The perceptual relevance of balance, evenness, and entropy in musical rhythms

2 authors:

Andrew J. Milne
Western Sydney University
52 PUBLICATIONS **355** CITATIONS

SEE PROFILE

Steffen A. Herff
Western Sydney University
28 PUBLICATIONS **108** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Perception of Time-Varying Sounds View project

Musica Facta View project

The Perceptual Relevance of Balance, Evenness, and Entropy in Musical Rhythms

Andrew J. Milne[*1] and Steffen A. Herff[1,2]

[1]The MARCS Institute for Brain, Behaviour and Development, Western Sydney
University, Penrith, Australia
[2]Digital and Cognitive Musicology Lab, École polytechnique fédérale de Lausanne,
Switzerland

Abstract

There is an uncountable number of different ways of characterizing almost any given real-world stimulus. This necessitates finding stimulus features that are *perceptually relevant* – that is, they have distinct and independent effects on the perception and cognition of the stimulus. Here, we provide a theoretical framework for empirically testing the perceptual relevance of stimulus features through their association with recognition, memory bias, and æsthetic evaluation. We deploy this framework in the auditory domain to explore the perceptual relevance of three recently developed mathematical characterizations of periodic temporal patterns: *balance*, *evenness*, and *interonset interval entropy*. By modelling recognition responses and liking ratings from 177 participants listening to a total of 1,252 different musical rhythms, we obtain very strong evidence that all three features have distinct effects on the memory for, and the liking of, musical rhythms. Interonset interval entropy is a measure of the unpredictability of a rhythm derived from the distribution of its durations. Balance and evenness are both obtained from the discrete Fourier transform (DFT) of periodic patterns represented as points on the unit circle, and we introduce a teleological explanation for their perceptual relevance: the DFT coefficients representing balance and evenness are relatively robust to small random temporal perturbations and hence are coherent in noisy environments. This theory suggests further research to explore the meaning and relevance of robust coefficients such as these to the perception of patterns that are periodic in time and, possibly, space.

Keywords: music; rhythm; temporal cognition; pattern cognition; discrete Fourier transform; entropy

1 Introduction

Every day we are exposed to countless stimuli across all sensory modalities. However, what we perceive as, say, a bird song is a complex physical event that can be described with an abundance of different *features* – characterizations of their physical attributes. Some such stimulus features directly relate to our perception. For example, the frequency of air pressure oscillation relates to pitch, which is a distinct and independent percept. However, many different features are highly correlated (or otherwise dependent) such that it is impossible to independently change one feature without changing many others. Despite these physical correlations, it is still possible that they may have distinct perceptual effects; for this reason, it is both useful and difficult to identify *perceptually relevant* features that have distinct perceptual roles. In this paper, we provide a theoretical framework capable of providing an operationalizable definition of "perceptual relevance" and we use it to design and conduct experiments that test the perceptual relevance of three stimulus features in the auditory domain, two of which are highly correlated.

Within the auditory domain in particular, time plays a special role. All auditory stimuli unfold over time, and organization of sound into patterns lies at the heart of auditory information,

---

[*]Corresponding author at: The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia.
Email address: a.milne@westernsydney.edu.au

experience, and communication – temporal patterns are fundamental to the perception of speech and, of course, to the art of music. Here, we use musical rhythms in a series of experiments in order to explore how features relating to the organization of sound into *periodic* (repeating) temporal patterns may influence perception.

Rhythm, the temporal pattern of onsets, is one of the main components of music across cultures [1, 2], and any rhythm can be represented by an abundance of *rhythmic features*. These features can range from simple functions of the time that passes between events (e.g., the mean or variance of the rhythm's interonset intervals) to more complicated functions such as their predictability. However, some rhythmic features may have negligible or indistinct effects on perception, while others may be quite salient and have distinct perceptual impacts (to repeat, features that are highly correlated may still have distinct perceptual impacts). Indeed, the links between physical attributes of rhythms and their perception remains elusive. Yet, identifying the features that drive rhythm perception would provide significant insight into how exactly the brain deconstructs and analyses periodic temporal patterns. Here, we aim to closely examine three theoretically promising rhythmic features in an attempt to explore their potential relevance to rhythm perception.

In recent work, we have investigated three novel mathematical characterizations of musical rhythms and scales: *balance* [3, 4], *evenness* [5, 4], and *interonset interval entropy* [5]. Balance and evenness are, respectively, linearly related to the magnitudes of the zeroth and first coefficients of the discrete Fourier transform (DFT) of the rhythm when that rhythm is represented as points on a circle using complex numbers. Interonset interval entropy (entropy, for short) characterizes the unpredictability of interonset intervals, or sequences of interonset intervals. Entropy's perceptual relevance in particular has been of interest across different modalities [6, 7]. This is because entropy – a measure of unpredictability – is an important concept in information processing accounts of perception, and can function as a measure of stimulus complexity. These three features are detailed in Section 2.

In our own and other's previous work, including those above, the focus has been on using balance and evenness as parameters for: (a) composing and performing new musical beats, loops, and melodic hockets – as exemplified by the software application XronoMorph [8]; (b) understanding and analyzing rhythms and tonalities of existing music – such as the use of balance, and other Fourier coefficients, to identify tonal transitions in Debussy's "Les sons et les parfums tournent dans l'air du soir" [9], and the structure of a dance rhythm of the Aka Pygmy tribe [10, 4].

These mathematical characterizations have seemed fruitful in these respects [11]; however, it has not previously been tested whether these three features are *perceptually relevant*. As fully detailed in Section 3, a feature's perceptual relevance can be operationalized through its systematic and independent impact on memory or æsthetic evaluation. In this context, it means that we are testing whether balance, evenness, and entropy (a) influence listeners' memories of rhythms; (b) influence listeners' æsthetic evaluations of rhythms; (c) have independent effects on memory or independent effects on liking.

This last point is a tricky one because balance and evenness are highly correlated across the space of all possible rhythms – sets of rhythms typically have balance and evenness values with rank correlations greater than 0.93 [4] (see also Fig. 4). This means that, even if they do have independent perceptual effects, obtaining precise estimates of these in an experimental setting requires a large sample size. (For example, the commonly used "variance inflation factor" implies that the standard errors of the coefficients of two regression predictors with a correlation r are $\sqrt{1/(1-r^2)}$ greater than if they were completely uncorrelated). (An alternative approach would be to use the principal components of balance and evenness – their sum and their difference – as predictors; we do not do this because the resulting predictors lose their straightforward interpretations.)

Three experiments were conducted, each with different sets of rhythms and participants, but otherwise identical. Participants answered whether or not they had heard the rhythm before in the experiment, and gave a rating on a seven-point scale for how much they liked the rhythm. In total, 177 participants took part, 1,252 different rhythms were tested, and 24,992 observations were recorded for each type of question. (For smaller sample sizes, the above mentioned PCA approach could be advantageous.) For these data, multilevel Bayesian regressions strongly indicate that balance, evenness, and entropy have a substantial effect on memory bias, and a small but distinct impact on the liking of rhythms.

In Section 2, we define the rhythmic features (balance, evenness, and interonset interval entropy) that we investigate; in Section 3, we detail the meaning of perceptual relevance, and how recognition and æsthetic evaluation are indicators for it; in Section 4 we summarize the experi-

ments' design, aims, and hypotheses, the rhythms we use to test them, and the overall statistical approach (Bayesian multilevel modelling) used for all of the data analysis and provide a "meta-analysis" of the data for all three experiments combined (the experiments are individually detailed and modelled in Supplementary Materials, Secs. 3, 4, and 5). In the Discussion (Section 5), we summarize the overall implications of the results and we propose an interesting and novel explanation for why balance and evenness (and, more generally, the discrete Fourier transform of rhythms) are perceptually relevant features; we conclude in Section 6.

The principal contributions of this article are: (a) for the first time, we experimentally test the perceptual relevance of rhythmic balance, evenness, and interonset interval entropy; (b) we provide a theoretical framework for empirically testing perceptual relevance of stimulus features; (c) we demonstrate that each of these features has independent effects on memory and independent effects on the æsthetic evaluation of rhythms; (d) we present a new theory – the robustness of features under random perturbations – for why balance and evenness are perceptually important.

2 Rhythmic Features: Balance, Evenness, and Entropy

In this section, we define balance, evenness, and interonset interval entropy. To help ensure these rhythmic features are understandable, we provide straightforward verbal definitions as well as more formal mathematical definitions.

Before proceeding, three quick definitions are required: we use the term *beats* to refer to sounded events within a rhythm. If a rhythm consists of 7 prominent sounded events, we refer to it as having 7 beats. We use the term *pulses* to indicate the smallest equal (isochronous) division of a rhythm's period that captures all the beats. In a given rhythm, pulses may or may not be sounded in addition to the beats, although in our experiments they always were. For instance, the clave rhythm discussed in the next section and depicted in Figure 1, has 5 beats (black disks) in 16 pulses (white disks). Most musical cultures use rhythms where beats' onsets are subsets of isochronous onsets (pulses) so, from a musical standpoint, these are useful rhythms to explore. Finally, the term *interonset interval* is the amount of time between two consecutive beats or two consecutive pulses (the context will make clear).

2.1 A Circular Representation of Periodic Patterns

There is a long history of circular depictions of both musical rhythms and musical scales: from the 13th century rhythmic cycles of Safi al-Din al-Urmawi [12] and Newton's synæsthetic illustration of the diatonic scale [13], through early iterations of music psychology [14, 15, 16], to contemporary music cognition [17, 18, 19]. This is because a *periodic* (repeating) pattern – such as a musical rhythm – can be naturally and parsimoniously represented by points distributed on a circle: each point's angular position indicates an event's onset time (modulo the period), while the circularity represents the pattern's periodicity. Periodic patterns also occur in the visual domain; for example, visual rhythms are periodic in time (as surveyed in [20]), while frieze patterns are periodic in space; hence the techniques and features introduced in this section may have relevance beyond audition.

In this section, we summarize *Argand vectors* [4], which define any such circular pattern – to arbitrary precision – with complex numbers on the unit circle. Under this circular representation, balance and evenness can be intuitively understood; furthermore, the Argand vector allows balance and evenness to be straightforwardly calculated with the discrete Fourier transform.

Consider the periodic clave rhythm, which can be musically notated as ... ♩. ♩. ♩ ♩ ♩ ... (an understanding of musical notation is not necessary for this article – numerical representations are used subsequently – but an accessible summary of musical symbols and terminology is available on Wikipedia [21]). The onsets of these beats can be represented numerically as a set of fractions of the period (which is 16 quavers in duration): $X = \left\{0, \frac{3}{16}, \frac{6}{16}, \frac{10}{16}, \frac{12}{16}\right\}$. These fractions can then be straightforwardly converted into a sequence of complex numbers, ordered by angular position, by multiplying them by $2\pi i$ and taking their exponential: $\boldsymbol{z} = \left(\exp\left(\frac{0}{16} \times 2\pi i\right), \exp\left(\frac{3}{16} \times 2\pi i\right),\right.$ $\left.\exp\left(\frac{6}{16} \times 2\pi i\right), \exp\left(\frac{10}{16} \times 2\pi i\right), \exp\left(\frac{12}{16} \times 2\pi i\right)\right)$. The vector $\boldsymbol{z}$ is termed the rhythm's *Argand vector* [4]; its K entries, which are indexed by $k = 0, 1, \ldots, K-1$, specify positions on the unit circle, as illustrated in Figure 1.

This can be expressed more formally: for a periodic K-onset rhythm with period p, let $x_k = t_k/p \bmod 1$, where t_k is the time of the kth onset, and $k = 0, 1, \ldots, K-1$. Hence, each x_k is a real
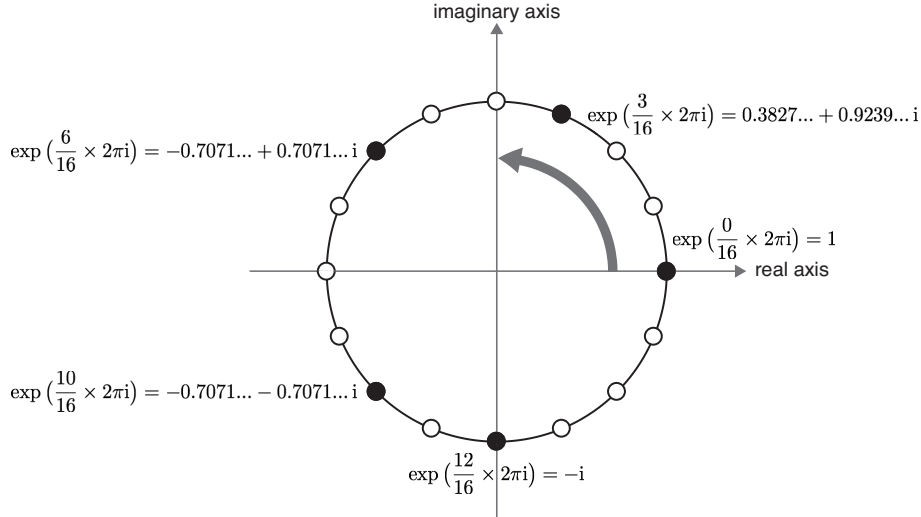
3

Figure 1: A circular representation of the clave rhythm. The rhythmic period is divided into 16 equal parts, as shown with the white circles. The clave onsets – depicted in the main text with musical notation and the numerical set X – are shown by the black disks. For a unit circle, the complex numbers in the Argand vector $\boldsymbol{z}$ precisely correspond to the beats' circular locations. The arrow is a reminder that, in the complex plane, angles increase counterclockwise – this results from the convention of having a horizontal real axis and a vertical imaginary axis.

number in the unit interval. For our features, it is of no consequence as to which onset is deemed the first (i.e., identified by $k = 0$), but they must be sequenced such that $0 \leq x_0 \leq x_1 \leq \cdots \leq x_{K-1} < 1$. The entries of the Argand vector are given by

$$z_k = \mathrm{e}^{2\pi \mathrm{i} x_k}. \tag{1}$$

For patterns of rhythmic onsets, Argand vectors have some advantages over more traditional time-domain representations. The time-domain representation of onsets is an *indicator vector*, which periodically samples for the presence or absence of an onset; for the clave rhythm above, this is the 16-entry vector $(1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0)$, or any upsample of this. Clearly, the time domain representation has a temporal resolution that is dependent on its sample rate; in this example, 1/16 of the period (notated above as quavers). For a higher temporal resolution, more entries would be required; for example, semiquaver resolution would need at least 32 entries. The Argand vector, conversely, has arbitrary timing precision because there is no sample rate; importantly, this means that subtle timings in musical rhythms (or intonations in scales) – which are common and have distinct perceptual impacts [22, 23, 24, 25, 26, 27] – can be easily represented without needing to change the entire structure of that representation (e.g., by having to change the sample rate). Crucially, despite its high precision, the Argand vector is highly parsimonious (sparse) because it requires only as many entries as there are onsets; for example, the Argand vector representation of the above clave rhythm requires only 5 entries rather than the 16 required by the indicator vector.

