all events in a composition (or part thereof), the remaining three measures apply only to thematically relevant material—material relating to any number of *thematic candidates* (i.e., potential themes) identified by the model within the musical piece. The theme detection method implemented in Hall and Pearce (2021) aimed to identify thematic candidates in a cognitively plausible manner; this constraint excludes many methods developed in the field of Music Information Retrieval (Janssen et al., 2014; Laaksonen & Lemström, 2019; Melkonian et al., 2019; Ren et al., 2017) as it necessitates thematic candidates to be detected dynamically as a composition unfolds. Instead, potential candidates were identified within a composition through a process of theme detection based on the *internal unpredictability* of the STM.³

Thematic candidates were identified as the onset of substantially novel pitch interval material, based on sequential comparisons of unpredictability of musical phrases with that of the preceding material within the composition. Two thresholds dictate the magnitude of change in unpredictability needed for a phrase to be identified as a thematic candidate, chosen with the intention of providing robust identification for a wide range of compositions: A phrase was declared a thematic candidate if: (1) its mean unpredictability was greater than the cumulative mean of the material preceding it by more than half a standard deviation; and (2) there was an absence of highly predictable events, such that the event with the lowest unpredictability was greater than one cumulative standard deviation of that of the preceding phrase. Composition beginnings were considered implicitly to be thematic candidates. Given the difficulty of identifying the precise length of a thematic candidate, all thematic candidates were taken as two phrases long. Phrase boundaries were identified computationally using the rule-based *Grouper* algorithm of Temperley (2001); *Grouper* identifies boundaries based

³ It is not the purpose of the research presented in this paper to specifically evaluate the process of identifying thematic candidates. Rather, it aims to test the four measures that are the model's primary output (of which, some employ these theme detection methods).

on large interonset intervals, consistent phrase lengths, and consistent metrical positioning.⁴

It should be noted that this method identifies thematic candidates purely on the basis of their intra-opus novelty and so does not correspond entirely to the notion of theme in music analysis. True themes may be considered to possess additional properties that add to their perceptual salience, such as differences in harmony, tempo and texture.

Thematic Repetition

The identified thematic candidates were used as training for TTM, with a separate model created for each candidate identified. Each TTM was trained on the thematic candidate and then used to predict the remainder of the composition commencing at that candidate's onset. The note-by-note information contents thus generated were divided into two clusters using Gaussian Mixture modeling, with the lower (more predictable given the TTM) cluster labelled *thematic material*. *Thematic repetition* reflects the proportion of events in a musical stimulus that are labelled as thematic material by any of the TTMs.

Thematic Variation

The extent to which the thematic material varies from its corresponding thematic candidate can be quantified when clustering TTM values at the phrase level. An information-theoretic measure of *compression distance* (Li et al., 2004; Pearce & Müllensiefen, 2017)—the normalized summed information content of the notes making up a phrase, given a TTM trained on the corresponding candidate—gives the *dissimilarity* between each phrase and its corresponding theme. *Dissimilarity* values were averaged for each melody to give a measure of how far thematic material develops from the parent thematic candidate; higher *thematic variation* indicates greater divergence between the thematic material and the parent themes within a composition.

Stylistic Unpredictability

The stylistic content of the thematic material was modeled using an LTM trained on a corpus of Western tonal music (described below). The mean LTM information contents for thematic note-events provides a measure of the *stylistic unpredictability* of thematic material.

THE PRESENT EXPERIMENTS

The two experiments presented in this paper were designed to test the extent to which these model-based

measures account for listeners' perception of thematic structure over different timescales—respectively, over short timescales in which thematic relationships manifest immediately, and over long timescales in which such relationships are numerous and occur over substantial spans of time. By testing these measures, we had three general aims: first, to examine the extent to which listeners can perceive differences in thematic structure between compositions; second, to compare the competing structural properties of music on which such perception is based; and third, to assess the hypothesis that perception of thematic structures relies on psychological processes of statistical learning (Hall & Pearce, 2021).

Music is a highly complex stimulus, containing features in the musical surface that can be represented in a large number of ways. While the above-described model (and the IDyOM components underpinning it) can apply to many different representations of music, to allow for these experiments and analyses to be tractable, this paper concerns itself with two—the sequential interval between pitches (in semitones), and the inter-onset interval between note events. This provides representations of melodic pitch and rhythm domains and allows us to test their independent effects. Of these, the pitch interval representation was prioritized in stimulus selection to avoid complications identified in Hall and Pearce (2021) when modeling thematic structure with rhythmic representations. It is viable in Western classical music for compositions to be rhythmically isochronous to varying degrees, such as consisting of only a single note duration (examples of which are common in Bach's works for solo violin and cello), consisting only of a single short pattern, or containing substantial rhythmically isochronous passages. Such rhythmic material can mask the identification of more nuanced thematic repetition. With the use of these two representations, it should be noted that these experiments do not aim to provide a complete account of all possible thematic relationships available within melodies, such as material that may only be related through shared implicit harmonies.

The model-based measures were designed to apply to complete pieces of music but the present experiments probe responses to excerpts from complete compositions of varying duration. Therefore, the four primary measures were divided into corresponding sub-measures (referred to as “experiment measures”) that reflect their precise application within the context of the two experiments (see Table 1).

Specifically, Experiment 1 focused on perception of thematic relations on a very short timescale, between

⁴ An example of this process of identifying thematic candidates is given in Hall and Pearce (2021), alongside discussion of its output.

TABLE 1. Summary of Model Measures and Derived Experiment Measures used in the Experiments

Model measure	Experiment measure	Experiment / task	Description
<i>Internal unpredictability</i>	<i>Internal unpredictability</i> ^a	2 / unity	The average unpredictability of all note-events for a model trained dynamically within a composition
	<i>Internal unpredictability of moment</i>	2 / recognition	The unpredictability of the phrase at a recognition moment when trained on the composition preceding it
	<i>Internal unpredictability at moment</i>	2 / recognition	The average internal unpredictability of a composition before a given moment
<i>Thematic repetition</i>	<i>Thematic repetition</i> ^a	2 / unity	The proportion of note-events identified as thematic material within a composition
	<i>Thematic repetition at moment</i>	2 / recognition	The average thematic repetition in a melody before a given moment
<i>Thematic variation</i>	<i>Dissimilarity</i>	1	The normalized compression distance between a parent thematic candidate and a subsequent phrase identified as a repetition
	<i>Thematic variation</i> ^a	2 / unity	The average dissimilarity between all thematic phrases and parent thematic candidates within a composition
	<i>Thematic variation at moment</i>	2 / recognition	The average thematic variation in a melody before a given moment
<i>Stylistic unpredictability</i>	<i>Stylistic unpredictability</i> ^a	2 / unity	The average unpredictability of thematic note-events for a model trained on a corpus of compositions
	<i>Stylistic difference at moment</i>	2 / recognition	The average stylistic unpredictability of a melody before a given moment
	<i>Stylistic difference</i>	1	The absolute difference in stylistic unpredictability for two given sequences
	<i>Mean stylistic unpredictability</i>	1	The (grand) mean of stylistic unpredictability for two given sequences

^aExperiment measure matches that of the model

a thematic candidate and a repetition of the candidate identified by the model. It was important to examine whether the model-based measures could predict perception on a local level before extending the enquiry to large-scale thematic structure—the testing of which introduces multiple complications. This experiment specifically examined the effects of *dissimilarity* (the individual component of *thematic variation*), *stylistic difference*, and *mean stylistic unpredictability* (both derived from *stylistic unpredictability*) on the extent to which pairs of thematic candidates and repetitions were perceived as being derived from the same piece of music. The framing of this task as a question of belonging to a shared work reflects the fact that such relationships are influenced by more than purely the similarity between the two; stylistic and structural considerations provide important ways in which thematic material may be considered to be related.

Because thematic repetition in music is not restricted to this immediate timescale, Experiment 2 investigated whether the model-based measures could account for perception of thematic structure over larger timescales. This experiment tested the influence of the four model measures on two different indicators of thematic structure: first, the ability of participants to identify musical

material at specified moments as being a repetition of material that appeared earlier in the piece (the recognition task); and second, the perception of a composition's structural unity (the unity task). For the recognition task, the four primary measures were applied to the preceding portion of a composition before each recognition moment and, additionally, the unpredictability of the moment itself within the composition was investigated.

For both experiments, we hypothesized that the measures of thematic structure would account significantly for participants' perception of thematic relationships and structure. Corroboration of this general hypothesis would provide evidence that large-scale thematic structure can be perceived in a systematic way, and that this perception can be simulated in terms of psychological mechanisms of statistical learning.

More specifically, for Experiment 1, three hypotheses were advanced for the relationships between participant ratings and the experiment measures: (1) that *dissimilarity* would have the strongest effect on participant responses versus the stylistic measures, with stimulus pairs showing high *dissimilarity* more likely to be judged as unrelated; (2) that *stylistic difference* would have a secondary effect with stimulus pairs showing

high *stylistic difference* being more likely to be perceived as unrelated, particularly when *dissimilarity* is high and so does not immediately identify the two as related; and (3) that overall *stylistic unpredictability* would be associated with a tendency to perceive the excerpts as unrelated—in other words, if the stimuli are stylistically novel to a participant, their ability to judge similarity would be reduced.

For Experiment 2, two hypotheses were proposed for the effects of experiment measures on participants' abilities to recognize repetition and perceive differences in structural unity: (1) that material with low *internal unpredictability* would be recognized as thematically related repetition; and (2) that low *internal unpredictability*, high *thematic repetition*, or low *stylistic unpredictability* would each increase the ability of participants to perceive repeated thematic structure and, therefore, increase perceived sense of unity for these substantial extracts from compositions.

For both experiments, we recorded measures of participants' musical backgrounds, hypothesizing that increased music training—through greater exposure to music and increased learning of stylistic conventions—would increase the impact of stylistic measures on perception of thematic structure. We made no specific hypotheses as to the effects of musical background on intra-opus learning.

Experiment 1

METHOD

Participants

Forty participants were recruited to participate through the online platform Prolific (prolific.co). No exclusion criteria were applied other than a required first language of English and normal or corrected to normal hearing. Participants had a mean age of 31.98 years ($SD = 10.76$) and 24 were women. Participants were of eight nationalities, with 30 of UK nationality. There were no prerequisites on music training; 20 participants reported having received some formal music training on at least one instrument, of which four reported training for more than 10 years.

Stimuli

Stimuli were selected from a large corpus of Western-classical tonal melodies (Hall & Pearce, 2021) using the output of the above-described model of thematic structure (stimuli, as well examples and the data collected for this experiment, can be found at osf.io/zmqh2). Stimuli took the form of pairs of a thematic candidate and a later phrase from the same composition identified by the

model as being related (or belonging) to that candidate. These theme–phrase pairs varied in three experiment measures: (1) *dissimilarity*, the normalized compression distance between the two stimuli; (2) *stylistic difference*, the absolute difference in LTM information content between the two stimuli; and (3) *mean stylistic unpredictability*, the mean LTM information content of the pair (see Table 1 for relationship with the primary model measures).