Despite these advantages, a possible drawback with Argand vectors is that there is no straightforward metric between them when they represent rhythms with differing numbers of onsets (because the Argand vectors will have differing dimensions); given fixed temporal resolution, indicator vectors can always be compared with a standard metric (e.g., Euclidean). This issue is easily solved by extracting scalar features directly from the Argand vectors – here, we use the features of balance, evenness, and entropy to allow Argand vectors to be directly compared, regardless of their numbers of onsets. It is well-established that "features" – simplified representations of complex phenomena – play an important role in statistical learning and memory in different modalities [28, 29, 30, 31].

For temporal pattern perception, it seems plausible that mechanisms of human cognition would gravitate towards an Argand-like representational system that is simple and sparse [32, 33], and which can be straightforwardly summarized with scalar features; hence, it seems worth investigating the relevance of balance, evenness, and entropy derived directly from the Argand vector.

2.2 Balance

The *balance* of a rhythm is simply a quantification of the proximity of that rhythm's "centre of mass" (the mean position of the points) to the centre of the unit circle. Balance is normalized to lie between 0 and 1, so that when balance is 1, the rhythm's centre of mass is at the centre of the circle (any such rhythm is denoted *perfectly balanced*); when it is 0, the rhythm's centre of mass is on the circle itself. This definition of balance allows it to be intuitively understood – one can imagine a horizontal disc with weights placed on its rim: only certain distributions of such weights will allow the disc to be perfectly balanced about its centre; some distributions will be more unbalanced than others as the centre of gravity moves away from the centre of the disc; when all the weights are on top of each other, the disc is maximally unbalanced (balance is 0). Interestingly, balance can also be understood in a different way – it is precisely equivalent to the circular variance of the rhythm, which is given by the magnitude of the circular distribution's first moment.

Mathematically, the balance of a rhythm is given by 1 minus the magnitude of the zeroth coefficient of the DFT of the rhythm's Argand vector $\boldsymbol{z}$ (see Sec. 5.2.1 and Eq. (4) for an overview of all of the DFT's coefficients). $B(\boldsymbol{z})$ denotes the balance of that rhythm, and is given by

$$B(\boldsymbol{z}) = 1 - \left| \sum_{k=0}^{K-1} z_k \right| / K$$

$$= 1 - |(\mathcal{F}\boldsymbol{z})_0|/K.$$

Because only the magnitude of the DFT coefficient is used, balance is insensitive to the phase of the rhythm and to its temporal direction. In other words, balance does not change when the rhythm is started at different onsets; nor does it change when the rhythm is played backwards.

2.3 Evenness

The *evenness* of a rhythm is a quantification of the lack of variance of its interonset intervals: if the rhythm's interonset intervals are all similar in duration, the rhythm will have high evenness; if the rhythm's interonset intervals are very different in size, its evenness will be low. Like balance, evenness is normalized to lie between 0 and 1, so that when evenness is 1, all interonset intervals are equal (such rhythms are denoted *perfectly even* or *isochronous*). It is worth noting that all perfectly even rhythms are also perfectly balanced; crucially, however, perfectly balanced rhythms are not necessarily perfectly even [4]. Unlike balance, evenness does not correspond to any statistical moment of the onsets' circular distribution; but, in circular statistical terms, it can be expressed as the circular concentration of the angular deviations between the rhythm's onsets and a discrete uniform distribution, where those angular deviations are between the kth members of both circularly ordered sets [3].

Mathematically, evenness is given by the magnitude of the first coefficient of the DFT of the rhythm's Argand vector $\boldsymbol{z}$ (see Sec. 5.2.1 and Eq. (4) for an overview of all of the DFT's coefficients). $E(\boldsymbol{z})$ denotes the evenness of that rhythm, and is given by

$$E(\boldsymbol{z}) = \left| \sum_{k=0}^{K-1} z_k \, e^{-2\pi i k/K} \right| / K$$

$$= |(\mathcal{F}\boldsymbol{z})_1|/K.$$

As with balance, the use of only the magnitude of the coefficient implies that evenness is insensitive to the phase of the rhythm and its temporal direction.

2.4 Interonset Interval Entropy

Entropy quantifies the unpredictability of a probability mass function. *Normalized entropy* is a unit-less form of entropy suitable for comparisons across probability mass functions with differing supports. The normalized entropy of the N-entry probability mass function $\boldsymbol{x}$ is given by $H(\boldsymbol{x}) = -\sum_{n=1}^{N} x_n \log(x_n)/\log(N)$, where $\log(N)$ is the maximum possible entropy. If the mass function is dominated by a small number of different outcomes with relatively high probabilities, new outcomes are easier to predict correctly and the entropy is low; if the mass function has a large number of different outcomes with similar probabilities, new outcomes are harder to predict correctly and the entropy is high.

In order to apply entropy to an object of perception (such as a rhythm), it is necessary to somehow express that object as a probability mass function. Typically there is no single way to do this, but it would seem most useful to choose probabilities of features that are likely to be perceptually salient. In rhythms, one such feature is the set of *interonset intervals* (IOIs) in the rhythm – the durations between consecutive onsets. In the Argand vector representation, these IOIs are the set of angular differences between successive elements of $\boldsymbol{z}$, which is all $\phi\big(z_{k+1}z_k^{-1}\big)$ for $k = 0, 1, \ldots, K - 1$, where ϕ is the phase (argument) of the complex number, and subscripts are taken modulo K. Given that every rhythm tested here had beats occurring at a subset of sounded regular pulses (as outlined at the start of this section), we define the support as the set of all $\phi(z_n)$ for $n = 0, 1, \ldots, N - 1$, where N is the number of regular pulses in the rhythmic period. (In the absence of a sounded pulse, a continuous probability distribution would be an alternative approach.) For example, in the clave rhythm shown above (Sec. 2.2), there are: no IOIs of duration $0/16$; no IOIs of duration $1/16$; one IOI of duration $2/16$; two IOIs of duration $3/16$; two IOIs of duration $4/16$; and no IOIs of longer durations. Taking account of the fact that there are a total of $K = 5$ onsets, this defines a probability mass function over the $N = 16$ possible IOI durations in the rhythm: $\boldsymbol{x} = (0, 0, 1, 2, 2, 0, 0, \ldots, 0)/5$. The normalized entropy of this probability mass function can then be calculated using the equation in the previous paragraph.

Like balance and evenness, IOI entropy is invariant with respect to phase and temporal direction. Unlike balance and evenness, entropy is also invariant with respect to the precise durations of the interonset intervals so long as the ranking of their durations does not change. It is also invariant to the order in which the different IOIs occur. For example, the periodic rhythms $\ldots$ ♩. ♩. ♩♩♩ $\ldots$, $\ldots$ ♩. ♩. ♩♩♩ $\ldots$, and $\ldots$ ♩♩𝅝 ♪ 𝅝 $\ldots$, have identical IOI entropies.

2.5 Feature Summary and Limitations

Clearly, the phase blindness of all three features – balance, evenness, and IOI entropy – means that they are unlikely to provide a complete description of the rhythmic percept; for instance, starting a given periodic rhythm at different beats often has a strong impact on where a listener assigns the metrical downbeat (weight) and, therefore, the "feel" of the rhythm. Yet, despite this limitation, we will show that these phase invariant features impact on rhythm perception.

In summary, it is worth noting that the three features – balance, evenness, and IOI entropy – are calculated based on the number of pulses, number of beats, and their temporal relation to one another. As a result, the perceptual relevance of these features needs to be critically investigated under several different beat and pulse configurations so as to determine how general their effects are. In this article, we test four different configurations over three experiments, as detailed in Section 4.1. In the next section, we elaborate on how we operationalize *perceptual relevance*.

3 Perceptual Relevance: Recognition and Liking

The concept of perceptual relevance aims at an often overlooked detail of high-level theories of perception. For example, Gestalt and related theories describe the perceptual grouping of stimulus features [34, 35] but, crucially, before we can generalize across stimulus features in order to achieve grouping, we need to understand what those stimulus features actually are. Stimulus features, such as spectral power distributions (that translate into colour perception) or sound pressure level (that contributes to loudness perception) have been discovered and are well defined, yet it would be ignorant to assume that we already have a comprehensive understanding of all stimulus features that are relevant to human perception. For example, research has only recently established the perceptual relevance of quasi-stationary sinusoidal components in acoustic object identification [36].

Important work in the fields of empirical aesthetics [37, 6, 38] and item memorability [39, 40] approach perceptual relevance from a behavioural perspective by establishing a feature's systematic effect on observers' responses. This approach seems intuitive; however, findings can be difficult to integrate with one another due to differences in methodology. We propose a theoretical framework that allows us to make an operationalizable definition of perceptual relevance – a definition that implies necessary requirements for an experiment to test the perceptual relevance of a set of features. This framework – detailed below – is domain-independent; we will then apply this framework in the field of musical rhythm perception.

Below we argue that of all possible systematic responses, those related to stimulus evaluation (e.g., liking) and memory (e.g., recognition) are particularly important [41]. The decision to operationalize perceptual relevance through measures of evaluation and memory is informed by the theoretical framework of the *regenerative multiple representations (RMR) conjecture* [42].

The RMR conjecture assumes a crucial link between prior experience, perception, and new memory formation. Over the course of time, observers learn a weighting of the importance of all possible stimulus features (also see Pertinence in [43, 44]). The RMR conjecture also states that, informed by prior experience, specific features can be combined to create a perception of a new higher-order feature. For example, listening experience within a specific musical tradition may lead to a sequence of pitches with specific interonset intervals being perceptually merged into the integrated percept of a melody [42]. The resulting percept – for example, a melody – can then form a new memory and can be evaluated in the light of prior experience. Unfortunately, the actual feature weightings are hidden and cannot be measured directly. However, as perception interacts with memory and prior experience, the points of interaction offer markers of perceptual relevance. In other words, if a feature affects memory responses, or responses that relate to stimulus evaluation in light of prior experience, then it can be deduced that this feature has a non-zero perceptual weighting [41]. Here, we deem features with a weighting strong enough to evoke a systematic response in listeners to be *perceptually relevant*. A schematic summary of the RMR framework is depicted in Figure 2.
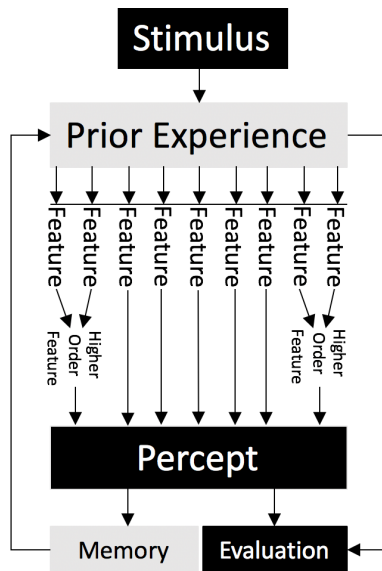

Figure 2: A schematic summary of the RMR conjecture. Prior experience provides information about how to weight and combine different stimulus features. Taken together, this forms a percept that can be evaluated and be subjected to memory, a system concerned with the encoding and retrieval of information. If stimulus features such as balance, evenness, entropy, or their interactions, show systematic effects on evaluation (liking) or memory (recognition) or both, this indicates that those features are weighted sufficiently to be of perceptual relevance. In turn, memory can become part of experience, a system that is concerned with integrating any form of information to inform behaviour, cognition, and perception. Experience incorporates memory but also other sources of information, for example sensory habituation [45].