Theme–phrase pairs were selected from the full set of possible pairs extracted from the corpus by the model. Outliers more distant than two SD from the mean for each measure were removed. Phrases that, in their entirety, were exact matches with any sub-part of the thematic candidate were discarded; for the paradigm used, these phrases would not return useful information. To avoid any possible overlap of material between stimuli, selection was limited to one phrase per candidate and one candidate per composition—the first thematic candidate was used in all cases. Following these constraints, 100 theme–phrase pairs were selected at random from the subset of 6,687 available.⁵ Audio files were generated for the selected stimuli in a piano timbre with uniform loudness, dynamics, and articulation, using MuseScore notation software (musescore.org). Original pitches, rhythms, and tempos were preserved.

In the selected stimuli, thematic candidates had a mean duration of 8.32 seconds ($SD = 4.92$) and phrases had a mean duration of 4.93 seconds ($SD = 1.50$).

Procedure

The experiment was implemented using the jsPsych JavaScript library (jpspsych.org)—with a custom designed task—to be conducted online through the participant's web browser (de Leeuw, 2015).

The experiment consisted of two parts. For the first (the more substantial task, taking around 30 minutes for completion), participants were presented with all 100 of the theme–phrase pairs (theme first, separated by a short gap) in a randomized order. Stimulus pairs were presented in four equal blocks with a 30-second mandatory rest period in-between each. Participants were not informed of the theme–phrase relationship between the pairs of stimuli.

⁵ Our analyses for this experiment are primarily concerned with testing differences between stimuli, based on model measures. To estimate the stimulus sample size needed, a statistical power analysis was performed for a multiple linear regression using the three model measures as predictors. Due to a lack of previous research that closely resembles the experiment's paradigm, and the likely weak effect of stylistic measures, a weak effect size of 0.2 was assumed, using Cohen's (1988) criteria. With an alpha of .05 and power of 0.8, the projected sample size needed with this effect size is 33 (G*Power, Faul et al., 2007). The 100 stimuli selected are more than adequate for the analyses of this experiment.

Instead, they were instructed that pairs may, or may not, be excerpts from two different compositions. Participants were asked to give a rating answering the question, “to what extent do these two excerpts sound like they are from the same piece?” Ratings were given on a moveable slider between “same piece” (returning a value of 0) and “different pieces” (returning 100). Participants were required to listen fully to both excerpts before submitting a rating. Additionally, as an experimental check for responses given by participants with specific veridical knowledge of the pieces used, participants were asked to indicate if the piece corresponding to either excerpt was known to them.

In the second part of the experiment, participants gave their responses to the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report questionnaire for musical sophistication (Müllensiefen et al., 2014).

Statistical Analysis

Ratings for stimuli were treated differently in five separate analyses. For all but the final analysis, the effects of measures in the pitch and rhythmic representations are presented separately.

The first analysis was categorical, comparing participant responses between high and low values of each experiment measure. Ratings for theme–phrase pairs were split into two at the median of each of the three measures, giving high/low *dissimilarity* (dissimilar/similar), high/low *stylistic difference* (stylistically distant/close), and high/low *mean stylistic unpredictability* (stylistically unpredictable/predictable) separately for both pitch and onset models. For each participant, ratings for stimuli in each cell of the resulting factorial design were averaged. A three-way repeated-measures ANOVA was used for each representation (pitch, onset) to test main and interaction effects of the three categorical measures on participant ratings.

The remaining analyses were continuous, the second using correlation coefficients (Pearson’s r) between the mean participant results for each theme–phrase pair and the pair’s corresponding value for *dissimilarity*, *stylistic difference*, and *mean stylistic unpredictability* in each representation. Third, a mixed-effect linear regression analysis was used, for each representation, to test the combined abilities of *dissimilarity*, *stylistic difference*, and *mean stylistic unpredictability* to predict the ratings of theme–phrase pairs. Fourth, a mixed-effects regression analysis compared the effects of measures between pitch and onset representations on participant ratings.

In the final analysis, participant responses to Gold-MSI questions were aggregated by summing participant responses (accounting for question phrasing) to

produce a measure of *general sophistication* for each participant, and aggregated for question subsets relating to participants’ *perceptual abilities* and *music training*. Correlation coefficients were also calculated between mean ratings for each participant and their aggregated Gold-MSI scores, and between linear regression slopes for individual participants in each measure and Gold-MSI scores.

RESULTS

The results are presented individually for each analysis as follows: (1) a summary of participant ratings, their distributions, and Gold-MSI scores; (2) the ANOVA for interactions between ratings and model measures; (3) correlation and multiple linear regression analyses of relationships between ratings and model measures; (4) the multiple linear regression analysis comparing measures from models using pitch and rhythm representations; and (5) the analysis of correlations between ratings and model measures, and between participants’ Gold-MSI scores, slopes and their ratings.

All 40 participants returned ratings for all 100 theme–phrase stimuli pairs and all gave full responses to the Gold-MSI questionnaire. Participants gave ratings of the extent to which the two phrases belonged to the same composition with a mean of 47.36 ($SD = 34.29$), using 98% ($SD = 3.73$) of the scale on average. Out of all responses, 1.5% of pairs were reported as both excerpts being known to the participant; 1.6% reported only the theme being known; 1.4% reported only the phrase being known.

After aggregation of responses to the Gold-MSI self-report questionnaire into scores for each participant, participants had a mean score for *general sophistication* (scale range 18–126) of 53.30 ($SD = 17.19$), a mean score for *perceptual abilities* (out of a possible scale range of 9–63) of 38.48 ($SD = 6.30$), and a mean score for *music training* (scale range 7–49) of 14.53 ($SD = 9.58$).

Categorical Analysis

For the pitch interval and interonset interval representations, stimuli were categorized as being either “high” or “low” in each of their three model measures, split at the median. Participants’ ratings were averaged for the stimuli in the eight combinations of categories produced, giving one value per participant in each condition. Table 2 summarizes the ratings for these categories.

A three-way repeated-measures ANOVA was used to test the interaction between the three categorical measures and participant ratings for each representation. For pitch interval, individually all three measures

TABLE 2. Descriptive Statistics of the Ratings for Stimuli in the Factorial Experimental Conditions of Dissimilarity (High, Low), Stylistic Difference (High, Low), and Mean Stylistic Unpredictability (High, Low)

		Dissimilarity			
Stylistic difference	Mean stylistic unpredictability	High		Low	
		M	SD	M	SD
Pitch interval					
High	High	59.09	7.27	43.60	12.82
	Low	60.67	8.93	42.20	11.64
Low	High	61.60	14.13	34.08	10.30
	Low	50.38	8.46	33.48	7.82
Interonset interval					
High	High	63.15	7.98	28.62	11.21
	Low	66.18	11.00	37.12	9.12
Low	High	60.11	9.20	37.81	9.85
	Low	59.08	15.15	27.43	9.51

showed significant main effects: *dissimilarity*, $F(1, 39) = 230.40$, $p < .001$, $\eta_p^2 = .59$; *stylistic difference*, $F(1, 39) = 34.06$, $p < .001$, $\eta_p^2 = .14$; and *mean stylistic unpredictability*, $F(1, 39) = 14.41$, $p < .001$, $\eta_p^2 = .03$. For all three measures, mean ratings were higher (i.e., indicating different pieces) for the “high” category. There was a significant two-way interaction between *dissimilarity* and *stylistic difference*, $F(1, 39) = 9.72$, $p < .01$, $\eta_p^2 = .03$, such that the effect of *stylistic difference* was greater when *dissimilarity* was low. There was also a significant interaction between *stylistic difference* and *mean stylistic unpredictability*, $F(1, 39) = 12.79$, $p < .001$, $\eta_p^2 = .03$, such that the effect of *mean stylistic unpredictability* was greater when *stylistic difference* was low. There was a significant three-way interaction between all model measures, $F(1, 39) = 13.18$, $p < .001$, $\eta_p^2 = .04$, such that when *dissimilarity* was high and *stylistic difference* low, there was a greater effect of *mean stylistic unpredictability* (see Figure 1).

For interonset interval, significant effects were shown for individual measures of *dissimilarity*, $F(1, 39) = 870.89$, $p < .001$, $\eta_p^2 = .76$ and *stylistic difference*, $F(1, 39) = 7.13$, $p < .01$, $\eta_p^2 = .03$, but not for *mean stylistic unpredictability*. For both significant measures, mean ratings were higher for the “high” category. There was a significant two-way interaction between *dissimilarity* and *stylistic difference*, $F(1, 39) = 5.87$, $p = .02$, $\eta_p^2 = .02$, such that the effect of *stylistic difference* was greater when *dissimilarity* was high—the converse relationship to that of pitch interval. There was also a significant interaction between *stylistic difference* and *mean stylistic unpredictability*, $F(1, 39) = 33.16$, $p < .001$, $\eta_p^2 = .11$, such that when *stylistic difference* was low, pairs with low *mean stylistic unpredictability* would

be given lower ratings (i.e., indicating the same parent piece), with the inverse true at high *stylistic difference*. A significant three-way interaction between all model measures, $F(1, 39) = 13.84$, $p < .001$, $\eta_p^2 = .05$, shows that the interaction between *stylistic difference* and *mean stylistic unpredictability* is greatest when *dissimilarity* is low (see Figure 1).

Continuous Analyses

Participants’ ratings for each stimulus were averaged to give a mean rating for each theme–phrase pair. These mean ratings were then analyzed using the three model measures in each representation as predictors. For both pitch interval and interonset interval, the mean ratings had a significant positive correlations both with *dissimilarity*, $r(98) = .58$, $p < .001$ and $r(98) = .56$, $p < .001$, respectively, and with *stylistic difference*, $r(98) = .37$, $p < .001$ and $r(98) = .27$, $p < .01$, but had no correlation with *mean stylistic unpredictability*, $r(98) = .05$, $p = .64$ and $r(98) = .17$, $p = .08$. The extent to which measures themselves were correlated is given in Table 3.

Using mixed-effects multiple linear regression, ratings were analyzed using the three measures as predictor variables, accounting for the random effects of participant and stimulus pair. Due to the over-fitting of data when using the maximal random effects structure, only random intercepts were included. Summaries for analyses of both representations are shown in Table 4. For pitch interval, *dissimilarity* and *stylistic difference* were found to be significant predictors, with greater *dissimilarity* or *stylistic difference* predicting higher ratings that pairs were not thought to come from the same piece, $\beta^* = 0.33$ and $\beta^* = 0.14$, respectively. The pitch interval model accounted for 41% of the total variance

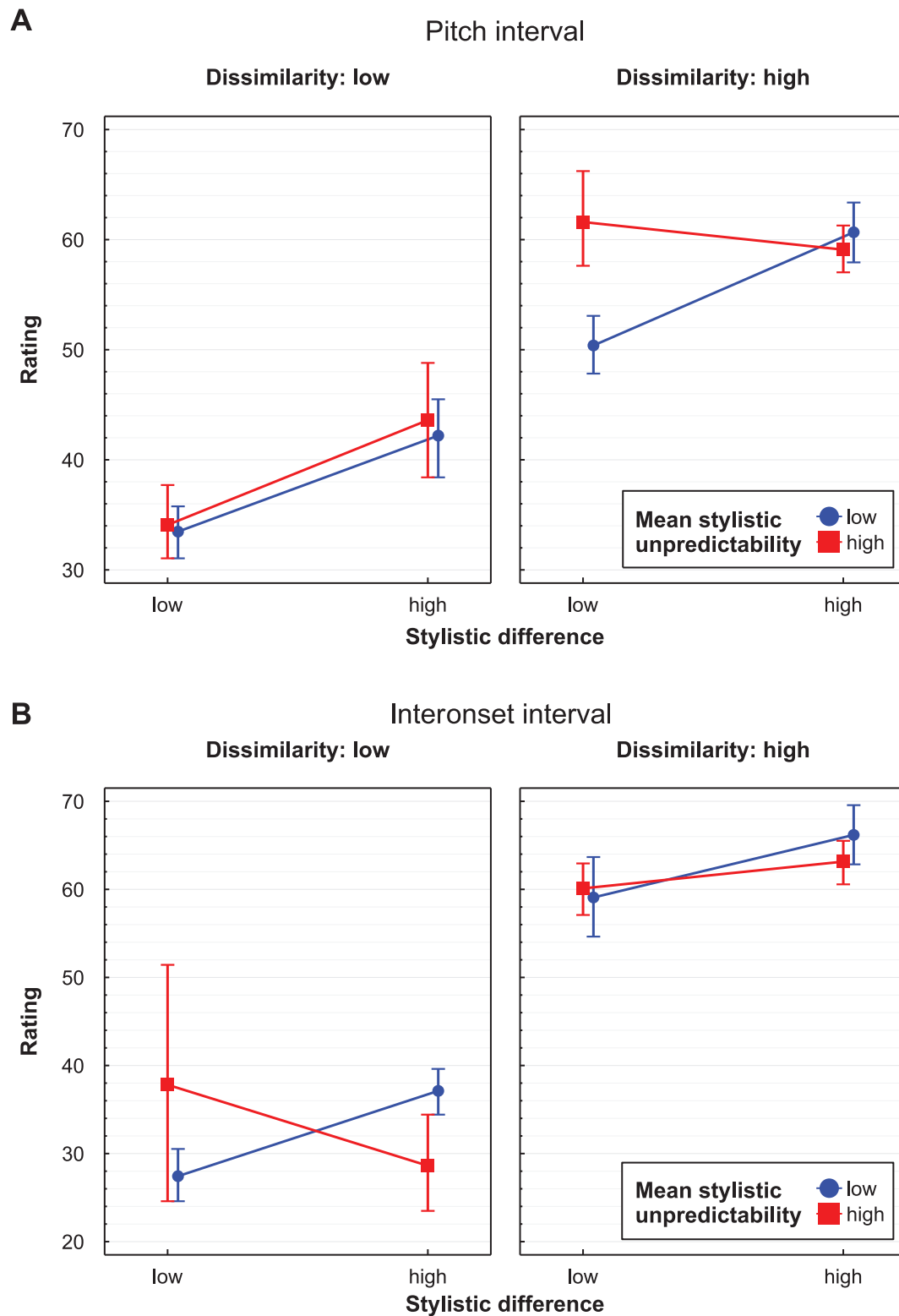


FIGURE 1. Interaction effects of categorical measures on mean participant ratings for (A) pitch interval and (B) interonset interval. Note: Error bars show 95% confidence intervals.

TABLE 3. *Pearson's r Correlations Between Experiment Measures and Participant Ratings*

Variable	<i>M</i>	<i>SD</i>	1	2	3
Pitch interval					
1. <i>Dissimilarity</i>	0.88	0.27	–		
2. <i>Stylistic difference</i>	0.73	0.50	.29	–	
3. <i>Mean stylistic unpredictability</i>	3.20	0.53	.04	.09	–
Ratings (stimuli means)	47.36	21.69	.58	.37	.05
Interonset interval					
1. <i>Dissimilarity</i>	1.56	1.19	–		
2. <i>Stylistic difference</i>	1.18	1.13	.23	–	
3. <i>Mean stylistic unpredictability</i>	1.93	0.81	.04	.25	–
Ratings (stimuli means)	47.36	21.69	.56	.27	.17

TABLE 4. *Mixed-Effects Linear Regression Analyses Predicting Participants' Belongingness Ratings by Experiment Measures for Each Pitch and Rhythmic Representation, Accounting for Participant and Stimulus Differences*

Predictor	β	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Pitch interval					
Intercept	2.28	12.02	97.04	0.19	.85
<i>Dissimilarity</i>	42.21	6.84	96.00	6.17	< .001
<i>Stylistic difference</i>	9.63	3.68	96.00	2.61	.01
<i>Mean stylistic unpredictability</i>	0.24	3.32	96.00	0.07	.94
Interonset interval					
Intercept	23.21	5.19	101.48	4.47	< .001
<i>Dissimilarity</i>	9.62	1.55	96.00	6.22	< .001
<i>Stylistic difference</i>	2.31	1.68	96.00	1.37	.17
<i>Mean stylistic unpredictability</i>	3.32	2.28	96.00	1.46	.15

in the data; participant intercepts *SD* = 5.59, stimulus intercepts *SD* = 16.79. For interonset interval, only *dissimilarity* was a significant predictor, with greater *dissimilarity* associated with higher ratings, $\beta^* = 0.33$. The interonset interval model accounted for 41% of the total variance in the data; participant intercepts *SD* = 5.59, stimulus intercepts *SD* = 17.28.

A final mixed-effects linear regression analysis was used to test the relative ability of all measures across both representations to predict participant's ratings, again with random intercepts of participant and stimulus. As shown in Table 5, *dissimilarity* in both pitch interval and interonset interval was found to be a significant predictor of ratings ($\beta^* = 0.24$, $\beta^* = 0.24$), with higher *dissimilarity* corresponding with higher ratings. Additionally, pitch interval *stylistic difference* was found to have a significant effect ($\beta^* = 0.11$). The combined representation model accounted for 42% of the total variance in the data; participant intercepts *SD* = 5.59, stimulus intercepts *SD* = 14.42.

To test whether participant's musical backgrounds influenced ratings, Gold-MSI scores were correlated both with the mean ratings for each participant and with coefficients (slopes) from simple linear models

predicting each participant's ratings from each of the model measures in each of the representations. No significant correlation was found between any of the three scores of *general sophistication*, *perceptual abilities*, and *music training*, and participants' mean ratings or slopes for the pitch interval model. A significant correlation was found between *general sophistication* and slopes using interonset interval *stylistic difference*, $r(38) = .37$, $p = .02$, such that ratings from participants with higher Gold-MSI scores were more influenced by the *stylistic difference* of pairs.

DISCUSSION

The results of this experiment support the original hypotheses, showing that experiment measures using both pitch and rhythmic representations have a significant influence on listeners' perception of theme–repetition relationships, operationalized as the extent to which the two excerpts sounded like they came from the same piece of music. The measure of *dissimilarity* between the thematic candidate and repetition had the greatest effect on this relationship such that—as hypothesized—high *dissimilarity* resulted in pairs being less likely to be judged as belonging to the same piece. This finding was

TABLE 5. Mixed-Effects Linear Regression Analyses Predicting Participants' Belongingness Ratings by Experiment Measures Across Both Pitch and Rhythmic Representations, Accounting for Participant and Stimulus Differences

Predictor	β	SE	df	t	p
Intercept	-3.16	10.67	94.27	-0.30	.77
Pitch interval					
Dissimilarity	31.71	6.49	93.00	4.89	< .001
Stylistic difference	7.90	3.26	93.00	2.42	.02
Mean stylistic unpredictability	-0.70	3.06	93.00	-0.23	.82
Interonset interval					
Dissimilarity	6.88	1.40	93.00	4.91	< .001
Stylistic difference	0.54	1.48	93.00	0.36	.72
Mean stylistic unpredictability	3.93	2.04	93.00	1.93	.06

confirmed for pitch and rhythm models separately, and when measures from both representations were tested together. Given that the task is closely related to one providing direct similarity ratings between musical passages, and the measure of *dissimilarity* is calculated as the compression distance between the pair of excerpts, this result provides further evidence that compression distance accurately simulates perceived similarity between pairs of melodies (Pearce & Müllensiefen, 2017).

Compression distance provided a direct, encapsulated measure of dissimilarity that was found to be independent of any relationship the excerpts had with other music or musical styles. This effect of *dissimilarity* provides some strong support for the importance of intra-opus statistical learning in the perception of thematic relationships in music. The finding that these stimulus pairs can be judged as being related solely through the unpredictability of one, given the other, supports the hypothesis that similar relationships may be perceived when pairs are situated in large pieces of music.

The results also suggest that there is a stylistic element in determining the relationship between passages of music, albeit confined to the pitch domain. Increasing the pitch-based *stylistic difference* between the two passages decreased the likelihood that they would be considered to belong to the same piece.

The results provide little evidence that the musical backgrounds of listeners influence their judgements of relationships between passages. Only one significant difference was found with participants' *general sophistication* influencing the importance of the *stylistic difference* measure for rhythm only. This finding provides some tentative support that listeners' musical backgrounds, through their greater music training and exposure, influences their sensitivity to stylistic elements. The lack of other findings based on musical background should perhaps be treated with some caution; it was not a primary aim of this experiment to explicitly compare participants with and without

music training, and so the participants have a narrower range of musical expertise and background than would otherwise be chosen. However, the lack of any effect of musical background on any intra-opus measures for this task corroborates the findings of Lamont and Dibben (2001), and Ziv and Eitan (2007), both of which explicitly tested for—and did not find—effects of music training on similarity judgements.

These findings provide initial corroboration of three measures from the Hall and Pearce (2021) model of the perception of large-scale thematic structure in music. Specifically, they provide evidence that measures based on intra-opus compression distance between two thematic excerpts and stylistic regularity acquired through long-term statistical learning influence how listeners perceive relationships between the subsections of music making up an overall piece. This experiment, however, is limited to testing these measures on an immediate timescale and without any repetition of material, as typically occurs in music. Experiment 2 investigated whether these measures can account for perception of thematic structure over longer timespans allowing for multiple (approximate) repetitions of thematic material.

Experiment 2

METHOD

Participants

As with Experiment 1, 40 participants were recruited using Prolific, with the same criteria. There was no overlap in participants between the two experiments. Participants had a mean age of 33.58 years ($SD = 14.03$) and 21 were women. Participants were of nine nationalities, with 25 of UK nationality. Again, there were no prerequisites on music training; 27 participants reported having received some formal music training on at least one instrument, of which three reported training for more than 10 years.

Stimuli

Stimuli for this experiment consisted of 40 two-minute-long melodies.^{6,7} Melodies were taken from the same corpus of longer compositions (described in Hall & Pearce, 2021), trimming over-length compositions and disregarding those shorter.

The possibility that a composition may be entirely rhythmically isochronous posed several obstacles for selecting stimuli and identifying repetitions of musical material for this experiment. For this reason, pitch interval only was used in the stimulus selection process.

For each of the four model measures, compositions more distant than three *SD* from the mean were removed and 40 melodies were randomly sampled from the remaining pool. To minimize disruption to the perception of large-scale structure by local effects of the stimuli ending abruptly, endings were manually identified within an additional 15-second window; endings were identified (with descending order of preference) at either a section ending marked in the score, before a substantial gap in the melody, or at a strong phrase ending—i.e., with a perfect cadence. Audio files were generated for the selected stimuli in a piano timbre with uniform loudness, dynamics, and articulation, using Apple's GarageBand software with original pitches, rhythms, and tempos preserved.