A new memory is informed by the percept. As a result, any feature that (either directly or after being merged into a higher-order feature) shows a systematic influence on memory must have been perceived and have perceptual relevance. It is important to note that a response is necessarily task and domain specific. This means that perceptual relevance can only be established within a given task and domain. For example, a feature that is relevant in auditory stimuli, may be irrelevant in visual stimuli and some features may have little perceptual relevance to a given task whereas the same feature is important to another task.

In the present study, we use a *recognition* rather than a *recall* task to measure memory. Recognition refers to the identification of a stimulus as a stimulus that has been previously encountered, whereas recall refers to the retrieval and reproduction of a previously encountered stimulus. Importantly, recognition tasks allow the separate quantification of *sensitivity*, *specificity*, and underlying memory *biases*. Sensitivity refers to the ability to detect stimuli that have been presented before. Specificity refers to the ability to detect stimuli that have not been presented before. Here, mem-

ory bias refers to a greater or lesser tendency to think a specific stimulus has occurred previously, whether or not it actually has.

For the present experiments, measuring recognition means that if balance, evenness, or IOI entropy systematically influence recognition responses, then they are of relevance to rhythm perception. Importantly, the same criteria can be used to determine whether different features are perceptually distinct. For instance, if one feature systematically increases memory performance or a memory bias, whereas another decreases it, then this is evidence that these two features are both perceptually relevant, but are meaningfully different from one another.

However, it is possible for a feature to be of perceptual relevance to a task despite not being consciously remembered. For example, if an observer systematically prefers blue paintings over red paintings, despite not remembering the color when prompted, then this is still evidence that color is perceptually relevant. Within the framework of the RMR conjecture, perceptual weightings determine the percept. Any form of evaluation of this percept always takes place in light of prior experience. If a feature systematically influences evaluation, then this is evidence that the feature must have influenced percept formation and that this feature, informed by prior experience, is of perceptual relevance. As a result, another way to establish perceptual relevance of a stimulus feature is to demonstrate a systematic influence of that feature on observers' evaluation of a stimulus. For the present study, we use liking to capture evaluation. We chose liking because we consider liking to be great importance to day-to-day music perception and easier to relate to other non-rhythmic stimuli compared to other forms of evaluation, for example perceived complexity or "groove". For the present experiments, this means that if evenness, balance, or IOI entropy show a systematic effect on liking, they are of relevance to rhythm perception.

In summary, in three experiments, we assess the relevance of evenness, balance, and IOI entropy to rhythm perception by measuring both the liking and the recognition of musical rhythms that systematically vary in these three features. Liking and recognition can both potentially establish the relevance of a feature to rhythm perception. However, it is possible to remember a stimulus feature without showing a systematic æsthetic response to it and vice versa so, by measuring both, we compensate for the shortcomings of the respective other.

3.1 Recognition: When a Feature Affects Memory Formation

Our intention is to measure the effects of evenness, balance, and entropy on recognition in order to assess their perceptual relevance. To measure recognition, we deploy a *continuous recognition paradigm* [46]. The continuous recognition paradigm is well established within [42, 47, 48, 49, 50, 51, 52] and outside the domain of music [53, 54, 46, 55].

In each trial, participants listen to stimuli (here rhythms) some of which have been presented before ($IsOld = 1$), and some of which have not been presented before ($IsOld = 0$). After each stimulus presentation, the participant indicates whether they think they have heard this stimulus in this experiment before. The participant can either respond "old" ($SayOld = 1$), if they think they have heard this rhythm in this experiment before; alternatively, they can respond "new" ($SayOld = 0$), if they think this rhythm has not been previously presented in this experiment. Depending on the combination of $SayOld$ and $IsOld$, four possible response outcomes can be observed, as depicted in Figure 3.

	IsOld = 0	IsOld = 1
SayOld = 0	Correct Rejection (CR)	Miss (M)
SayOld = 1	False Alarm (FA)	Hit (H)
Rates (R)	C**R**R = CR / (CR + FA) FA**R** = FA / (CR + FA)	M**R** = M / (M + H) H**R** = H / (M + H)

Figure 3: Response contingency table in a recognition task. The table shows the four possible outcomes in a recognition task: *correct rejection* (CR), *miss* (M), *false alarm* (FA), *hit* (H). These are recorded as *counts* (non-negative integers) across a set of observations. The bottom row shows how these counts are converted into *rates* (proportions between 0 and 1). For example, to calculate the *correct rejection rate* (CRR), correct rejections (CR) need to be divided by number of responses given when $IsOld = 0$; that is, the sum of correct rejections (CR) and false alarms (FA).

Within signal detection theory (SDT), the *rates* of the four outcomes in a recognition task –

shown in Figure 3 – can be fully characterized by two variables: *discriminability* (also known as d', sometimes referred to as *sensitivity*) and *bias* (also known as the *criterion*). Discriminability is a measure of how well old rhythms can be correctly distinguished from new ones. In SDT, discriminability is calculated by subtracting the z-scored false alarm rate (FAR) from the z-scored hit rate (HR); that is, $discriminability = \Phi^{-1}(HR) - \Phi^{-1}(FAR)$ (note that Φ^{-1}, the probit function, converts probabilities into z scores) [56].

Bias reflects the overall tendency to choose one answer ("old" or "new") over the other, regardless of whether or not the to-be-classified stimulus has been encountered before. In SDT, bias is calculated as half the negative of the sum of the z-scored FAR and the z-scored HR; that is, $bias = -\frac{\Phi^{-1}(HR) + \Phi^{-1}(FAR)}{2}$ [56]. A negative bias is a greater tendency to say "old", and a positive bias is a lesser tendency to say "old", regardless of anything else [57]. For example, given a completely unrecognizable stimulus: when bias is negative, there will be more "old" than "new" responses; when bias is positive there will be more "new" than "old" responses; when bias is zero, there will be an equal number of "old" and "new" responses.

There are at least two possible mechanisms that may account for a negative bias (greater tendency to say "old") induced by a given rhythmic stimulus. First, such a rhythm may be familiar to the participant – perhaps because it is similar to a rhythm they have heard many times before in their day-to-day life. Such a participant may, therefore, identify the rhythm as having occurred before in the experiment when, in fact, it is has actually occurred outside the experiment. They are likely, therefore, to respond "old" whether or not that rhythm has previously occurred in the experiment. Clearly, if participants are members of a shared musical culture, these biases will become population-level effects. Secondly, a negative-bias-inducing rhythm may be perceptually clustered with other rhythms. A participant may find rhythms within such perceptual clusters hard to distinguish and so will think they have been heard before, both when they have been heard before and when a different member of the cluster has been heard before. In both cases, perception was affected, yet the effect would only be captured in the response bias, rather than memory performance. As a result, any systematic change in either bias or discriminability induced by a feature is indicative of the perceptual relevance of that feature.

To model the data, we use a binary probit model, which is a generalized linear model with a Bernoulli probability distribution and a probit link function. This is a common choice for modelling binary data and is closely related to a logistic model, which uses a logit link instead of a probit link. The binary probit model assumes that binary responses $y \in \{0, 1\}$ (here, saying "new" is coded 0, and "old" is coded 1) are motivated by a normally distributed latent (unobserved) random variable $y^* \in \mathbb{R}$. Due to the probit link, this latent variable is the z-scored probability of saying "old", and can be thought of as the real-valued "utility" of saying "old"; the mean of y^* is predicted by a linear combination of known predictors x_j with unknown weights ("effects") β_j such that, for the ith observation,

$$y_i = \begin{cases} 0, & y_i^* < 0 \\ 1, & y_i^* > 0 \end{cases}$$
$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \varepsilon_i, \text{ and}$$
$$\varepsilon_i \sim \mathcal{N}(0, 1).$$

The error term follows a standard normal distribution; hence, for a fitted model, a unit change in x_j predicts that, all else being equal, y^* (the z-scored probability of saying "old") changes by β_j; crucially, it is this that allows the following model to estimate the effect of its predictors on discriminability and bias.

Our model takes the general form shown in (2), where x is the value of a feature of interest such as evenness, balance, entropy, trial number, or one of their interactions (there may be multiple such predictors included in the model but, for simplicity, only one is shown in the equation below), and *IsOld* is a binary variable coded 0 if the rhythm is new and 1 if it is old:

$$y^* = \beta_0 + \beta_1 x + \beta_2 IsOld + \beta_3 (x \times IsOld) + \varepsilon. \tag{2}$$

The above model's coefficients β_j allow us to estimate how the predictor x influences HR, MR,

CRR, and FAR, as well as discriminability and bias:

$$\Phi^{-1}(HR) = discriminability/2 - bias \qquad = \beta_0 + \beta_1 x + \beta_2 + \beta_3 x + \varepsilon$$
$$\Phi^{-1}(MR) = bias - discriminability/2 \qquad = -(\beta_0 + \beta_1 x + \beta_2 + \beta_3 x + \varepsilon)$$
$$\Phi^{-1}(CRR) = bias + discriminability/2 \qquad = -(\beta_0 + \beta_1 x + \varepsilon)$$
$$\Phi^{-1}(FAR) = -(bias + discriminability/2) = \beta_0 + \beta_1 x + \varepsilon$$

while

$$discriminability = \Phi^{-1}(HR) - \Phi^{-1}(FAR) \qquad = \beta_2 + \beta_3 x + \varepsilon$$
$$bias = -\frac{\Phi^{-1}(HR) + \Phi^{-1}(FAR)}{2} \qquad = -\left(\beta_0 + \frac{\beta_2}{2} + (\beta_1 + \frac{\beta_3}{2})x + \varepsilon\right). \qquad (3)$$

To understand these equations, it may be useful to note that, when $IsOld = 0$, any predictor term involving $IsOld$ must equal 0 and so can be dropped; when $IsOld = 1$, it multiplies by 1 so can be omitted. Furthermore, $\Phi^{-1}(p) = -\Phi^{-1}(1-p)$, which implies $\Phi^{-1}(HR) = -\Phi^{-1}(MR)$ and $\Phi^{-1}(FAR) = -\Phi^{-1}(CRR)$.

As stated above, a systematic change in either the discriminability or bias induced by x is indicative of the perceptual relevance of x. Crucially, this implies that if either $\beta_1 \neq 0$ or $\beta_3 \neq 0$, then x is perceptually relevant.

As detailed in Section 4.3, our principal model of memory responses includes more than just the single predictor x shown in (2). The model includes all interactions and conditional main effects of balance, evenness, interonset interval entropy, and $IsOld$, along with the the conditional main effect of trial number and its interaction with $IsOld$. For clarity, this model is expressed below in the familiar model formula syntax used in many statistical software packages (in this syntax, $\sim$ means "is modelled by", the $*$ operator stands for all possible interactions and conditional main effects (e.g., $a * b * c$ means $a + b + c + (a \times b) + (a \times c) + (b \times c) + (a \times b \times c)$), and $(a \mid d)$ is a group-level (random) effect of a grouped by the categorical variable d):

```
SayOld ~ (Balance * Evenness * Entropy + Trial) * IsOld
       + ((Balance * Evenness * Entropy + Trial) * IsOld | ID)
       + (Trial * IsOld | RhythmID).
```

Note that this model has a maximal random effects structure [58] in that all within-subject variables are allowed to vary between participants (ID), and all within-item variables – "items", here, being rhythms ($RhythmID$) – are allowed to vary between rhythms. All non-binary variables are standardized. The features of balance, evenness, and interonset interval entropy and their interactions are included to test their perceptual relevance. $Trial$ and its interaction with $IsOld$ are also important to include because it is likely that participants will correctly figure out that repetitions are more likely as the experiment continues; omitting it would lead to omitted variable bias (because $Trial$ causally effects both the predictor $IsOld$ and the outcome $SayOld$), which would incorrectly increase the estimated positive effect of $IsOld$, and its interactions, on $SayOld$; it is only by including both $Trial$ and $IsOld$ in the regression that their direct causal effects can be estimated [59].

As detailed in Section 4.3, the above model – which has a three-way interaction between balance, evenness, and entropy – performed better (for out-of-sample predictions) than alternatives that used additional squared terms of those three variables either with and without their interactions.

3.2 Liking: When a Feature Affects Evaluation

Here, we deem features perceptually relevant if they show a systematic effect on memory or evaluation. Above, we described how we capture the perceptual relevance of evenness, balance, and entropy in the context of *memory*. In order to assess the perceptual relevance of evenness, balance, and entropy on *evaluation*, we collect liking ratings after each stimulus presentation. Here, liking is operationalized in form of a 7-point horizontal Likert-item labeled "Strongly Dislike" on the left, "Neutral" in the middle, and "Strongly Like" on the right.

To model the data, we use an ordinal probit model, which is a generalized linear model using a cumulative normal distribution with a probit link. This type of model is analogous to the previously discussed binary probit model in that it assumes the existence of a normally distributed latent variable $y^* \in \mathbb{R}$ that underlies the observed discrete Likert responses $y \in \{1, 2, 3, 4, 5, 6, 7\}$.

However, to capture the seven different response options (rather than the two options in the recognition model), there are six additional "cutpoint" parameters θ_1 to θ_6 (in the summaries of the fitted models, these are denoted "intercept[1]" to "intercept[6]", etc.). As before, the mean of y^* is predicted by a linear combination of known predictors x_j with unknown weights ("effects") β_j such that, for the ith observation,

$$y_i = \begin{cases} 1, & y_i^* \leq \theta_1 \\ 2, & \theta_1 < y_i^* \leq \theta_2 \\ 3, & \theta_2 < y_i^* \leq \theta_3 \\ & \vdots \\ 7, & \theta_6 < y_i^*. \end{cases}$$
$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \varepsilon_i, \text{ and}$$
$$\varepsilon_i \sim \mathcal{N}(0, 1).$$

The error term follows a standard normal distribution; hence, for a fitted model, a unit increase in x_j predicts that, all else being equal, the underlying liking variable increases by β_j standard deviations (put differently, the z-score of the latent liking variable increases by β_j). Hence the interpretation of the model and its coefficients are straightforward.

As detailed in Section 4.3, our principal model for liking takes the same form as the memory model – it includes all interactions and conditional main effects of balance, evenness, interonset interval entropy, and *IsOld*, along with the the conditional main effect of trial number and its interaction with *IsOld*. It also uses a maximal random effects structure and all non-binary variables are standardized:

```
Liking ~ (Balance * Evenness * Entropy + Trial) * IsOld
       + ((Balance * Evenness * Entropy + Trial) * IsOld | ID)
       + (Trial * IsOld | RhythmID).
```

The features of balance, evenness, and interonset interval entropy and their interactions are included to test their perceptual relevance. *IsOld* is included in the liking model to control for the *mere exposure* effect – a well-established phenomenon of increased liking for previously encountered stimuli [57]. For this reason, liking is sometimes considered to be a proxy for mere exposure and, hence, *implicit memory* (which refers to when a person has no awareness of memory retrieval) [60]; this differs from *explicit memory* (which refers to when a person is aware of the memory retrieval), which is directly measured by the previously described recognition task. *Trial* is included to control for the effect of *boredom* as the experiment progresses, which may lead to participants giving progressively lower ratings. As with the *SayOld* model, it is only by including both *Trial* and *IsOld* in the regression that their direct effects can be estimated without bias. Mere exposure and boredom effects do not support the perceptual relevance of evenness, balance, or entropy. However, if a feature systematically moderates, or is moderated by, the mere exposure effect, then this indicates that the feature influences evaluation and so is perceptually relevant. Clearly, any feature that shows a direct systematic effect on liking ratings also has perceptual relevance. As with the memory model, we included the interaction between balance and evenness but not their squared terms (see Section 4.3).

4 Experiments 1–3

In order to assess the perceptual relevance of balance, evenness and interonset interval entropy, three experiments were conducted. As detailed below, these experiments differed only in numbers of participants, and the actual rhythms presented. Before detailing the methods used, we first explain the choices of rhythms made.

4.1 Choices of Rhythms

Our aim is to estimate the individual perceptual relevances of balance, evenness, and entropy; however, as can be seen in the scatter plot in Figure 4, balance and evenness are highly correlated. This cannot be avoided and, as noted in Section 1, is why it was anticipated that so much data would
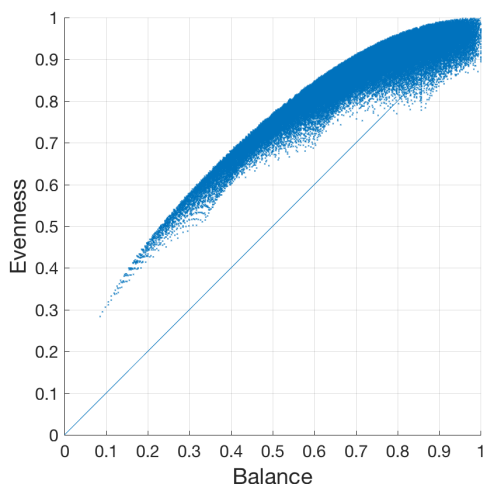
Figure 4: The balance and evenness of all 146,777 primitive rhythms (up to rotation and reflection) that are 7-beat subsets of 8, 9, . . . , 30 pulses (equal divisions of the rhythmic period). Similar results occur with differing numbers of beats. As a visual reference, the blue diagonal line passes through all points where balance and evenness are equal.

be required to accurately assess whether balance and evenness have independent effects on memory and liking. In Experiments 1 and 2, rhythms were selected so that many had almost identical balance values but widely different evenness values, and many had almost identical evenness values but widely differing balance values. Visually, this can be gleaned in Supplementary Materials Figures 7 and 14 from the way that the selected rhythms are arranged in vertical and horizontal "strips". This was done to allow for the possibly independent effects of balance and evenness to be investigated with precision at approximately constant values either of balance or of evenness.

To establish whether evenness, balance, and entropy have consistent perceptual effects across differing numbers of pulses and beats, and across familiar and unfamiliar metrical structures, we tested four broad categories of rhythms. In Experiment 1, we tested rhythms with 7 beats in 24 pulses. In Experiment 2 we tested 7 beats in 18 pulses. In Experiment 3 we used rhythms with either 6 or 7 beats in 23 pulses.

For Experiment 1, we chose a 24-pulse period for two reasons. First, there are an unusually large number of distinct perfectly balanced 7-beat rhythms that are periodic subsets of 12, 18, 20, 24, 28, . . . pulses (see [4] for a continuation of this list). Including perfectly balanced rhythms is particularly compelling because, by definition, they have the maximal possible balance values and so may be particularly influential; furthermore, they are generalizations of the polyrhythms found in much Sub-Saharan African music, as detailed in [4]. Secondly, 24 is a familiar rhythmic period – for example, it corresponds to the number of quavers (eighth notes) in two bars of $\frac{12}{8}$, or the number of semiquavers (sixteenth notes) in two bars of $\frac{3}{4}$, or the number of quavers in three bars of $\frac{4}{4}$ (bear in mind that rhythmic patterns commonly extend over more than one bar).

In Experiment 2, we tested 7 beats in 18 pulses. As with 24-pulse rhythms, 18 pulses also allow for a large number of perfectly balanced rhythms (see the above list) and form a relatively common metrical periodicity. Experiment 2, therefore, functions as an extension of Experiment 1. Experiment 2 allows us to assess whether any potential perceptual relevances of balance, evenness, and entropy established in Experiment 1, are limited to the particular beat and pulse configuration used in Experiment 1.

In Experiment 3, we explore 6 and 7 beats in 23 pulses. We use both 6 and 7 beats to reduce between-rhythm similarity in Experiment 3. Lower between-rhythm similarity effectively reduces the difficulty of the memory task. The present task is very difficult, and reducing the difficulty may increase the effects of the features on memory performance. This, for example, could manifest in reduced fatigue or less performance decline as the experiment progresses. Furthermore, rhythms in 23 pulses are very uncommon compared to rhythms in 24 (Experiment 1) or 18 (Experiment 2) pulses. Therefore, Experiment 3 allows us to see if the perceptual relevance of evenness, balance, and entropy vary as a result the commonness (hence familiarity) of the number of pulses in the rhythmic period.

All of the rhythmic patterns used in the experiments are known as *primitive*, which means they

do not have rotational symmetry when arranged in a circle, as in Figure 1; that is, rotating it by any amount other than a (multiple of) a whole turn never results in precisely the same set of onset times. The reason for avoiding non-primitive rhythms is because they are actually periodic over a smaller interval than the desired period. For example, Figure 5 shows two rhythms: one without rotational symmetry, and one with 3-fold rotational symmetry – we can see that the latter rhythm has a true period of 8 pulses rather than the desired 24 pulses.

(a) A primitive 6-beat rhythm in 24 pulses. The only rotation over which the rhythm is invariant is a full turn.

(b) A non-primitive 6-beat rhythm in 24 pulses. This rhythm is the same when rotated by any multiple of 8 pulses (one third of a turn).

(c) The same rhythm as in (b), but represented over its true 8-pulse period – it is now primitive.
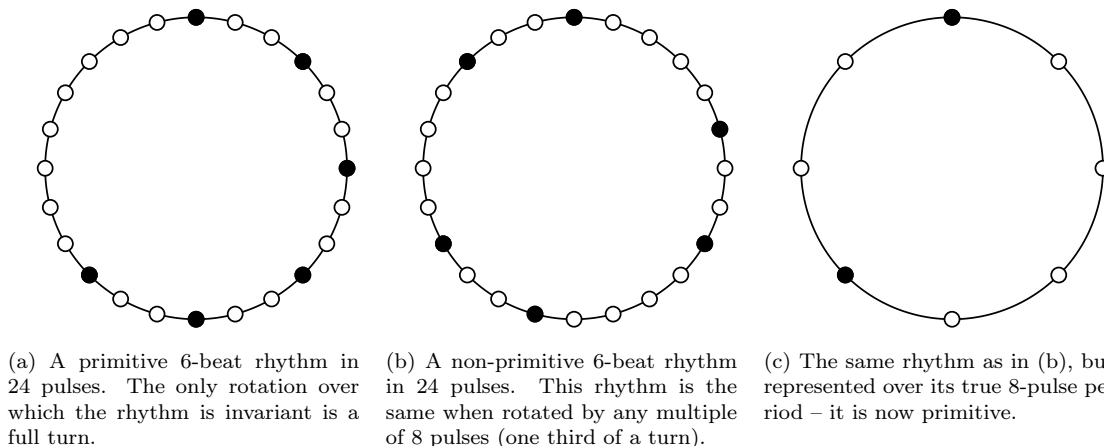
Figure 5: The rhythm in (a) is primitive; the rhythm in (b) is non-primitive because a rotation by less than a full turn (here 8 pulses or one third of a turn) results in exactly the same pattern; the rhythm in (b) in its primitive form. Note that the balance of the rhythm as represented in (b) is not the same as the balance of the rhythm as represented in (c); neither is its evenness. Hence it is useful to ensure every rhythm is written in its primitive form; that is, it is circular with respect to its smallest period.

As detailed in Section 4.3, the effects of interest did not meaningfully differ between the three experiments. This means it makes sense for us to focus on all three experiments as a single set of data. Doing so ensures our estimates of the effects' sizes are as tight as possible, and also simplifies the exposition. The individual experiments, and models fitted to them, are detailed in Supplementary Materials, Sections 3, 4, and 5.

4.2 Method

The method used in all three experiments was the same except for the numbers of participants and the sets of rhythms used, as outlined below.

4.2.1 Participants

Both Experiments 1 and 2 had 62 participants each, while Experiment 3 had 53 participants (details are provided in Supplementary Materials, Sections 3.1.1, 4.1.1, and 5.1.1). Across all three experiments, there were 177 participants: 36 male, 141 female, with a mean years-of-age of 22.1 with a standard deviation of 6.3. Expertise can have a large impact on perception. To account for a possible influence of musical expertise we deployed the Goldsmith musical sophistication index (GMSI). The GMSI is an index of musical sophistication derived from subscales measuring active musical engagement, music perceptual abilities, musical training, emotional engagement to music, and singing abilities, such that individuals with high GMSI undertake more varied musical behaviours more frequently and execute them with greater ease [61]. The present sample showed a mean GMSI of 65.3 with a standard deviation of 15.0. The GMSI score was relatively homogeneous across the three experiments and, interestingly, 0.8 standard deviations lower than the mean of 81.6 obtained from a large sample (147,663) of British participants [61]. All participants were first-year students, currently enrolled in a Psychology undergraduate degree. Recruitment took place through Western Sydney University's online system. Participation was reimbursed with course credits. Participants were required to have normal or corrected hearing. All experiments were approved by Western Sydney University's Human Research Ethics Committee (reference number H12030).

4.2.2 Materials

A total of 1,252 different rhythms were tested. Every such rhythm comprised two different musical sounds that we here term *pulses* and *beats*. The pulses were sounded by a synthesized hi-hat cymbal and the beats were sounded by a synthesized drum (detailed subsequently). The pulses were the same for every trial, and comprised isochronous (perfectly even) onsets with an interonset interval of 117 ms, which is very close to the speed of a semi-quaver (sixteenth note) in music with a tempo of 128 beats per minute (a common tempo in pop music [62]). In each trial, the beats were sounded at different subsets of the pulses, as now detailed.

In the first experiment, the beats were sounded at 136 different subsets of seven of the just mentioned pulses, and each such sequence of beats repeated every 24 pulses. In the second experiment, the beats were sounded at 144 different subsets of seven of the pulses, and each such sequence of beats repeated every 18 pulses. In the third experiment, the beats were sounded at 972 different subsets of either six or seven of the pulses, and each such sequence of beats repeated every 23 pulses. Figure 6 shows the relationships between the *balance*, *evenness*, and *entropy* values for all of the tested rhythms (more details, including scatter plots, for the sets of rhythms used in each of the three experiments can be found in Supplementary Materials, Sections 3.1.2, 4.1.2, and 4.1.2).
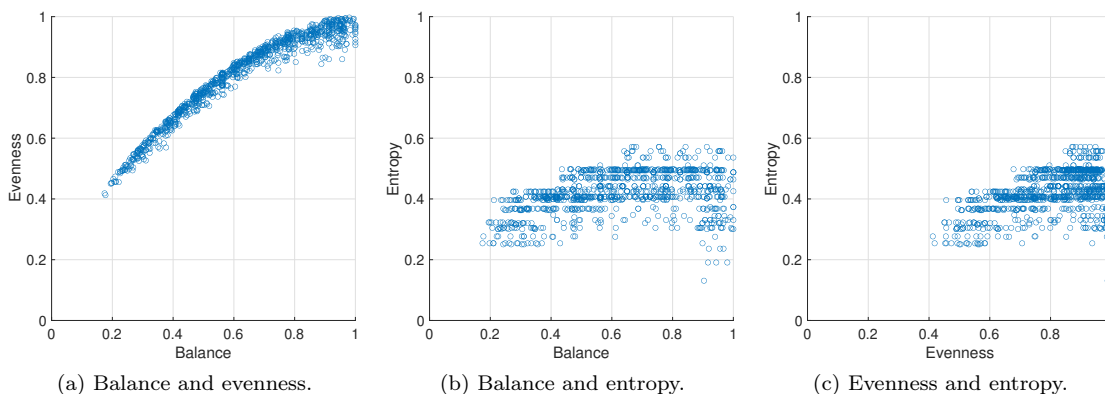

Figure 6: Scatter plots showing the relationships between the balance, evenness, and entropy values in the 1,252 different 7-in-24, 7-in-18, 6-in-23 and 7-in-23 rhythms tested in all three experiments: (a) shows the relationship between balance and evenness; (b) shows the relationship between balance and entropy; (c) shows the relationship between evenness and entropy.