Within the second minute of each stimulus, four phrases were selected for the recognition task, described below. These phrases were selected to span a wide range of *internal unpredictability* (mean *unpredictability* for the phrase) when compared to the cumulative median *unpredictability* for the stimulus. To ensure this, phrases were selected iteratively alternating between high—above the cumulative median—and low—below the median—states (with the initial state randomized); at each iteration, the phrase furthest from the median in the stated direction was selected. Phrases within two seconds of those already identified were excluded in order to minimize potential disruption caused by moments being presented in quick succession. This process was repeated within the second minute of each

stimulus until four phrases were identified for each. These selected phrases are referred to as “recognition moments” or simply “moments.”

The selected stimuli had a mean duration of 127.37 seconds (*SD* = 5.12). Recognition moments had a mean duration of 2.85 seconds (*SD* = 1.74).

Procedure

As with Experiment 1, an online experiment was created using jsPsych. The experiment comprised a first part containing the main experimental task (taking around 50 minutes for completion) and a second part of the Gold-MSI self-report questionnaire in the same form as in Experiment 1.

Four sets of 20 stimuli each were created, distributed randomly, with each stimulus present in exactly two sets. Participants heard one set each—10 hearing each set—with order of presentation randomized individually for each participant.

Individual trials contained a single stimulus melody and involved two tasks, the first a recognition memory task at four moments while listening to the melody, and the second a rating of coherent unity for the melody in its entirety.

For the recognition tasks, a color-coded visual indicator was used. A red indicator notified participants that they should just listen to the stimulus, an amber indicator gave notice that a recognition moment was approaching, and a green indicator identified the passage of music for which the task was to be conducted. Participants were asked to state whether they thought they had heard the exact musical passage playing at the indicated moment (while the green indicator was displayed) at any point previously in the stimulus.⁸ Responses were after the indicator had changed back to red using the “y” (heard before) and “n” (not heard before) keys. Four such recognition moments occurred during the second half of each stimulus, as described above. To avoid any overlap between responding to one moment and attending to another, participants were asked to provide their responses as quickly as possible.⁹

After listening to each stimulus, participants were asked to provide a rating of the level of structural unity they perceived in the stimulus. Specifically, they were

⁶ These stimuli, as well examples and the data collected for this experiment can be found at <https://osf.io/e86ys/>.

⁷ As with the first, this experiment is primarily concerned with testing differences between stimuli. A statistical power analysis was performed for the unity rating task where it was hypothesized the weakest effect would be found. The analysis was based on a multiple linear regression using the four model measures as predictors. As reviewed in the introduction, previous similar research suggests a weak effect of structure on unity ratings should be assumed; 0.2 was used, based on Cohen's (1988) criteria. With an alpha of .05 and power of 0.8, the projected sample size needed with this effect size is 33 (G*Power, Faul et al., 2007).

⁸ Recognition of exact repetition was specified to reduce ambiguity and potential disruption due to increased cognitive load (see the Discussion of Experiment 2). It was assumed that highly similar repetitions would also be captured by this task due to the imperfect memory of participants and the time constraints involved.

⁹ While participants were asked to respond in a speeded manner, these timings were not intended as a measure of performance. Timings cannot be generalized between the different task moments within and across stimuli—the differences in material across these moments introduces different cognitive processing constraints.

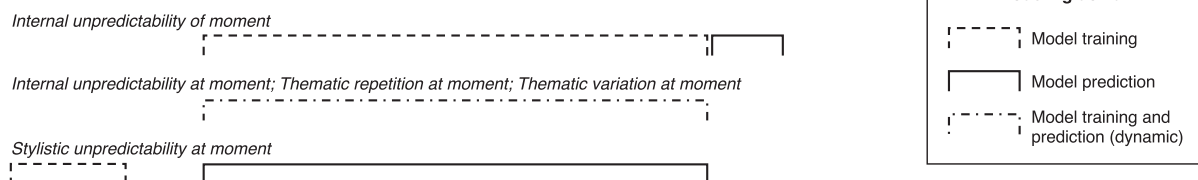
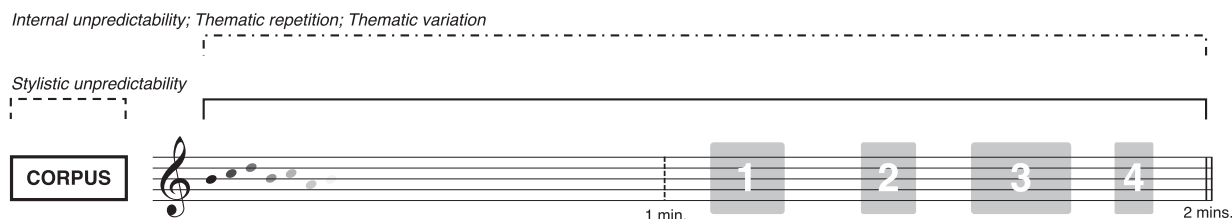
Recognition moment task**Unity rating task**

FIGURE 2. Illustrative training and prediction domains for Experiment 2 measures. *Recognition moment task*: For a given recognition moment (shown here for the first moment but applied to all four), domains used for model training only are shown with dashed lines and model prediction only with a solid line; domains in which both training and prediction are applied dynamically are shown with dash-dotted lines. *Unity rating task*: As in the previous task, dash-dotted lines illustrate measures based on training and prediction within a stimulus composition; *stylistic unpredictability* is trained on a corpus and used to predict events in the composition.

asked to rate “the extent to which you think the different parts of the piece unify into a coherent whole” on a continuous scale from “not very unified” (returning a value of 0) to “very unified” (returning 100). Participants were also given the following guidance (adapted from the experimental materials used in Tan & Spackman, 2005):

For a unified piece, even though there may be many different ideas in it, the music still sounds like one well-integrated, whole, single composition. A piece that is not unified, on the other hand, is one that sounds like unconnected fragments of music that do not seem to belong together, so that they do not hold together as one well-integrated, whole, single piece of music.

As with Experiment 1, participants were asked to indicate whether each piece was known to them. Again, this was included as an experimental check for responses given by participants with specific veridical knowledge of the works used.

Statistical Analysis

Data collected for this experiment contained three levels of detail: (1) responses for each recognition moment; (2) unity ratings for each stimulus; and (3) Gold-MSI questionnaire responses for each participant. Each of these levels was analyzed in turn for both pitch interval and interonset interval representations.

The independent variables for the analyses consisted of model-based characterizations of the stimuli both at the

level of the entire stimulus and at the level of specific moments. At the stimulus level, melodies were characterized in terms of the four basic model measures described above: (1) *internal unpredictability*; (2) *thematic repetition*; (3) *thematic variation*; and (4) *stylistic unpredictability*. In addition to the “high” or “low” categorization from the selection process for each moment (and an analogous categorisation made using interonset interval),¹⁰ moments were also characterized by five experiment measures. First, the *unpredictability of moment* gave the mean unpredictability for the phrase when trained on the preceding portion of the stimulus.¹¹ Furthermore, moments were also characterized in terms of the following properties of the melody up to the start of the recognition moment (corresponding to the four stimulus-level characteristics): the *unpredictability at moment*, the *thematic repetition at moment*, the *thematic variation at moment*, and the *stylistic unpredictability at moment*. Figure 2 illustrates how these measures apply to the stimuli (see

¹⁰ Due to the effect of completely isorhythmic content for a small number of stimuli (the factor that excluded this representation from use in moment selection) there is a small discrepancy between numbers of “high” and “low” unpredictability categorizations for inter-onset interval.

¹¹ As the phrase to be recognized does not appear in isolation to the listener, but rather with some context of the melody immediately preceding it, this measure is practically calculated as the mean internal unpredictability (which is dynamically trained throughout the melody) for the note-events of the moment.

Table 1 for a summary of the relation of experiment sub-measures to primary model measures). Separate measures for pitch interval and interonset interval were calculated for stimulus-level and moment-level measures.

For the recognition task, data for pitch and rhythm models were first analyzed using a chi-squared test comparing participant responses to the moments' high/low *internal unpredictability*. Chi-squared tests were then used to test responses separately against binary classifications for the remaining five measures for each representation, splitting each at the median value for that measure across all recognition moments. In all cases missing responses were dropped. Second, recognition-moment data were analyzed using mixed-effects logistic regression analyses with participants' responses as the dependent variable. The independent variables were *unpredictability of moment* and the four "at moment" measures for both pitch interval and interonset interval representations.

Analysis of participant unity ratings first tested for correlation (Pearson's r) between ratings and the four interval model measures for each representation. Mixed-effects linear regression analyses were then used to examine whether the four measures were significant predictors of unity ratings for each representation.

As with Experiment 1, participant responses to Gold-MSI questions were aggregated to produce measures of *general sophistication*, *perceptual abilities*, and *music training* for each participant. To examine the effect of participants' musical backgrounds on their recognition responses, a multiple logistic regression analysis was conducted predicting responses from Gold-MSI scores. The correlation between Gold-MSI scores and both mean unity rating per participant and slopes from simple linear regression models between ratings and measures was also examined.

RESULTS

The results for this experiment are presented individually for each level of data collected for both pitch interval and interonset interval representations—at the levels of (1) recognition moments, (2) entire stimuli, and (3) participant Gold-MSI scores.

Recognition Moments

Each of the 40 participants heard 20 melodies, each containing four recognition moments, with 40 melodies across all participants. In total, 1,642 responses were given for moments being heard before and 1,514 for moments not heard before, with 44 missing responses. Responses were aggregated to give the proportion of "heard before" to "not heard before" responses for each stimulus moment (missing responses were not included). Correlations were found for corresponding measures between pitch and rhythmic representations; measures involving intra-opus training showed this positive correlation, only *stylistic unpredictability at moment* did not.

For chi-squared tests, proportions were rounded to produce a single binary response for each moment. Due to the multiple comparisons being made, a Bonferroni-adjusted alpha value of .004 was used. Classification of the moment according to *internal unpredictability* (high or low), relative to that of the melody before the moment, was found to have a significant effect on responses for both pitch interval, $\chi^2(1, N = 80) = 34.25, p < .001$, and interonset interval, $\chi^2(1, N = 80) = 32.33, p < .001$. For binary classifications of the remaining five measures (see Table 6), *unpredictability of moment* was found to be significant for pitch interval, $\chi^2(1, N = 80) = 30.64, p < .001$, and interonset interval, $\chi^2(1, N = 80) = 18.24, p < .001$, and *unpredictability at*

TABLE 6. Descriptive Statistics for High and Low Categories of Individual Experiment Measures

Measure	High		Low	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Pitch interval				
<i>Unpredictability of moment</i>	4.78	1.03	1.00	0.55
<i>Unpredictability at moment</i>	3.10	0.47	1.70	0.42
<i>Thematic repetition at moment</i>	0.55	0.09	0.30	0.08
<i>Thematic variation at moment</i>	0.91	0.15	0.46	0.09
<i>Stylistic unpredictability at moment</i>	3.55	0.37	2.65	0.26
Interonset interval				
<i>Unpredictability of moment</i>	2.23	0.80	0.83	0.36
<i>Unpredictability at moment</i>	1.65	0.25	1.11	0.30
<i>Thematic repetition at moment</i>	0.81	0.10	0.40	0.16
<i>Thematic variation at moment</i>	1.06	0.20	0.59	0.14
<i>Stylistic unpredictability at moment</i>	1.78	0.62	0.62	0.24

moment was found to be significant for pitch interval only, $\chi^2(1, N = 80) = 12.08, p < .001$. For all three of these measures, participants were more likely to mark high *unpredictability* moments as not being heard before. The remaining measures were not found to be significant for either pitch or rhythm.