The rhythmic sounds were synthesized by Chromaphone 2 [63], which is a commercial physical modelling synthesizer capable of producing naturalistic percussion sounds. Crucially, for rhythms with many fast repetitions, these do not suffer from an artificiality – the so-called "machine gun" effect – caused by the unnatural lack of timbral variance over repeated strikes when using sampled drum sounds. Physical modelling synthesizers produce subtle timbral changes with every strike, in the same way as a physical drum. Natural percussion sounds were chosen over the ramped sinusoids that are sometimes used in rhythm experiments for two reasons: (a) to ensure the stimuli had greater ecological validity and (b) to make them relatively pleasurable to listen to.

Two different percussion sounds were used in every trial: a short (closed) *hi-hat* cymbal sound, which was slightly adapted from Chromaphone's built-in preset "Hi-Hat Trick 5"; a sound from a modelled tubular *drum* with a skin head, which was slightly adapted from the "Arabian Perc" preset. The first author chose these sounds because he thought they both had clear (sharp) onsets, they were not reminiscent (particularly in combination) of any specific musical genre or cultural origin, and they both had a sonically appealing sound quality.

4.2.3 Procedure

Prior to listening to the stimuli, participants read an information sheet, signed a consent form and completed an online questionnaire, which consists of demographic questions and the Goldsmiths Musical Sophistication Index [61]. Each experiment lasted approximately 40–45 minutes, and the questionnaire took an additional 5–10 minutes to complete. Participants were advised they could take a break at any time.

The 62 participants in Experiment 1, were randomly allocated to two groups (one of size 32,

one of size 30). Each group listened to a distinct set of 68 different rhythms, each heard twice. For each participant, each rhythm was started with one of its seven beats and in a random phase; that is, there were seven different beats that might have been sounded first for each rhythm. For each given participant, the second presentation of the rhythm was in the same phase as the first presentation. As discussed in Section 2, all of our features are insensitive to phase differences, which is why we randomized over phase, so as to minimize potential confounds associated with differing phases. Each rhythm had a period of $117\,\text{ms/pulse} \times 24\,\text{pulses} = 2{,}808\,\text{ms}$ and was played five times through (a total of $14{,}040\,\text{ms}$), after which the participant gave their responses, as described subsequently.

The procedure of Experiment 2 closely followed that of Experiment 1. There were just two substantive differences. As before, the 62 participants were split into two groups (one of size 32, the other of 30 size) but, this time, each group listened to a distinct set of 72 different rhythms, rather than 68. Secondly, each rhythm had a period of $117\,\text{ms/pulse} \times 18\,\text{pulses} = 2{,}106\,\text{ms}$ and was played five times through (a total of $10{,}530\,\text{ms}$) prior to the participant giving their responses.

The procedure of Experiment 3 also closely followed Experiment 1. There were three substantive differences. First, the 53 participants were split into two groups of 28 and 25. Secondly, each group listened to rhythms randomly adapted from the same set of 72 rhythms, but this random adaptation meant that every participant heard a different set of rhythms (although some rhythms may well have been heard by more than one participant). Thirdly, each rhythm had a period of $117\,\text{ms/pulse} \times 23\,\text{pulses} = 2{,}691\,\text{ms}$ and was played 5 times through (a total of $13{,}455\,\text{ms}$) prior to the participant giving their responses.

As detailed in Section 3.1, a continuous recognition paradigm [46] was deployed in all three experiments. Closely following previous melody recognition experiments [42, 47, 48, 49, 50, 51, 52], in each trial participants listened to a rhythm and were asked to indicate whether or not they think they have heard this rhythm in this experiment before by clicking on one of the two words "Old" or "New". After that, participants were prompted to indicate their liking of rhythm on a horizontal 7-point Likert-item labelled "Strongly Dislike" on the left, "Neutral" in the middle, and "Strongly Like" on the right. The rhythms as well as their total number differed between the three experiments. The order of the rhythms within each experiment as well as the number of intervening items until a target rhythms was presented for the second time was randomized between participants.

4.2.4 Statistical Approach

To analyse the relationships between the predictors and the outcomes, we used multilevel Bayesian regression. As detailed below, this provides a number of advantages over a classical statistical approach. In the next three sections, we summarize Bayesian regression, our choices of priors, and our method for comparing models. The R package brms [64, 65] – a high-level interface to Stan [66] – was used to perform the regressions. All data and models can be downloaded from https://osf.io/d3vjf/.

Multilevel Bayesian Regression   A multilevel (also known as hierarchical) Bayesian regression model is the Bayesian analog of a classical mixed effects model: it contains *population-level effects* and *group-level effects* which are analogous, respectively, to the definitions of *fixed effects* and *random effects* commonly used in classical statistical packages in R, such as lme4 (there are numerous alternative and contradictory definitions for fixed and random effects [67], which motivates our preference for the former pair of terms). In our context, we have two such "groups": "subjects" (our participants) and "items" (the rhythms), hence the population-level effects are the estimated means of the effects across all participants and rhythms tested; the group-level effects (standard deviations and correlations) estimate the amount these effects co-vary from the mean across participants, and also across rhythms (beyond that accounted for by the model's fixed effects for each rhythm, which are *balance*, *evenness*, *entropy*, and their interactions).

All of our models use a *maximal group-level (random) effects structure* [58]; that is, we allow all within-subject variables to vary by subject (here, participant), and all within-item variables to vary by item (our "items" being rhythms). Maximal random effects structures such as these often fail to converge with classical statistical approaches but can be successfully estimated using the Monte Carlo sampling methods used for Bayesian regression. Another important advantage of Bayesian regression is that, given the observed data and a prior distribution (see the next section for a discussion of priors), it calculates the whole *posterior probability distribution* of each effect rather

than only a point-estimate of the most probable effect of each predictor. This allows for *credibility intervals* to be calculated; unlike the confidence intervals in classical regression, credibility intervals have a straightforward and intuitive meaning: given the model and the data, the 95% credibility interval of an effect is the interval we can be 95% certain contains the effect's true value. It also allows *evidence ratios* to be calculated; these are probability ratios (odds) in favour of directional hypotheses (such as a given effect being greater than zero). For example, if the integral of the posterior distribution over the interval $(0, \infty)$ is p, the evidence ratio in favour of the effect being greater than 0 is $\frac{p}{1-p}$; so, if the lower boundary of a (one-sided) 95% credibility interval is precisely zero, this implies there is a 5% probability the effect is less than zero and a 95% probability it is greater than zero, hence the evidence ratio is $.95/.05 = 19$ – strong evidence in favour of a positive effect.

Due to our use of a Bayesian approach, we do not report classical *p*-values; instead, we focus on effect sizes and evidence ratios. However, if required, a 95% credibility interval can serve a role analogous to the classical *significant/non-significant* binary classification. If the (two-sided) 95% credibility interval does not contain zero this means that less than 2.5% of the posterior distribution is located on the other side of zero, and the effect could be considered "significant" at an $\alpha = .05$ level.

Priors The priors used in all of our models are what are termed *weakly informative priors* [68]. These reflect a minimal amount of knowledge that we possess about the effects' sizes, as now detailed. All of our non-binary predictors are standardized (they have a standard deviation of 1; this immediately implies that we would be unlikely to see effect sizes with very high magnitudes. For example, although we would not be surprised to see an effect size of, say, 0.1 or -1.2 it would be surprising to see an effect size of 10 or -10, while an effect size of 1,000,000 or $-1,000,000$ would be so unlikely it would immediately suggest a coding error [69]. A prior can, and should, reflect this prior knowledge.

For all population-level effects, except the intercept, we used a prior with a Student's *t*-distribution with 3 degrees of freedom, a mean of 0, and a scale of 1. Crucially, the zero mean indicates that our prior beliefs weakly favour the null hypothesis of zero effect size and – in comparison to using a flat uninformative prior – regularizes the estimations [70], thereby reducing overfitting (that said, in the presence of a large number of observations – as in our data, which comprises a total of 24,992 observations – posterior distributions are relatively insensitive to the choice of prior). The prior on the intercept is the brms package's default, which is a Student's *t*-distribution with 3 degrees of freedom, mean of 0, and a scale of 10; the latter reflecting that, in standardized models, intercepts can take on larger sizes than the effects.

The group-level effects (which are standard deviations) are given the default brms prior, which is a half-Student's *t*-distribution with 3 degrees of freedom, a mean of 0, and a scale of 10 (hence reflecting weak prior support for zero standard deviations); the correlation matrix for the group-level effects has an lkj(1) prior which corresponds to a uniform distribution for correlations and standard deviations (respectively, off-diagonals and diagonals in the multivariate normal distribution's covariance matrix).

Model Comparisons with PSIS-LOO We performed a limited number of model comparisons in order to test whether squared terms for evenness, balance, entropy, or their interactions should be included, and also to test whether effects differed between experiments (these comparisons are outlined in Sec. 4.3 and detailed in Supplementary Materials, Sec. 1). The method we used for comparing models is approximate leave-one-out cross-validation achieved with Pareto smoothed importance sampling (PSIS-LOO) [71]. PSIS-LOO provides a useful estimate of the ability of a model to generalize to out-of-sample data – if the model is (a) excessively flexible and is fitting the noise in the data rather than the signal, or (b) it is too simple and not adequately capturing the signal in the data, it will have a poorer fit to out-of-sample data. In the brms package, PSIS-LOO is quantified by the expected log pointwise predictive density (ELPD), which is higher for better models. An ELPD difference of Δ means that the likelihood ratio of the two models, given the priors and out-of-sample data, is e^{Δ}; hence, given an ELPD difference of 3, the model with the higher ELPD is approximately 20 times more likely than the other model (given the data and priors); given an ELPD difference of 10, the model with the higher ELPD is approximately 22,026 times more likely.

McKelvey-Zavoina Bayesian R-squared   For binary or ordinal responses, a model's overall fit to the data can be usefully assessed with the McKelvey and Zavoina pseudo-R^2 [72], here denoted R^2_{mz}, which closely approximates the traditional R^2 that would have been obtained if the latent continuous variable underlying the discrete responses had have been directly observed and fitted with a standard linear model [73, 74, 75]. For a probit link function, and in a Bayesian context, $R^{2(s)}_{\mathrm{mz}} = \mathrm{Var}^{(s)}_{\mu} / (\mathrm{Var}^{(s)}_{\mu} + 1)$, where $\mathrm{Var}^{(s)}_{\mu}$ is the variance of the fitted data using the sth draw of the posterior distribution of the linear part of the model (i.e., $\mathbf{X}\boldsymbol{\beta}$). The total set of draws of $R^{2(s)}_{\mathrm{mz}}$ can be conveniently summarized by their mean, denoted R^2_{mz} and its 2.5% and 97.5% percentiles.

4.3   Results

The models reported here use the same set of predictors, and their interactions, for both the memory and the liking data. The only difference being that the former is a Bernoulli (logistic) regression, the latter is a cumulative normal (ordinal) regression. The respective model formulæ, in the standard notation used in R and some other statistical packages, are provided in Tables 1 and 6.

This set of predictors and their interactions was selected using PSIS-LOO cross-validation within four sets of model comparisons. It was feasible to perform the first set of comparisons only for the memory models because the liking models take too long to fit. These comparisons are detailed in Supplementary Materials, Section 1, and are summarized below. In all cases, the predictors highlighted below were interacted with *IsOld* (whether the rhythm is old or new), and also included the interaction between *Trial* (each participant's progress through the experiment) and *IsOld*. They were all fitted to the data combined from the three experiments, and had maximal group-level structures (as detailed in Sec. 4.2.4).

1. Due to the high correlation between balance and evenness, we compared models with and without interactions between them and models with and without squared *Balance* and/or *Evenness* terms. The model including an interaction between *Balance* and *Evenness* and with no squared terms probably performed best at out-of-sample predictions. Having said that, the outcomes were not decisive enough to be sure whether the interaction between *Balance* and *Evenness* may, in actuality, be a (partial) proxy for some combination of quadratic nonlinear responses to balance and evenness. This uncertainty is unsurprising given the high correlation between balance, evenness, and their squares.

2. Entropy is often considered to exert a nonlinear inverted "U" effect on responses so we compared a model with a squared *Entropy* term to a model without. For the memory data, the two models performed almost identically and, given that the squared term's estimated effect was minuscule, we chose the simpler model. For the liking data, the simpler model without the squared term performed better.

3. To check that the model resulting from the above two sets of comparisons did not overfit the data, we compared it with a model having no interactions between balance, evenness and entropy, and a model with balance, evenness, and entropy completely removed. For both the memory data and the liking data, the first model performed substantially better at out-of-sample prediction. This confirms the models do not overfit.

4. A final comparison explored whether effects vary between the three experiments. For both the *SayOld* and *Liking* data, models with a 3-level categorical factor for the three experiments – this factor interacting with all other predictors – were found to predict out-of-sample data worse than models without this factor. Importantly, this confirms that the effects of *Balance*, *Evenness*, *Entropy*, *IsOld*, and *Trial*, are broadly consistent across the three experiments, which is why we do not include this experiment factor in the models reported in this section. The overall consistency of effects can also be visually confirmed by comparing the conditional effects plots for *SayOld* in each of the three experiments shown in Figures 9, 16, and 23; and comparing the conditional effects plots for *Liking* in each of the three experiments shown in Figures 12, 19, and 26; all of which are in Supplementary Materials, Sections 3–5.

In psychological experiments, there can be autocorrelations among successive responses (i.e., each response may be similar to the previous response). Such autocorrelations can be modelled by including a lagged dependent variable as a predictor, but doing so complicates the interpretation of

the all other predictors' effects and can create bias in random effects models [76]. For every model reported in this article, we ran an alternative that also included a lagged dependent variable; in every case, all other population-level effects and their credibility intervals were almost identical (standardized effects differing by no more than 0.01). For simplicity, therefore, we report only models without a lagged dependent variable.

4.3.1 *SayOld* Model and Results

Table 1 summarizes the posterior distributions of the memory model's coefficients; it also quantifies the evidence that each such effect is greater than or less than zero (a complete table of the model's coefficients, including the group-level effects, is in Supplementary Materials, Sec. 6). Subsequently, Table 5 summarizes the model in terms of bias and discriminability using Equation (3). The fitted model has a Bayesian McKelvey-Zavoina R-squared of .33 with a 95% credibility interval of .32 to .35; in other words, it explains about one third of the variance in the unobserved continuous tendency to say "old" that underlies the observed binary ("old" or "new") responses.

For the hypotheses that balance, evenness, and entropy are relevant to memory, we obtain very strong support from the evidence ratios for the effects of balance, evenness, entropy, and the interaction between balance and evenness (all of which are greater than 30), and strong support from the evidence ratio for the interaction between balance and entropy (greater than 10).

```
Family: bernoulli
Links: mu = probit_approx
Formula: SayOld ~ (Balance * Evenness * Entropy + Trial) * IsOld
          + ((Balance * Evenness * Entropy + Trial) * IsOld | ID)
          + (Trial * IsOld | RhythmID)
Data: BalEveAllData (Number of observations: 24992)
Samples: 4 chains, each with iter = 3500; warmup = 1000; thin = 1;
         total post-warmup samples = 10000
```

Hypothesis	Mean	SE	95% CI		Evid.Ratio	Post.Prob	Star
(Intercept) > 0	0.42	0.05	0.32	0.51	> 9999.00	1.00	*
(Balance) > 0	0.29	0.07	0.16	0.41	9999.00	1.00	*
(Evenness) < 0	−0.54	0.08	−0.67	−0.40	> 9999.00	1.00	*
(Entropy) > 0	0.13	0.03	0.08	0.18	> 9999.00	1.00	*
(Balance:Evenness) < 0	−0.15	0.03	−0.20	−0.11	> 9999.00	1.00	*
(Balance:Entropy) > 0	0.11	0.08	−0.02	0.23	10.59	0.91	
(Evenness:Entropy) < 0	−0.09	0.08	−0.22	0.04	6.35	0.86	
(Balance:Evenness:Entropy) < 0	−0.02	0.02	−0.06	0.03	2.81	0.74	
(Trial) > 0	0.36	0.03	0.32	0.40	> 9999.00	1.00	*
(IsOld) > 0	0.07	0.04	0.01	0.14	34.71	0.97	*
(Balance:IsOld) > 0	0.04	0.10	−0.12	0.20	1.90	0.66	
(Evenness:IsOld) < 0	−0.04	0.11	−0.21	0.14	1.78	0.64	
(Entropy:IsOld) < 0	−0.02	0.04	−0.08	0.05	1.80	0.64	
(Balance:Evenness:IsOld) > 0	0.02	0.04	−0.04	0.08	2.18	0.69	
(Balance:Entropy:IsOld) > 0	0.07	0.10	−0.09	0.24	3.35	0.77	
(Evenness:Entropy:IsOld) < 0	−0.10	0.11	−0.28	0.08	4.71	0.82	
(Balance:Evenness:Entropy:IsOld) < 0	−0.02	0.03	−0.08	0.03	3.01	0.75	
(Trial:IsOld) < 0	−0.16	0.03	−0.21	−0.12	> 9999.00	1.00	*

Table 1: Summary of directional hypothesis tests for the effects in the *SayOld* model. The columns from left to right are: the hypothesis being tested; the estimated mean of the standardized effect; its standard error; the 95% credibility interval; the evidence ratio in favour of the hypothesis; the posterior probability for the hypothesis; a star, if 0 lies outside the 95% CI.

Figure 7 visualizes how interactions of balance, evenness, and entropy affect the probability of saying "old" for both new rhythms and old rhythms. In the following subsections, these effects are verbally summarized by showing how each feature (balance, evenness, or entropy), as conditioned on the other two features, influences saying "old"; we also summarize the mutual effects of *IsOld* and *Trial*. More detailed analyses are provided in the Supplementary materials, Sec. 2.

Balance. Typically, balance moderately increases "old" responses. This increase in "old" responses is diminished with increasing evenness, and enhanced with increasing entropy. Table 2 gives an overview of the direction in which balance affects "old" responses, as conditioned by the
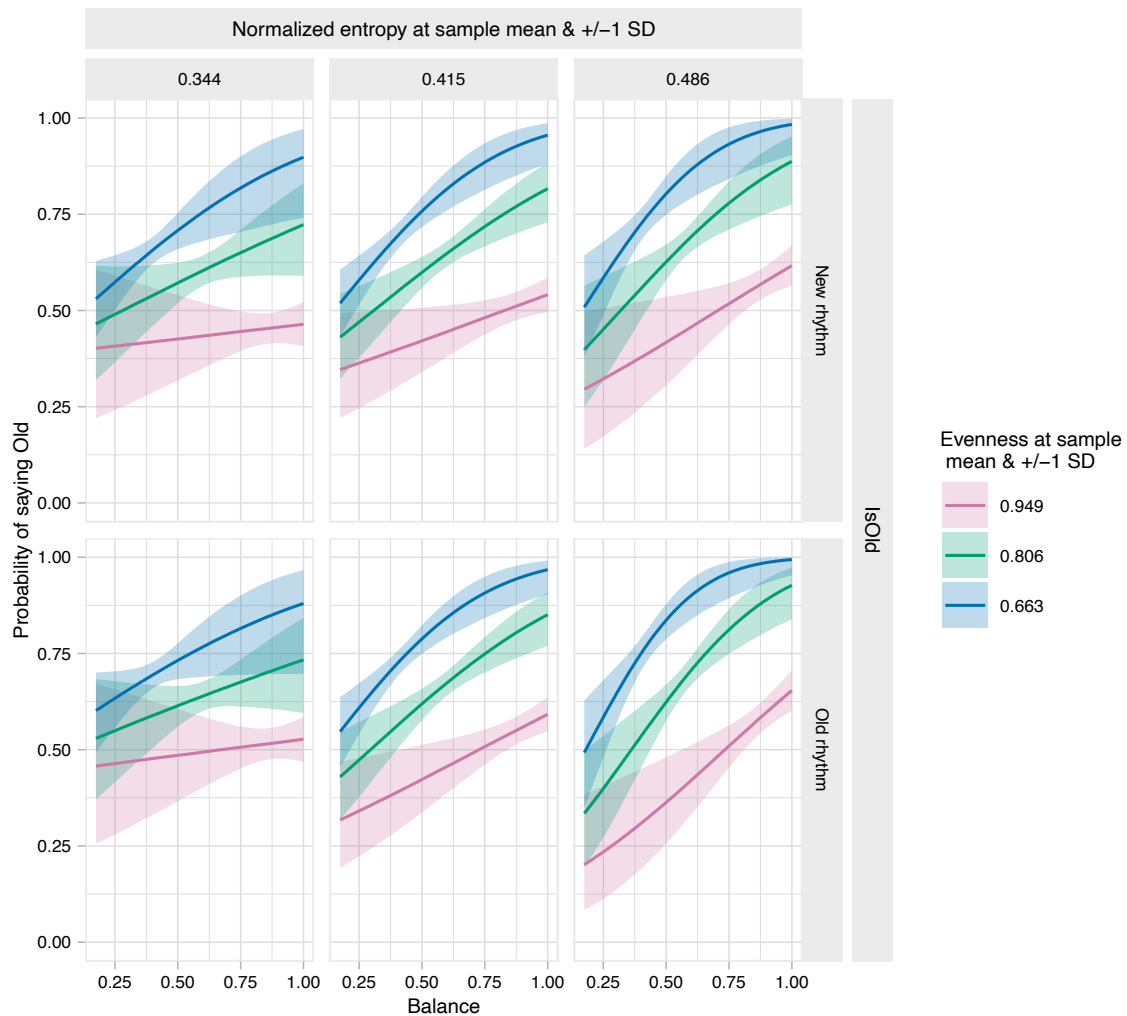
Figure 7: Conditional effects for *Balance*, *Evenness*, *Entropy*, and *IsOld*, on *SayOld* for all tested rhythms. All other predictors at their means. *SayOld* here refers to the expected probability of saying "old", given the predictors' values and the posterior distribution of their effects. The ribbons show the posterior distributions' 95% credibility intervals.

values of evenness and entropy (the magnitudes of these effects, and the effects discussed subsequently, can be assessed from Table 1 or Figure 7).

	$Ent = -1$	$Ent = 0$	$Ent = 1$
$Eve = -1$	+	+	+
$Eve = 0$	+	+	+
$Eve = 1$	0	+	+

Table 2: Effect directions of *balance* conditioned across different values of *evenness* and *entropy* (their means and 1 standard deviation below and above).

Evenness. Generally, evenness strongly decreases "old" responses. This effect tends to increase with increasing balance. Table 3 gives an overview of the direction in which evenness affects "old" responses, as conditioned by the values of balance and entropy.

	$Ent = -1$	$Ent = 0$	$Ent = 1$
$Bal = -1$	–	–	–
$Bal = 0$	–	–	–
$Bal = 1$	–	–	–

Table 3: Effect directions of *evenness* conditioned across different values of *balance* and *entropy* (their means and 1 standard deviation below and above).

Entropy. Typically, entropy weakly increases "old" responses. This effect is even more pronounced when balance is high and when evenness is low. Table 4 gives an overview of the direction in which entropy affects "old" responses, as conditioned by the values of balance and evenness.

	$Eve = -1$	$Eve = 0$	$Eve = 1$
$Bal = -1$	+	0	0
$Bal = 0$	+	+	0
$Bal = 1$	+	+	+

Table 4: Effect directions of *entropy* conditioned across different values of *balance* and *evenness* (their means and 1 standard deviation below and above).

IsOld and Trial. At the start of the of the experiment, participants are capable of recognizing the rhythms but this ability decreases over time. Generally, the probability of saying "old" increases over time, with a greater increase for new rhythms.