Using mixed-effects logistic regression, recognition-moment responses were analyzed using the five pitch interval measures (in continuous form) as predictor variables. The model accounted for the random effects of participants and moments (nested in their respective stimulus). However, due to the over-fitting to data when using the maximal random-effects structure, these differences were modeled using random intercepts only. As shown in Table 7, predictor *unpredictability of moment* was found to be highly significant. Higher values (increased unpredictability) favored the response “not heard before.” The 2.76 odds ratio (OR^{-1}) associated with

increasing *unpredictability of moment* indicated that for every 1-point increase, the odds of a “not heard before” response increased by 276%. Participant intercepts had $SD = 0.62$, moment intercepts $SD = 1.02$. For a logistic regression model predicting responses using interonset interval measures with random intercepts of participant and moment (see Table 7), a similar significant effect of *unpredictability of moment* was found. The odds ratio indicated that for every 1-point increase in *unpredictability of moment*, the odds of a “not heard before” response increased by 311%. Participant intercepts had $SD = 0.62$, moment intercepts $SD = 1.02$.

To test both representations together, a mixed-effects logistic regression using all measures from both, as well as accounting for the random intercepts of participants and stimulus moments, was used to predict recognition response proportions. As shown in Table 8, both pitch interval and interonset interval *unpredictability of*

TABLE 7. Mixed-Effects Logistic Regression Analyses Predicting Recognition-Moment Responses by Experiment Measures for Each Pitch and Rhythmic Representation, Accounting for Participant and Stimulus Differences

Predictor	β	OR^{-1}	SE	z	p
Pitch interval					
Intercept	2.08	0.86	0.76	2.72	< .01
Unpredictability of moment	-0.49	2.76	0.05	-10.44	< .001
Unpredictability at moment	-0.25	1.23	0.13	-1.90	.06
Thematic repetition at moment	-0.02	1.00	0.72	-0.23	.98
Thematic variation at moment	-0.10	1.03	0.42	-0.24	.81
Stylistic unpredictability at moment	0.05	0.97	0.20	0.24	.81
Interonset interval					
Intercept	2.25	0.90	0.59	3.78	< .001
Unpredictability of moment	-1.21	3.11	0.12	-10.44	< .001
Unpredictability at moment	-0.20	1.08	0.27	-0.72	.47
Thematic repetition at moment	-0.68	1.18	0.45	-1.51	.13
Thematic variation at moment	0.59	0.85	0.38	1.53	.13
Stylistic unpredictability at moment	-0.06	1.05	0.15	-0.42	.67

TABLE 8. Mixed-Effects Logistic Regression Analyses Predicting Recognition-Moment Responses by Experiment Measures Across Both Pitch and Rhythmic Representations, Accounting for Participant and Stimulus Differences

Predictor	β	OR^{-1}	SE	z	p
Intercept	2.46	0.89	0.80	3.06	< .01
Pitch interval					
Unpredictability of moment	-0.30	1.86	0.05	-5.85	< .001
Unpredictability at moment	-0.02	1.02	0.15	-0.16	.87
Thematic repetition at moment	0.34	0.95	0.68	0.51	.61
Thematic variation at moment	-0.21	1.06	0.40	-0.52	.60
Stylistic unpredictability at moment	0.05	0.97	0.20	0.26	.80
Interonset interval					
Unpredictability of moment	-0.76	2.03	0.13	-5.91	< .001
Unpredictability at moment	-0.27	1.11	0.32	-0.83	.41
Thematic repetition at moment	-0.70	1.19	0.47	-1.50	.13
Thematic variation at moment	0.67	0.83	0.39	1.73	.08
Stylistic unpredictability at moment	-0.13	1.10	0.15	-0.88	.38

TABLE 9. Mixed-Effects Linear Regression Analyses Predicting Stimulus Unity Ratings by Experiment Measures for Each Pitch and Rhythmic Representation, Accounting for Participant and Stimulus Differences

Predictor	β	SE	df	t	p
Pitch interval					
Intercept	88.87	12.41	36.15	7.16	< .001
Internal unpredictability	-12.82	2.77	35.28	-4.63	< .001
Thematic repetition	-3.87	11.72	35.01	-0.33	.74
Thematic variation	-4.30	8.01	35.14	-0.54	.60
Stylistic unpredictability	3.30	3.28	35.13	1.00	.32
Interonset interval					
Intercept	99.73	7.53	37.83	13.24	< .001
Internal unpredictability	-23.40	3.49	35.13	-6.70	< .001
Thematic repetition	-7.94	6.43	35.25	-1.24	.26
Thematic variation	2.34	5.78	34.65	0.41	.69
Stylistic unpredictability	-0.89	2.40	37.00	-0.37	.71

moment were significant predictors. With odds ratios of 1.85 and 2.04, for every 1-point increase in the respective pitch interval or interonset interval measure, the odds of “not heard before” increased by 185% and 204%. An AIC score of 3495.00 for this combined model showed a better goodness of fit to the data than those using only pitch measures, 3525.40, or rhythmic measures, 3517.40 (where the rhythmic measure model showed a better fit than the pitch one). Participant intercepts had $SD = 0.62$, moment intercepts $SD = 0.89$.

Unity Ratings

All 40 participants returned ratings of unity for all 20 melodies presented to them. Out of all responses, 3% of stimuli were reported as being known to the participant. Across the 40 total stimuli, unity was rated with a mean of 57.50 ($SD = 27.88$) and the mean scale usage by participants was 82% ($SD = 15.43$). Unity ratings were combined by producing an average for each stimulus. As with the recognition task measures, corresponding measures between representations were positively correlated, excluding *stylistic unpredictability*.

Correlations between mean unity ratings and the four pitch interval model measures showed a highly significant correlation between *internal unpredictability* and unity, $r(38) = -.61$, $p < .001$. No significant correlation was found for the remaining three measures; *thematic repetition*, $r(38) = -.03$, $p = .86$; *thematic variation*, $r(38) = -.011$, $p = .50$; and *stylistic unpredictability*, $r(38) = -.02$, $p = .89$. Correlations between unity ratings and interonset interval measures showed the same pattern of results, with a highly significant correlation found with *internal unpredictability*, $r(38) = -.78$, $p < .001$, and no significant correlations found for *thematic repetition*, $r(38) = .11$, $p = .48$; *thematic variation*, $r(38) = -.011$, $p = .50$; and *stylistic unpredictability*, $r(38) = -.05$, $p = .06$.

Using mixed-effects linear regression, unity ratings were analyzed using the four measures as predictor variables, accounting for random intercepts of participant and stimulus. As shown in Table 9, for both representations the predictor of *internal unpredictability* accounted for a significant proportion of variance, with higher *internal unpredictability* corresponding to lower perceived unity for a melody ($\beta^* = 0.30$ and $\beta^* = 0.38$, respectively). The pitch interval model accounted for 32% of the total variance in the data and the interonset interval model accounted for 31% of the total variance in the data. Participant intercepts varied for pitch interval $SD = 9.58$ and interonset interval $SD = 9.71$. Stimulus intercepts varied for pitch interval $SD = 9.56$ and interonset interval $SD = 6.62$.

As the sole predictor of significance in the regression models for both representations, the relative ability of *internal unpredictability* to account for variance in participant ratings can be compared between representations (still accounting for stimulus and participant random effects). Pitch interval *internal unpredictability* accounted for 31% of variance in mean ratings ($\beta = 12.08$, $df = 38.28$, $t = 4.64$, $p < .001$) and interonset interval *internal unpredictability* accounted for 31% of variance, ($\beta = 22.73$, $df = 37.74$, $t = 7.88$, $p < .001$) as shown in Figure 3. However, the strong correlation between the measures of internal unpredictability for pitch and rhythm, $r(38) = .72$, $p < .001$, makes it difficult to ascertain which has the stronger effect.

Gold-MSI Scores

After averaging of Gold-MSI responses into scores for each participant, participants had a mean score for *perceptual abilities* (out of a possible scale range of 9–63) of 40.32 ($SD = 6.16$), a mean score for *music training* (scale range 7–49) of 16.53 ($SD = 8.66$), and

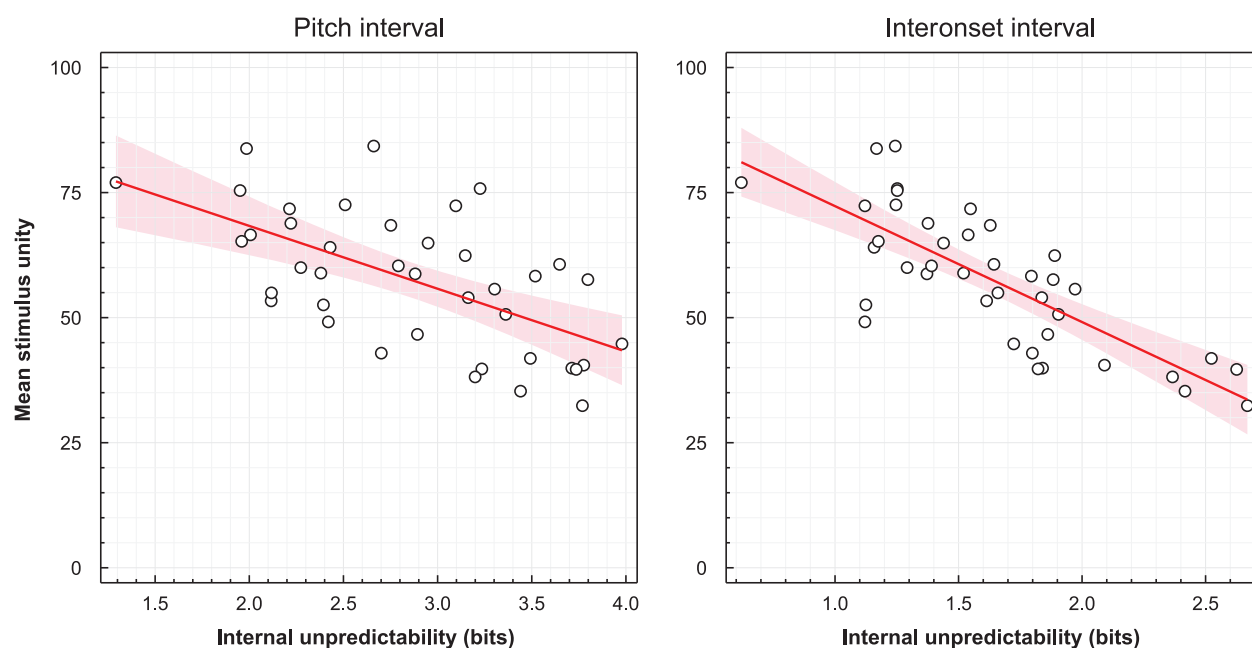


FIGURE 3. Fitted linear relationships of *internal unpredictability* predicting mean stimulus ratings using pitch interval and interonset interval representations. Note: Shaded regions show 95% confidence intervals.

a mean score for *general sophistication* (scale range 18–126) of 56.60 ($SD = 17.48$).

Multiple logistic regression was used to test for the effects of musical-sophistication scores on participants' recognition-moment responses. Participants' *perceptual abilities* were found to have a weakly significant effect, $b = -0.02$, $z(3155) = -2.14$, $p = .03$. The other scores were not significant predictors: *general sophistication*, $b = 0.004$, $z(3155) = 1.27$, $p = .20$; and *music training*, $b = -0.002$, $z(3155) = -0.39$, $p = .70$.