Something important to note about Figure 7 is that many of the values it depicts are physically/mathematically infeasible. As shown in Figure 6, balance and evenness are highly correlated and so it is simply impossible to have a rhythm with, for example, very low evenness and very high balance. The regions in Figure 7 where real-world rhythms actually live are indicated by the areas where the 95% credibility intervals (the ribbons) are narrower – they are narrower in these locations because there were more rhythms there to be fitted by the model. Figure 8 illustrates the model's estimated effects in a way that avoids this issue – it shows the predicted *SayOld* values for all rhythms actually tested in the experiment ("three-dimensional" versions of these plots are provided in Supplementary Materials, Sec. 8, Figure 31).
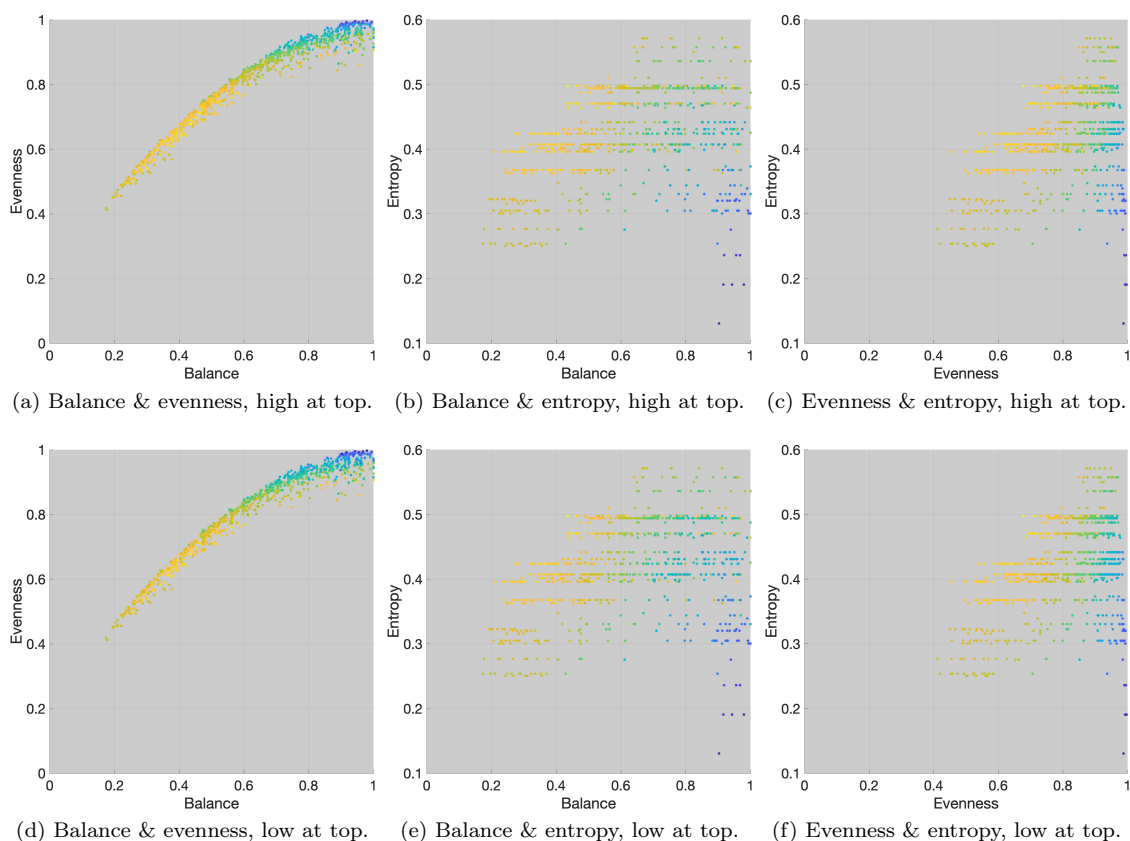
Figure 8: Scatter plots showing the fitted values of *SayOld* (yellow is high, green is intermediate, blue is low) across different pairs of balance, evenness, and entropy values in the 1,252 different rhythms tested in all three experiments. All other covariates were set to their mean, so these simulations are averaged over old and new rhythms. In the top row, rhythms with higher fitted values of *SayOld* are plotted on top of rhythms with lower fitted values of *SayOld*; hence these plots facilitate seeing the rhythms most likely to induce an "old" response. In the bottom row, rhythms with lower fitted values of *SayOld* are plotted on top of rhythms with higher fitted values of *SayOld*, which facilitates seeing the rhythms least likely to induce an "old" response.

Memory bias and discriminability. As detailed in Section 3.1, the coefficients of the model – discussed in detail in the previous subsection – can be used to calculate discriminability and bias within the signal detection theory framework. Table 5 provides estimates and hypothesis tests for the effects of *balance*, *evenness*, and *entropy* on *discriminability* (*disc*) and *bias*. Essentially, Table 5 is just a different way of summarizing the information in Table 1; we provide both to demonstrate how the modelling approach deployed here can be used to estimate, directly from the model's coefficients, important measures within the SDT framework. In summary, there is strong evidence (mostly very strong) that all predictors and their interactions affect memory bias. In terms of discriminability, we see strong evidence that participants are able to discriminate old from new items. However, as the experiment progresses, discrimination performance decreases. While we obtain evidence that balance, evenness, and entropy affect bias, we do not obtain convincing evidence that they affect discriminability.

```
Family: bernoulli
Links: mu = probit_approx
Formula: SayOld ~ (Balance * Evenness * Entropy + Trial) * IsOld
              + ((Balance * Evenness * Entropy + Trial) * IsOld | ID)
              + (Trial * IsOld | RhythmID)
Data: BalEveAllData (Number of observations: 24992)
Samples: 4 chains, each with iter = 3500; warmup = 1000; thin = 1;
         total post-warmup samples = 10000
```

Hypothesis	Mean	SE	95% CI		Evid.Ratio	Post.Prob	Star
bias < 0	−0.46	0.04	−0.53	−0.38	> 9999.00	1.00	*
bias(Balance) < 0	−0.31	0.06	−0.40	−0.21	> 9999.00	1.00	*
bias(Evenness) > 0	0.56	0.06	0.45	0.67	> 9999.00	1.00	*
bias(Entropy) < 0	−0.13	0.02	−0.17	−0.09	> 9999.00	1.00	*
bias(Balance:Evenness) > 0	0.15	0.02	0.11	0.18	> 9999.00	1.00	*
bias(Balance:Entropy) < 0	−0.14	0.06	−0.24	−0.04	118.05	0.99	*
bias(Evenness:Entropy) > 0	0.14	0.06	0.03	0.25	74.76	0.99	*
bias(Balance:Evenness:Entropy) > 0	0.03	0.02	−0.01	0.06	10.40	0.91	
bias(Trial) < 0	−0.28	0.02	−0.32	−0.24	> 9999.00	1.00	*
disc > 0	0.07	0.04	0.01	0.14	34.71	0.97	*
disc(Balance) > 0	0.04	0.10	−0.12	0.20	1.90	0.66	
disc(Evenness) < 0	−0.04	0.11	−0.21	0.14	1.78	0.64	
disc(Entropy) < 0	−0.02	0.04	−0.08	0.05	1.80	0.64	
disc(Balance:Evenness) > 0	0.02	0.04	−0.04	0.08	2.18	0.69	
disc(Balance:Entropy) > 0	0.07	0.10	−0.09	0.24	3.35	0.77	
disc(Evenness:Entropy) < 0	−0.10	0.11	−0.28	0.08	4.71	0.82	
disc(Balance:Evenness:Entropy) < 0	−0.02	0.03	−0.08	0.03	3.01	0.75	
disc(Trial) < 0	−0.16	0.03	−0.21	−0.12	> 9999.00	1.00	*

Table 5: Summary of directional hypothesis tests for the effects of the *SayOld* model's predictors on memory *bias* and *discriminability* (here denoted "disc"). These are calculated from linear combinations of the model's raw coefficients, as detailed in Section 3.1. For this reason, the estimated mean bias or discriminability for any set of predictor values is obtained by summing the estimated mean values shown in the second column, each multiplied by the value of their respective predictor(s); e.g., if all predictors are at zero (their means) except for *Balance* which is −0.5, and *Evenness* which is 0.25, the predicted mean bias would be $-0.46+(-0.31\times-0.5)+(0.56\times0.25)+(0.15\times(-0.5\times0.25)) \approx -0.18$. (Note that a negative bias means a greater tendency to say "old", while a positive bias means a lesser tendency to say "old", regardless of anything else; as explained in Sec. 3.1). The columns from left to right are: the hypothesis being tested; the estimated mean of the effect; its standard error; the 95% credibility interval; the evidence ratio in favour of the hypothesis; the posterior probability for the hypothesis; a star, if 0 lies outside the 95% CI.

4.3.2 *Liking* Model and Results

Table 6 summarizes the posterior distributions of this model's coefficients; it also quantifies the evidence that each such effect is greater than or less than zero. (A complete table of the model's coefficients, including the group-level effects, is in Supplementary Materials, Sec. 6.) The fitted model has a Bayesian McKelvey-Zavoina R-squared of .52 with a 95% credibility interval of .51 to .54 so it explains about half the variance of the continuous, but unobserved, "liking" that underlies the observed ordinal ratings.

For the hypotheses that balance, evenness, and entropy are relevant to æsthetic evaluation, we obtain very strong support (evidence ratios greater than 30) for evenness, the interaction between evenness and balance, and the interaction between balance and entropy, and strong support (evidence ratios greater than 10 but less than 30) for the interactions between evenness and entropy, between balance, evenness, and entropy, between balance and *IsOld*, and between evenness and *IsOld*.

Figure 9 visualizes how interactions of balance, evenness, and entropy affect the underlying liking of both new rhythms and old rhythms. In the following subsections, these effects are verbally summarized by showing how each feature (balance, evenness, or entropy), as conditioned on the other two features, influences the likng of rhythms; we also summarize the mutual effects of *IsOld* and *Trial*, which are respectively indicative of mere exposure and boredom. More detailed analyses are provided in the Supplementary materials, Sec. 2.

Balance For new rhythms, balance moderately increases liking for rhythms with low to medium evenness and medium to high entropy. For old rhythms, balance generally has a more positive effect. Tables 7 and 8 give an overview of the direction in which balance affects liking – firstly for new rhythms, then for old – as conditioned by the values of evenness and entropy (the magnitudes

Family: cumulative
Links: mu = probit_approx; disc = identity
Formula: Liking ~ (Balance * Evenness * Entropy + Trial) * IsOld
                  + ((Balance * Evenness * Entropy + Trial) * IsOld | ID)
                  + (Trial * IsOld | RhythmID)
Data: BalEveAllData (Number of observations: 24992)
Samples: 4 chains, each with iter = 3500; warmup = 1000; thin = 1;
        total post-warmup samples = 10000

Hypothesis	Mean	SE	95% CI		Evid.Ratio	Post.Prob	Star
(Intercept[1]) < 0	−2.54	0.08	−2.67	−2.40	> 9999.00	1.00	*
(Intercept[2]) < 0	−1.75	0.08	−1.88	−1.61	> 9999.00	1.00	*
(Intercept[3]) < 0	−0.92	0.08	−1.05	−0.79	> 9999.00	1.00	*
(Intercept[4]) > 0	0.32	0.08	0.18	0.45	> 9999.00	1.00	*
(Intercept[5]) > 0	1.38	0.08	1.25	1.51	> 9999.00	1.00	*
(Intercept[6]) > 0	2.35	0.08	2.21	2.48	> 9999.00	1.00	*
(Balance) > 0	0.04	0.07	−0.08	0.15	2.33	0.70	
(Evenness) < 0	−0.23	0.08	−0.36	−0.10	369.37	1.00	*
(Entropy) > 0	0.00	0.03	−0.05	0.04	1.35	0.57	
(Balance:Evenness) < 0	−0.11	0.03	−0.16	−0.07	> 9999.00	1.00	*
(Balance:Entropy) > 0	0.14	0.07	0.02	0.25	35.63	0.97	*
(Evenness:Entropy) < 0	−0.12	0.07	−0.24	0.01	16.36	0.94	
(Balance:Evenness:Entropy) < 0	−0.03	0.02	−0.07	0.01	10.05	0.91	
(Trial) < 0	−0.13	0.03	−0.18	−0.09	> 9999.00	1.00	*
(IsOld) > 0	0.06	0.03	0.01	0.11	40.67	0.98	*
(Balance:IsOld) > 0	0.14	0.08	0.01	0.27	24.97	0.96	*
(Evenness:IsOld) < 0	−0.13	0.09	−0.27	0.01	14.22	0.93	
(Entropy:IsOld) > 0	0.01	0.03	−0.04	0.06	1.81	0.64	
(Balance:Evenness:IsOld) < 0	0.00	0.03	−0.05	0.05	1.24	0.55	
(Balance:Entropy:IsOld) > 0	0.01	0.08	−0.12	0.13	1.15	0.53	
(Evenness:Entropy:IsOld) < 0	−0.03	0.08	−0.17	0.11	1.87	0.65	
(Balance:Evenness:Entropy:IsOld) > 0	0.00	0.02	−0.04	0.04	1.19	0.54	
(Trial:IsOld) > 0	0.07	0.02	0.03	0.11	624.00	1.00	*

Table 6: Summary of directional hypothesis tests for the effects in the *Liking* model. This model has six intercepts, which are the cut points on the latent scale that determine into which of the seven Likert ratings a response is categorized (as detailed in Sec. 3.2). The columns from left to right are: the hypothesis being tested; the estimated mean of the standardized effect; its standard error; the 95% credibility interval; the evidence ratio in favour of the hypothesis; the posterior probability for the hypothesis; a star, if 0 lies outside the 95% CI.

of these effects, and the effects discussed subsequently, can be assessed from Table 6 or Figure 9).

	$Ent = -1$	$Ent = 0$	$Ent = 1$
$Eve = -1$	0	+	+
$Eve = 0$	−	0	+
$Eve = 1$	−	−	0

Table 7: For new rhythms, the effect directions of *balance* conditioned across different values of *evenness* and *entropy* (their means and 1 standard deviation below and above).

	$Ent = -1$	$Ent = 0$	$Ent = 1$
$Eve = -1$	+	+	+
$Eve = 0$	0	+	+
$Eve = 1$	−	0	+

Table 8: For old rhythms, the effect directions of *balance* conditioned across different values of *evenness* and *entropy* (their means and 1 standard deviation below and above).