Correlations between Gold-MSI scores and mean unity ratings for each participant were tested; no significant correlations were found. Slope coefficients from linear models predicting each participant's unity ratings from model measures values for stimuli were also tested against Gold-MSI scores (shown in Table 10). For both representations, significant correlations were found between *general sophistication* and individual slopes for measures of *thematic variation*, and between *perceptual abilities* and slopes for measures of *internal unpredictability*. Additionally, participant slopes for interonset interval *thematic variation* was found to have a significant correlation with *music training*.

DISCUSSION

The results from the recognition-moment task provide convincing evidence that intra-opus *internal unpredictability* has an effect on the perception of thematic

repetition within a composition. The results indicate that this effect is realized in two ways; however, with some differences between pitch and rhythmic representations. First, if a passage of music is predictable, given the music that has occurred before it, it is likely to be perceived as a repetition for both pitch and rhythmic measures of *internal unpredictability of moment*. Second, and to a lesser extent, if the music occurring before the passage in question is itself predictable, the passage will be more likely to be perceived as a repetition, with the chief evidence for this measure found using the pitch representation. These findings support our hypothesis regarding the relationship between *internal unpredictability* and repetition recognition.

The difference between pitch and rhythmic representations for the results of the repetition recognition task may be due, in part, to the presence of rhythmically isochronous stimuli; such stimuli would offer maximum repetition with minimum variation of interonset interval, increasing the likelihood that a passage would be perceived as having been heard earlier in the piece.

Although in this research we are concerned with the recognition of both exact and inexact repetition, exact repetition was specified in the recognition task of this experiment for two reasons. First, it was chosen to reduce ambiguity, making the task as clearly interpretable as possible. The task followed the assumption that, though the instruction states exact repetition, the

TABLE 10. Correlation with Gold-MSI Scores for Participant Mean Ratings and Slopes Predicting Participants' Unity Ratings by Model Measures for Each Pitch and Rhythmic Representation

Variable	<i>M</i>	<i>SD</i>	<i>General sophistication</i>	<i>Perceptual abilities</i>	<i>Music training</i>
Mean ratings	57.50	11.30	.07	-.14	.03
Pitch interval					
Slopes ^a					
<i>Internal unpredictability</i>	-11.85	9.36	-.09	-.37*	-.07
<i>Thematic repetition</i>	-5.87	33.04	-.17	-.10	-.12
<i>Thematic variation</i>	-10.65	22.86	-.41**	-.08	-.20
<i>Stylistic unpredictability</i>	-1.37	9.78	-.05	-.10	-.01
Interonset interval					
Slopes ^a					
<i>Internal unpredictability</i>	-22.67	13.76	-.14	-.34*	-.11
<i>Thematic repetition</i>	5.27	22.77	-.23	-.07	-.13
<i>Thematic variation</i>	-6.49	20.06	-.34*	-.19	-.32*
<i>Stylistic unpredictability</i>	-7.18	8.50	.06	-.26	.01

^aSlopes for linear regressions predicting each participant's unity ratings from a model measure

* $p < .05$; ** $p < .01$

imperfect memory of listeners and time constraints would result in merely highly similar repetition being treated as "exact." Second, this reduced ambiguity, and the simple nature of a binary response reduces the potential cognitive load imposed by the task, minimizing disruption caused by a single moment judgement on other tasks across the same stimulus. However, it should be noted that there is the potential for participants who possess both exceptionally good memory and highly accurate similarity judgement to disrupt this assumption—recognizing similar repetitions and correctly categorizing them as not identical—although the task results suggest that this was largely not the case.

The results for ratings of structural unity revealed a similar pattern of effects to those of the recognition task. There was a strong effect of *internal unpredictability* on the perception of structural unity for both pitch and rhythm. This result corroborates our hypothesis, that compositions that are internally predictable due to repetition of material are perceived as having a strong sense of unity. No significant influence of *thematic repetition*, *thematic variation*, or *stylistic unpredictability* was found for either representation.

The extent to which listeners' musical backgrounds influence how they perceive thematic structure is still somewhat uncertain. In both tasks, the majority of Gold-MSI score comparisons to participant responses did not yield any significant relationships, with three exceptions. First, there was a significant effect of *perceptual abilities* as a predictor of recognition moment identification. This appears to indicate that listeners with lower *perceptual abilities* were more likely overall to

consider the recognition moments as repetitions, perhaps reflecting less accurate encoding and therefore less finely discriminated recognition of familiar material. Second, participants with greater overall *general sophistication* (and, additionally, *music training* in the case of rhythm) showed a greater sensitivity to *thematic variation*, with greater sophistication associated with a stronger negative association between *thematic variation* and unity. Third, participants with higher scores of *perceptual ability* had greater sensitivity to the *internal unpredictability* of stimuli, such that greater *perceptual ability* was associated with a stronger negative effect of *internal unpredictability* on unity.

Overall, these findings provide evidence that large-scale thematic structure can be readily perceived by listeners, reflected both by an ability to identify the internal elements of structure over a relatively long time span, and by an ability to distinguish differences in inherent structural unity of different compositions. Furthermore, it was possible to predict differences in a perception of large-scale thematic structure between stimuli as a function of experiment measures. Importantly, the significant influence of *internal unpredictability* in accounting for these effects supports our hypothesis that statistical learning is a plausible psychological mechanism allowing large-scale thematic structure to be perceived.

General Discussion

The two experiments reported in this paper provide evidence that thematic structure in music can be

perceived by listeners, including structure emerging on the timescale of substantial portions of compositions. The results also provide some evidence that this perception is facilitated through the effects of dynamic statistical learning and probabilistic prediction. Both experiments aimed to test a computational model of the perception of thematic structure in music (Hall & Pearce, 2021) that uses statistical learning to generate four measures of *internal unpredictability*, *thematic repetition*, *thematic variation*, and *stylistic unpredictability*—hypothesized to be involved in perception of thematic structure—and compare their relative importance within the separate domains of pitch and rhythm.

Understanding how these measures interact and influence the perception of musical phrases presented in isolation as immediate pairwise comparisons of thematic material (in Experiment 1) was considered an important first step to understanding their efficacy for longer passages of music (Experiment 2). For thematic repetition to be identified over large timescales, the process of comparing incoming musical material to existing material already heard and stored in memory must happen many times throughout the course of listening to a composition. If the experiment measures did not influence perception of isolated instances of immediate thematic repetition, they surely would not scale up to perception of large-scale thematic structure. Therefore, the rationale of Experiment 1 was to test the model-based measures of thematic structure in the microcosm provided by single pairwise comparisons of musical passages from the same piece. Due to the nature of this experimental paradigm, only a subset of the primary model measures were applicable: *dissimilarity* (derived from *thematic variation*), *stylistic difference*, and *mean stylistic unpredictability* (both derived from *stylistic unpredictability*). *Thematic repetition* is not meaningful in this context, and *internal unpredictability*—insofar as its functional aspect in this paradigm is the *unpredictability* of the phrase given the thematic candidate—is encompassed by *dissimilarity* between the two musical passages.

The results from Experiment 1 showed that all three of the experiment measures tested, to some extent, significantly influence perception of the relationship between thematic candidate and repetition. *Dissimilarity* between the two musical passages had the strongest overall effect across both representations followed by *stylistic difference*, albeit with some differences between pitch and rhythm. These results indicate that the extent to which listeners perceive two musical passages as coming from the same piece of music depends on two factors: first, their *dissimilarity* in isolation, operationalized here in terms of

compression distance; and second, any differences in terms of their congruence with Western stylistic norms, internalized in memory via statistical learning during long-term prior musical experience. *Mean stylistic unpredictability* had a more subtle influence, suggesting that when the two passages showed little *stylistic difference*, greater incongruence with Western musical norms led listeners to perceive the two passages as coming from different compositions, perhaps on the basis of overall stylistic unfamiliarity. These findings provide important initial evidence in support of statistical learning as an underlying psychological mechanism for the perception of thematic structure. However, the local effects revealed in Experiment 1 might not necessarily have generalized to the perception of thematic structure on the much larger timescales found in musical compositions.

Therefore, Experiment 2 extended the scope to longer musical passages of around two minutes duration and, in doing so, investigated the effects of the model-based measures on the ability of listeners both to identify thematic repetitions of musical material and to perceive structural unity. With the full musical context available in these longer stimuli, the model-based measures were applied in the same form as they are hypothesized to operate in real-world musical listening. In contrast with Experiment 1, the analysis focused on how knowledge acquired dynamically via statistical learning throughout the entire piece would influence the perception of both repetition and unity. The findings of Experiment 2 present a more striking contrast between the influence of the model-based features. There was strong evidence for effects of *internal unpredictability* for both pitch and rhythm, such that phrases that are predictable given the preceding musical context are more likely to be perceived as having been heard before, especially when the preceding context was itself predictable. Furthermore, pieces that are internally predictable are perceived as having a greater sense of coherent unity. These results provide evidence that statistical learning of internal regularities within a piece of music is a plausible mechanism by which listeners can both detect thematic repetitions and perceive thematic unity.

The extent to which measures based on the isolation of thematic material influenced participants' perception of thematic structure was less apparent. We suggest two possible causes. First, we should consider potential differences between the way in which themes (or rather, thematic candidates) are detected in the model, and how they may be perceived in real-world listening. The model identifies potential themes as substantially novel material, working incrementally through a composition from beginning to end. This is consistent with

traditional understandings of “theme” but probably represents something of a simplification of perceptual theme identification. The extent to which this is the case is hard to discern given the relatively little empirical research on theme perception. Any discrepancies between the theme detection model and perception may add noise to measures relying on theme detection—*thematic repetition* and *thematic variation*—whereas *internal unpredictability* is computationally simpler and does not rely on theme detection.

However, a second explanation presents itself. In addition to being computationally simpler, *internal unpredictability* is also applicable to the entirety of a composition’s material, whereas *thematic repetition* and *thematic variation* apply only to the material identified as being thematic. From this, we can infer that material identified by the model as thematic is not the sole contributor to perception of thematic structure. Instead, the results suggest that *any* material that becomes predictable through repetition, no matter how insignificant, can contribute to a constantly accumulating perception of thematic structure. This provides important indicators about how repetition contributes to perception of large-scale structure. Relatively precise repetition of important material, the focus of most empirical investigations of repetition perception to date, may be secondary to the effects of any repetition at all, no matter how small in scale or approximate. In this light, we can trace the significant effect of *dissimilarity* in the first experiment to the significant effect of *unpredictability of moment* in the recognition-moment task of the second, rather than to *thematic variation*.

Taken together, the results of the two experiments also suggest some substantial disparities in how stylistic content is perceived at the two timescales investigated. Aside from *dissimilarity* and *thematic variation*, *stylistic unpredictability* was also tested directly in both experiments. In Experiment 1, this measure was subdivided into two elements, the *stylistic difference* within the pair and the *mean stylistic unpredictability* of the pair combined. In Experiment 2, the influence of *stylistic predictability* of the thematic material preceding a recognition moment or coherence judgement was investigated. The results obtained on the local scale of Experiment 1 indicate that style was a significant factor in determining whether a musical phrase is perceived as coming from the same piece of music as a presented thematic candidate. However, there was little evidence of a similar effect in the recognition-moment task of Experiment 2. This task is highly related to that of Experiment 1, asking participants to judge whether material is related to that presented before it. It is unlikely that this results

from the focus of stylistic predictability on thematic material because the stylistic predictability of the thematic material is highly correlated with overall stylistic predictability (including both thematic and non-thematic material). It seems more likely that the richer context of Experiment 2 increased the relative salience of predictability, so that any effects of stylistic congruence, apparent in the more impoverished context of Experiment 1, were rendered imperceptible.

In both experiments, while findings for pitch interval and interonset interval measures are largely in agreement, there is still some disparity between them. In some cases this difference should be treated cautiously, as the presence of rhythmically isochronous compositions may exaggerate the influence of certain features when testing using representations of rhythm through their maximum repetition and minimum variation. However, rhythmically isochronous compositions are relatively frequent in Western music, suggesting such compositions should not be unfamiliar for listeners or disrupt their perceptual processing of thematic structure. Instead, it seems likely that the psychological representations involved in the cognitive process of perceiving thematic structure vary based on their contextual relevance, as has been shown in research investigating the perception of other musical features (Prince, 2014; Prince et al., 2009). Additionally, analysis of the stimuli from the two experiments provides some insight into the relationship between pitch and rhythm representations over different timescales. The correlation between composition-trained (i.e., intra-opus, non-stylistic) measures across representations suggests these properties become more correlated as the length of excerpt they are applied to increases—barring any effects of isochrony. The cause of this correlation may be due to the repetition characterized by these measures. Repeated material usually maintains at least some of its pitch and rhythmic content between repetitions. As longer excerpts can afford more repetition—and increased repetition increases predictability—when material is repeated, predictability in both domains increases together.

There is one further way in which the results of the two experiments are in agreement: there is little substantial evidence that listeners’ musical backgrounds have an effect on the way structural elements of music are perceived. The only findings of any significance were in Experiment 2, where listeners with lower *perceptual abilities* more often perceived all recognition moments as being a repetition, listeners with greater *perceptual abilities* were more likely to show stronger negative effects of *internal unpredictability* on perception of unity, and listeners with greater *general*

sophistication were more likely to show stronger negative effects of *thematic variation*. However, it should be cautioned that testing effects of musical background was not the primary aim of this research, and participants were not specifically recruited on the basis of musical ability. Although participant groups showed a reasonably wide range of musical backgrounds, as measured by the Gold-MSI sub-scales, these data are primarily reported for ease of replication and comparison with other studies. The conclusions that may be drawn with respect to music training effects are, therefore, limited without future targeted confirmatory research, with participant groups carefully selected on the basis of musical ability and level of exposure to Western classical music. That said, we can at least be fairly confident that the present results do not depend critically on high levels of music training or specific familiarity with certain styles.

As an initial model-based study of perception of thematic structure, several constraints were imposed on both computational modeling and experimental design to ensure a tractable, interpretable empirical investigation. It is important, therefore, that the findings of these experiments are considered within the context of these limitations. Foremost among these are constraints on the musical information processed by the model.

First, the restriction to monophonic stimuli is an important reduction that avoids making problematic assumptions as to the way in which polyphonic music material is processed psychologically while also avoiding complexities in the modeling (e.g., certain elements of the original model, such as identifying phrase boundaries, can only accept monophonic material, as discussed in Hall & Pearce, 2021). Nonetheless, we argue that the melodies used in these experiments can function as full compositions in their own right, and so produce results that are generalizable to music more widely. Furthermore, the modeling approach, grounded in statistical learning and probabilistic prediction, is also generalizable in principle to polyphonic music (Goldman et al., 2021; Harrison & Pearce, 2020; Sears et al., 2019).

Second, the findings of these experiments should be considered in the context of the representations of musical surface used. To make the modeling of thematic structure in this research tractable, only a single pitch and a single rhythmic representation were used (pitch interval and interonset interval respectively). While these representations can accommodate a great deal of musical information—in which, we would argue, a substantial amount of thematically relevant structure can be found—they do not provide a complete account of thematic structure in music (nor, almost certainly, listeners' perception of it),

even within these monophonic stimuli. These two representations are by no means the only two relevant to their respective domains, and, due to the correlation between measures of different representations, the extent to which they can be compared is limited. Future research is needed that can provide an in-depth comparison of relative effects of representation on the perception of thematic structure, in particular, investigating the performance of more abstract representations such as pitch contour. In this context, it is worth noting the capability of IDyOM to represent and integrate multiple representations of the musical surface.

Similarly, while the focus on monophonic material removes features of *explicit* harmony, *implicit* harmonic relationships are still possible within the melodies used. As such, it should be noted that a listener may perceive a thematic connection between two passages that is not manifested in lower-level patterns of pitch interval, and so unavailable to the model. A relationship may be perceived based on patterns of implied harmonies, or based on melodic patterns of non-adjacent pitches that are given salience by the implied harmonic context. The construction of a representation of implied harmony in melodies is highly complex, with a current lack of research into developing and empirically validating any such techniques. Future research should address this issue, alongside the investigation of the effects of explicit harmony on the perception of thematic structure, for which suitable robust representations are already in existence (Cambouropoulos, 2016; Harrison & Pearce, 2020; Sears et al., 2019).

Also to be considered are the effects of musical features such as dynamics, timbre, texture, articulation, and pitch register. Previous research provides evidence that, when present, such features influence listeners' perception of similarity relationships in music (Eitan & Granot, 2009; Pollard-Gott, 1983). While, in the present experiments, effects of these features (with the exception of pitch register) were controlled either by their removal, or through being made uniform across stimuli, they provide an important avenue for future research. It should be further noted that the feature of pitch register cannot be experimentally controlled in the same manner, with notes in the stimuli occurring at the same pitch and octave as in the original compositions from which they were taken. The choice of a pitch interval representation used in the modeling of these stimuli prioritises an invariance to transposition above an ability to account for register. In addition to future experimentation investigating its role on the perception of thematic structure, future work is, therefore, needed to develop and test appropriate computational representations of pitch register.

It should also be noted that the stimuli used in Experiment 2 were limited to a duration of two minutes. This length was chosen as being long enough both to contain large-scale musical structure and to provide an appropriate amount of context for both tasks used, but short enough to collect sufficient experimental data within a reasonable experimental session. However, while there do exist many classical compositions of this length, it is certainly on the shorter end of the range. These findings are, therefore, limited when scaling up from these compositions of two minutes to those of quarter- or half-hour length or longer; model-based perception of thematic structure within works of this length remains to be investigated in future experimental research.

Finally, differences between the musical experiences of participants and the training corpus used may limit the extent to which stylistic enculturation can be accurately modeled. The psychological mechanisms of statistical learning and probabilistic prediction embodied in the model and experiment measures are thought to operate implicitly, reflecting a lifetime spent listening to music without necessarily having explicit music training. It is possible that participants lacking extensive exposure to Western-classical music would show lower correspondence with predictions made using the stylistic model trained on Western-classical melodies than individuals with greater exposure. However, the schematic stylistic properties that are learned in the training of a stylistic model span a broad range of musical features—such as tonal and metrical considerations—applicable to Western music outside of the classical canon and previous research has shown that training corpora need not precisely match the stylistic experience of listeners in order to achieve sufficiently accurate simulations (Pearce, 2018). Nonetheless, the ability to closely model specific groups of listen-

ers' music experiences to provide tailored models of their stylistic expectations requires further research and production of materials.

In summary, the two experiments presented provide evidence that intra-opus statistical learning while listening to music provides a plausible psychological mechanism underlying the perception of thematic structure—in other words, that thematic structure is perceived through the statistical regularities it creates within a piece of music. The results also suggest that statistically learned properties of music are important both in the perception of musical relationships on a local timescale, but also when perceiving similar relationships, in the form of repetition, over larger timescales and, finally, when perceiving the overall unity or coherence of a piece of music. This research builds on, and contributes to, a growing existing understanding of statistically learned elements of perception, extending it to psychological processing of large-scale thematic structure in music perception.

Author Note

This research was supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1). Data and materials for the experiments in this paper have been made available in a repository at <https://osf.io/hdx4c/>. We have no known conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Edward Hall (edward.hall@qmul.ac.uk), School of Electronic Engineering and Computer Science, Queen Mary University of London. Mile End Road. London, E1 4NS, United Kingdom.

References

- AGRES, K., ABDALLAH, S., & PEARCE, M. T. (2018). Information-theoretic properties of auditory sequences dynamically influence expectation and memory. *Cognitive Science*, 42, 43–76. <https://doi.org/10.1111/cogs.12477>
- AGUS, T. R., & PRESSNITZER, D. (2013). The detection of repetitions in noise before and after perceptual learning. *Journal of the Acoustical Society of America*, 134, 464–473. <https://doi.org/10.1121/1.4807641>
- BIANCO, R., HARRISON, P. M. C., HU, M., BOLGER, C., PICKEN, S., PEARCE, M. T., & CHAIT, M. (2020). Long-term implicit memory for sequential auditory patterns in humans. *eLife*, 9, 1–6. <https://doi.org/10.7554/eLife.56073>
- BUNTON, S. (1997). Semantically motivated improvements for PPM variants. *The Computer Journal*, 40, 76–93. https://doi.org/10.1093/comjnl/40.2_and_3.76
- CAMBOUROPOULOS, E. (2016). The harmonic musical surface and two novel chord representation schemes. *Springer International Publishing*. https://doi.org/10.1007/978-3-319-25931-4_2
- CHEUNG, V. K. M., HARRISON, P. M. C., MEYER, L., PEARCE, M. T., HAYNES, J. D., & KOELSCH, S. (2019). Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology*, 29, 4084–4092.e4. <https://doi.org/10.1016/j.cub.2019.09.067>

- CLEARY, J. G., & TEAHAN, W. J. (1997). Unbounded length contexts for PPM. *The Computer Journal*, 40, 67–75. https://doi.org/10.1093/comjnl/40.2_and_3.67
- COHEN, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- DE LEEUW, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47, 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- DELIÈGE, I. (2007). Similarity relations in listening to music: How do they come into play? *Musicae Scientiae*, 11, 9–37. <https://doi.org/10.1177/1029864907011001021>
- DOWLING, W. J., MAGNER, H., & TILLMANN, B. (2016). Memory improvement with wide-awake listeners and with nonclassical guitar music. *Psychomusicology: Music, Mind, and Brain*, 26(1), 26–34. <https://doi.org/10.1037/pmu0000106>
- DOWLING, W. J., TILLMAN, B., & AYERS, D. F. (2001). Memory and the experience of hearing music. *Music Perception*, 19, 249–276. <https://doi.org/10.1525/mp.2001.19.2.249>
- EGERMANN, H., PEARCE, M. T., WIGGINS, G. A., & McADAMS, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, and Behavioral Neuroscience*, 13, 533–553. <https://doi.org/10.3758/s13415-013-0161-y>
- EITAN, Z., & GRANOT, R. Y. (2008). Growing oranges on Mozart's apple tree: "Inner form" and aesthetic judgment. *Music Perception*, 25, 397–418. <https://doi.org/10.1525/mp.2008.25.5.397>
- EITAN, Z., & GRANOT, R. Y. (2009). Primary versus secondary musical parameters and the classification of melodic motives. *Musicae Scientiae*, 13, 139–179. <https://doi.org/10.1177/102986490901300107>
- FARBOOD, M. M., HEEGER, D. J., MARCUS, G., HASSON, U., & LERNER, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Frontiers in Neuroscience*, 9, 157. <https://doi.org/10.3389/fnins.2015.00157>
- FAUL, F., ERDFELDER, E., LANG, A.-G., & BUCHNER, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191. <https://doi.org/10.3758/BF03193146>
- GINGRAS, B., PEARCE, M. T., GOODCHILD, M., DEAN, R. T., WIGGINS, G., & McADAMS, S. (2016). Linking melodic expectation to expressive performance timing and perceived musical tension. *Journal of Experimental Psychology: Human Perception and Performance*, 42, 594–609. <https://doi.org/10.1037/xhp0000141>
- GOLD, B. P., PEARCE, M. T., MAS-HERRERO, E., DAGHER, A., & ZATORRE, R. J. (2019). Predictability and uncertainty in the pleasure of music: A reward for learning? *The Journal of Neuroscience*, 39, 9397–9409. <https://doi.org/10.1523/JNEUROSCI.0428-19.2019>
- GOLDMAN, A., HARRISON, P. M. C., JACKSON, T., & PEARCE, M. T. (2021). Reassessing syntax-related ERP components using popular music chord sequences. *Music Perception*, 39, 118–144. <https://doi.org/10.1525/mp.2021.39.2.118>
- GRANOT, R. Y., & JACOBY, N. (2011). Musically puzzling I: Sensitivity to overall structure in the sonata form? *Musicae Scientiae*, 15, 365–386. <https://doi.org/10.1177/1029864911409508>
- GRANOT, R. Y., & JACOBY, N. (2012). Musically puzzling II: Sensitivity to overall structure in a Haydn E-minor sonata. *Musicae Scientiae*, 16, 67–80. <https://doi.org/10.1177/1029864911423146>
- HALL, E. T. R., & PEARCE, M. T. (2021). A model of large-scale thematic structure. *Journal of New Music Research*, 50, 220–241. <https://doi.org/10.1080/09298215.2021.1930062>
- 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. T. (2016). "If you have to ask, you'll never know": Effects of specialised stylistic expertise on predictive processing of music. *PLOS One*, 11, e0163584. <https://doi.org/10.1371/journal.pone.0163584>
- HARRISON, P. M. C., & PEARCE, M. T. (2020). Representing harmony in computational music cognition. *PsyArXiv*.
- HURON, D. B. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT Press.
- JANSSEN, B., DE HAAS, W. B., VOLK, A., & VAN KRANENBURG, P. (2014). Finding repeated patterns in music: State of knowledge, challenges, perspectives. In M. Aramaki, O. Derrien, R. Kronland-Martinet, & S. Ystad (Eds.), *Sound, music, and motion* (pp. 277–297). Springer International Publishing. https://doi.org/10.1007/978-3-319-12976-1_18
- KARNO, M., & KONEČNÍ, V. J. (1992). The effects of structural interventions in the first movement of Mozart's Symphony in G Minor K. 550 on aesthetic preference. *Music Perception*, 10, 63–72. <https://doi.org/10.2307/40285538>
- KIVY, P. (1993). *The fine art of repetition: Essays in the philosophy of music*. Cambridge University Press.
- KIVY, P. (2017). On the recent remarriage of music to philosophy. *The Journal of Aesthetics and Art Criticism*, 75, 429–438. <https://doi.org/10.1111/jaac.12402>
- LAAKSONEN, A., & LEMSTRÖM, K. (2019). Transposition and time-warp invariant algorithm for detecting repeated patterns in polyphonic music. *Sixth International Conference on Digital Libraries for Musicology* (pp. 38–42). <https://doi.org/10.1145/3358664.3358670>
- LALITTE, P., & BIGAND, E. (2006). Music in the moment? Revisiting the effect of large scale structures. *Perceptual and Motor Skills*, 103, 811–828. <https://doi.org/10.2466/PMS.103.3.811-828>

- LAMONT, A., & DIBBEN, N. (2001). Motivic structure and the perception of similarity. *Music Perception*, 18, 245–274. <https://doi.org/10.1525/mp.2001.18.3.245>
- LARTILLOT, O. (2005). Multi-dimensional motivic pattern extraction founded on adaptive redundancy filtering. *Journal of New Music Research*, 34, 375–393. <https://doi.org/10.1080/09298210600578246>
- LERDAHL, F., & JACKENDOFF, R. (1983). *A generative theory of tonal music*. MIT Press.
- LI, M., CHEN, X., LI, X., MA, B., & VITÁNYI, P. (2004). The similarity metric. *IEEE Transactions on Information Theory*, 50, 3250–3264. <https://doi.org/10.1109/TIT.2004.838101>
- MARGULIS, E. H. (2012). Musical repetition detection across multiple exposures. *Music Perception*, 29, 377–385. <https://doi.org/10.1525/mp.2012.29.4.377>
- MARGULIS, E. H. (2013). Aesthetic responses to repetition in unfamiliar music. *Empirical Studies of the Arts*, 31, 45–57. <https://doi.org/10.2190/EM.31.1.c>
- MARGULIS, E. H. (2014). *On repeat: How music plays the mind*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199990825.001.0001>
- MCADAMS, S., VINES, B. W., VIEILLARD, S., SMITH, B. K., & REYNOLDS, R. (2004). Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Perception*, 22, 297–350. <https://doi.org/10.1525/mp.2004.22.2.297>
- MCDONALD, G., & WÖLLNER, C. (2022). Appreciation of form in Bach's Well-Tempered Clavier: Effects of structural interventions on perceived coherence, pleasantness, and retrospective duration estimates. *Music Perception*, 40(2), 150–167. <https://doi.org/10.1525/mp.2022.40.2.150>
- MELKONIAN, O., REN, I. Y., SWIERSTRA, W., & VOLK, A. (2019). What constitutes a musical pattern? *Proceedings of the 7th ACM SIGPLAN International Workshop on Functional Art, Music, Modeling, and Design* (pp. 95–105). <https://doi.org/10.1145/3331543.3342587>
- MEYER, L. B. (1956). *Emotion and meaning in music*. University of Chicago Press.
- MEYER, L. B. (1973). *Explaining music: Essays and explorations*. University of California Press.
- 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, e89642. <https://doi.org/10.1371/journal.pone.0089642>
- NARMOUR, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model*. University of Chicago Press.
- OMIGIE, D., PEARCE, M. T., & STEWART, L. (2012). Tracking of pitch probabilities in congenital amusia. *Neuropsychologia*, 50, 1483–1493. <https://doi.org/10.1016/j.neuropsychologia.2012.02.034>
- OMIGIE, D., PEARCE, M. T., WILLIAMSON, V. J., & STEWART, L. (2013). Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia*, 51, 1749–1762. <https://doi.org/10.1016/j.neuropsychologia.2013.05.010>
- PEARCE, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* [Doctoral thesis, 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, 378–395. <https://doi.org/10.1111/nyas.13654>
- PEARCE, M. T., & MÜLLENSIEFEN, D. (2017). Compression-based modelling of musical similarity perception. *Journal of New Music Research*, 46, 135–155. <https://doi.org/10.1080/09298215.2017.1305419>
- PEARCE, M. T., MÜLLENSIEFEN, D., & WIGGINS, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, 39, 1367–1391. <https://doi.org/10.1068/p6507>
- 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, 302–313. <https://doi.org/10.1016/j.neuroimage.2009.12.019>
- POLLARD-GOTT, L. (1983). Emergence of thematic concepts in repeated listening to music. *Cognitive Psychology*, 15, 66–94. [https://doi.org/10.1016/0010-0285\(83\)90004-X](https://doi.org/10.1016/0010-0285(83)90004-X)
- PRINCE, J. B. (2014). Contributions of pitch contour, tonality, rhythm, and meter to melodic similarity. *Journal of Experimental Psychology: Human Perception and Performance*, 40, 2319–2337. <https://doi.org/10.1037/a0038010>
- PRINCE, J. B., THOMPSON, W. F., & SCHMUCKLER, M. A. (2009). Pitch and time, tonality and meter: How do musical dimensions combine? *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1598–1617. <https://doi.org/10.1037/a0016456>
- REN, I. Y., KOOPS, H. V., VOLK, A., & SWIERSTRA, W. (2017). In search of the consensus among musical pattern discovery algorithms. In S. J. Cunningham, Z. Duan, X. Hu, & D. Turnbull (Eds.), *Proceedings of the 18th International Society for Music Information Retrieval Conference* (pp. 671–679). ISMIR Press.
- ROLISON, J. J., & EDWORTHY, J. (2012). The role of formal structure in liking for popular music. *Music Perception*, 29, 269–284. <https://doi.org/10.1525/mp.2012.29.3.269>
- SAUVÉ, S. A., SAYED, A., DEAN, R. T., & PEARCE, M. T. (2018). Effects of pitch and timing expectancy on musical emotion. *Psychomusicology: Music, Mind, and Brain*, 28, 17–39. <https://doi.org/10.1037/pmu0000203>
- SCHOENBERG, A. (1967). *Fundamentals of musical composition* (4th ed., G. Strang & L. Stein, Eds.). Faber & Faber.

- SEARS, D. R. W., PEARCE, M. T., SPITZER, J., CAPLIN, W. E., & MCADAMS, S. (2019). Expectations for tonal cadences: Sensory and cognitive priming effects. *Quarterly Journal of Experimental Psychology*, 72, 1422–1438. <https://doi.org/10.1177/1747021818814472>
- SMYTH, D. (1993). “Balanced Interruption” and the formal repeat. *Music Theory Spectrum*, 15, 76–88. <https://doi.org/10.2307/745910>
- TAN, S.-L., & SPACKMAN, M. P. (2005). Listeners’ judgments of the musical unity of structurally altered and intact musical compositions. *Psychology of Music*, 33, 133–153. <https://doi.org/10.1177/0305735605050648>
- TAN, S.-L., SPACKMAN, M. P., & PEASLEE, C. L. (2006). The effects of repeated exposure on liking and judgments of musical unity of intact and patchwork compositions. *Music Perception*, 23, 407–421. <https://doi.org/10.1525/mp.2006.23.5.407>
- TEMPERLEY, D. (2001). *The cognition of basic musical structures*. MIT Press.
- TEMPERLEY, D. (2007). *Music and probability*. MIT Press.
- TILLMANN, B., & BIGAND, E. (1996). Does formal musical structure affect perception of musical expressiveness? *Psychology of Music*, 24, 3–17. <https://doi.org/10.1177/0305735696241002>
- TILLMANN, B., BIGAND, E., & MADURELL, F. (1998). Local versus global processing of harmonic cadences in the solution of musical puzzles. *Psychological Research*, 61(3), 157–174. <https://doi.org/10.1007/s004260050022>
- TILLMANN, B., DOWLING, W. J., LALITTE, P., MOLIN, P., SCHULZE, K., POULIN-CHARRONNAT, B., ET AL. (2013). Influence of expressive versus mechanical musical performance on short-term memory for musical excerpts. *Music Perception*, 30(4), 419–425. <https://doi.org/10.1525/mp.2013.30.4.419>
- VAN DER WEIJ, B., PEARCE, M. T., & HONING, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in Psychology*, 8, 824. <https://doi.org/10.3389/fpsyg.2017.00824>
- ZIV, N., & EITAN, Z. (2007). Themes as prototypes: Similarity judgments and categorization tasks in musical contexts. *Musicae Scientiae*, 11, 99–133. <https://doi.org/10.1177/1029864907011001051>