Evenness. Typically, evenness moderately decreases liking for rhythms with medium to high balance and medium to high entropy. Table 9 gives an overview of the direction in which balance affects liking, as conditioned by the values of balance and entropy.
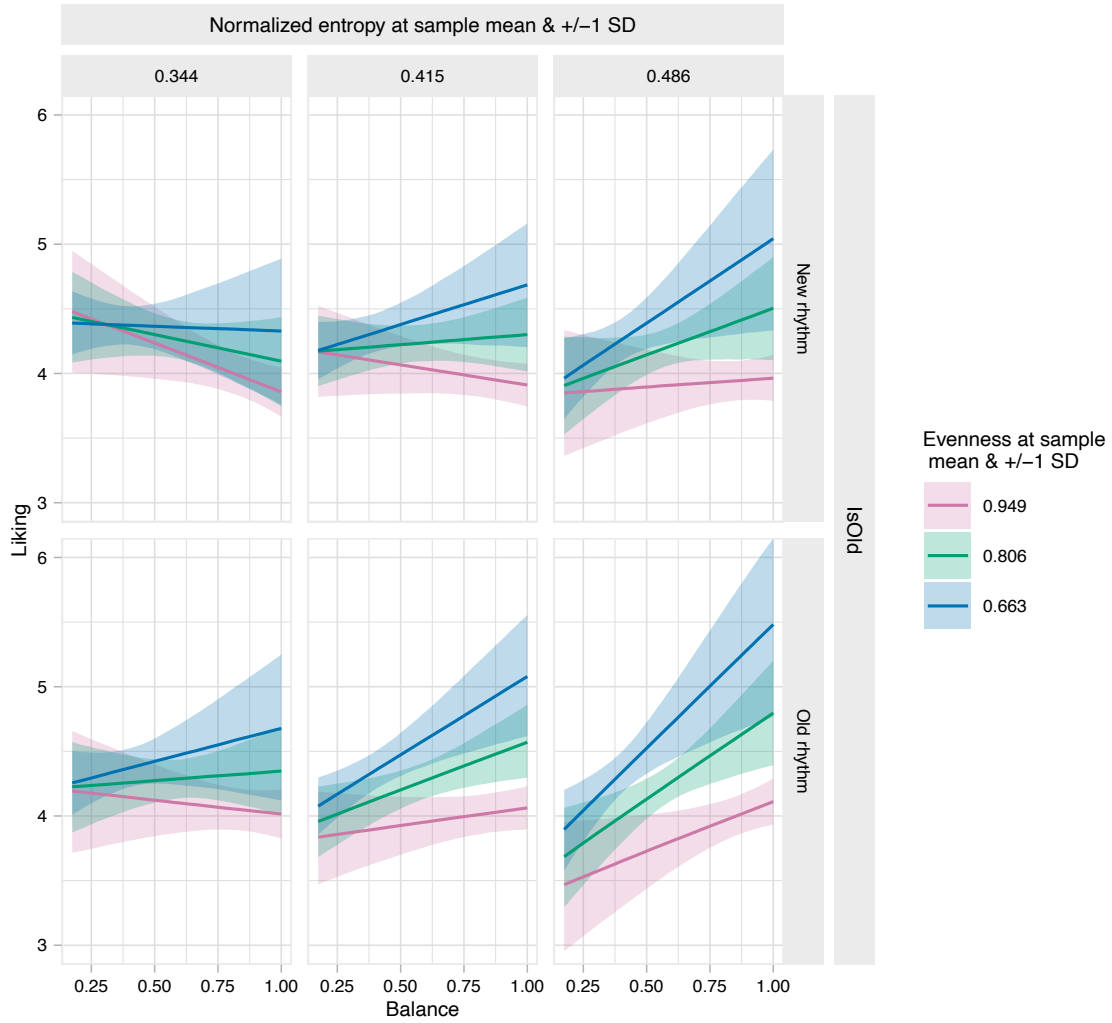
Figure 9: Conditional effects for *Balance*, *Evenness*, *Entropy*, and *IsOld*, on *Liking* for all tested rhythms. All other predictors at their means. *Liking* here refers to the expected predicted ordinal response, given the predictors' values and the posterior distribution of their effects. The ribbons show the posterior distributions' 95% credibility intervals.

	$Ent = -1$	$Ent = 0$	$Ent = 1$
$Bal = -1$	0	−	−
$Bal = 0$	0	−	−
$Bal = 1$	−	−	−

Table 9: Effect directions of *evenness* conditioned across different values of *balance* and *entropy* (their means and 1 standard deviation below and above).

Entropy. Generally, entropy moderately increases liking for rhythms with high balance and low evenness, but moderately decreases liking for rhythms with low balance or with high evenness. Table 10 gives an overview of the direction in which balance affects liking, as conditioned by the values of balance and evenness.

	$Eve = -1$	$Eve = 0$	$Eve = 1$
$Bal = -1$	0	−	−
$Bal = 0$	+	0	−
$Bal = 1$	+	+	0

Table 10: Effect directions of *entropy* conditioned across different values of *balance* and *evenness* (their means and 1 standard deviation below and above).

IsOld and Trial. There is a very small mere exposure effect, which is slightly enhanced for rhythms with high balance and towards the end of the experiment. For new rhythms, there is a slight reduction in liking as the experiment progresses (likely, a boredom effect); for old rhythms, the boredom effect is slightly diminished.

A useful summary of the model is provided by Figure 10 which, for all of the tested rhythms, show the fitted values for *Liking* as a function of *Balance*, *Evenness*, and *Entropy*. As noted earlier, this avoids the issue of plots like Figure 9, which contain regions that cannot be occupied by any real-world rhythm. Four views of a "three-dimensional" plot of these relationships are shown in Supplementary Materials, Sec. 8, Figure 32.

(a) Balance & evenness, high at top.	(b) Balance & entropy, high at top.	(c) Evenness & entropy, high at top.

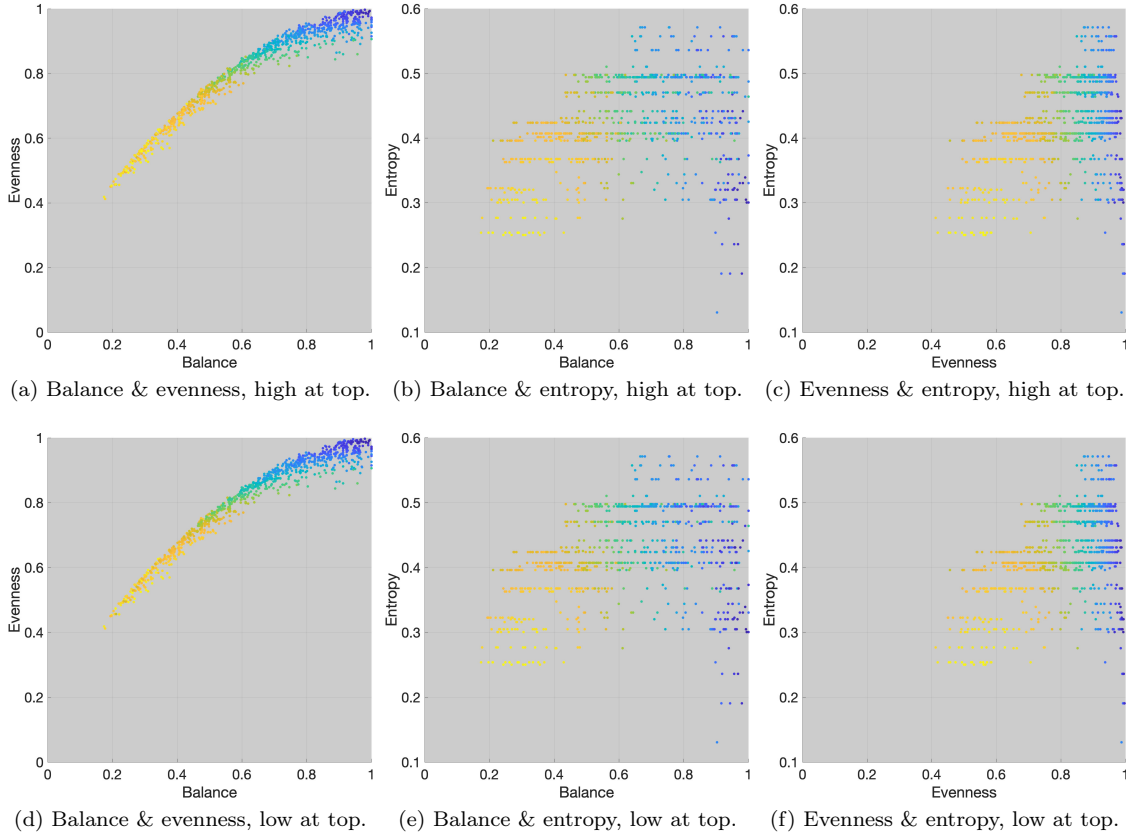| (d) Balance & evenness, low at top.	(e) Balance & entropy, low at top.	(f) Evenness & entropy, low at top.

Figure 10: Scatter plots showing the fitted values of liking (yellow is high, green is intermediate, blue is low) across different pairs of balance, evenness, and entropy values in the 1,252 different rhythms tested in all three experiments. All other covariates were set to their mean, so these simulations are averaged over old and new rhythms. In the top row, rhythms with higher fitted values of liking are plotted on top of rhythms with lower fitted values of liking; hence this plot facilitates seeing where the most liked rhythms occur. In the bottom row, rhythms with lower fitted values of liking are plotted on top of rhythms with higher fitted values of liking, which facilitates seeing where the least liked rhythms occur.

5   Discussion

In this article, we have explored how features characterizing the organization of sound into periodic temporal patterns influence perception. Over the course of three experiments, we found support for the distinct perceptual relevance of balance, evenness, and interonset interval entropy in the context of musical rhythms – they all impact on the memory and liking of musical rhythms. Under the RMR conjecture, perceptually relevant features lead to the formation of distinct and coherent percepts, hence a measure of memory gives insight into the link between perception and memory, whereas liking responses provide a window to an evaluation process that necessitates an interaction between perception and prior knowledge [41].

Here, models for memory and æsthetic evaluation, using the data from all 177 participants and 1,252 rhythms in three experiments, explained a meaningful proportion of the variance in

responses (for the *SayOld* model, $R^2_{\text{mz}} = .33$; for the *Liking* model, $R^2_{\text{mz}} = .52$). Furthermore, both models showed clear support for the perceptual relevance and distinctiveness of the three rhythmic features at the focus of this investigation: balance, evenness, and interonset interval entropy.

The memory model showed that participants were capable of recognizing previously presented rhythms above chance, and that balance, evenness, and entropy all affect the way we remember and æsthetically evaluate musical rhythms. These effects are broadly consistent between three different rhythmic periods of 24, 18, and 23 pulses (as demonstrated by the model comparisons at the start of Sec. 4.3). In summary, rhythms with low evenness, high balance, and high IOI entropy typically induce a negative memory bias (a greater tendency to say "old"); as explained in Section 3.1, this suggests that such rhythms are perceived as familiar (outside of the experiment) or that such rhythms are perceptually clustered. The results for æsthetic evaluation are a little more complicated to summarize because of the way the effects interact; but, in short, rhythms with low evenness are generally liked more, and low-evenness rhythms that also have relatively low IOI entropy are usually liked even more.

Most research on musical rhythm perception has focused on (a) how sensorimotor synchronization (typically assessed by tapping experiments [77, 78]) is negatively affected by increasing rhythmic complexity and how tap timings exhibit specific biases such as negative mean asynchrony [79] or attraction towards specific durational ratios [80, 81, 82]; (b) comparing the frequency content of a rhythm's envelope with its associated EEG neural response, thereby providing insight into top-down influences or nonlinearities in the auditory system [83, 84]; (c) fMRI responses to rhythms with differing complexities, which raise the possible importance of the basal ganglia in generating an internal beat (see [85] for an overview); (d) listeners' ratings of rhythmic "groove" (the urge to move in response to music) [86, 87, 25, 27, 88]. Although groove is strongly associated with pleasure [86], it is likely not the only attribute of rhythm that influences likeability or memorability. We are not aware of any research prior to this article that experimentally investigates either the memorability or the overall likeability of a large range of rhythms; nor any studies that directly explore the impact of balance, evenness, or interonset interval entropy on rhythm perception. Given our results, we feel that the approach demonstrated here – both in terms of the questions asked ("old"/"new" and "liking") and the modelling with DFT coefficients of the Argand vector – may prove useful for future investigations of rhythm perception and performance, and may provide useful insights into cognitive representations of temporal patterns.

We now explore some ramifications of our results in two ways: first (Sec. 5.1), we generalize the fitted models to unseen rhythms; secondly (Sec. 5.2), we posit teleological explanations for why balance, evenness, and entropy have perceptual relevance. The explanation for balance and evenness and, more generally, the discrete Fourier transform is novel and we discuss this in depth in Section 5.2.1. Explanations for the relevance of entropy are well established, and Section 5.2.2 serves as a brief overview and reminder of these.

5.1 Generalization of Fitted Models to Untested Rhythms

Figures 11 and 12 show our models' predicted *SayOld* and *Liking* values for all possible 6- and 7-beat rhythms in all periods of 8 pulses up to 32 pulses. This is a grand total of 319,763 different primitive rhythms up to rotation and reflection. Predictions, such as these – obtained from a generalized linear regression model used to predict data beyond the sample it was fitted to – must be taken with caution, more so when the model contains interactions and the predictors' values are beyond the range of the original data [89]. Despite this, we think it of interest to observe such predictions – they may be useful in demonstrating obvious flaws in the fitted model and, from the musician's perspective, they satisfy the inevitable curiosity about the musical potentials of unfamiliar rhythms, speculative though they may be. Clearly, as more data are observed, these predictions can be improved.

5.2 Why are Balance, Evenness, and Entropy Perceptually Relevant?

5.2.1 Balance, Evenness, and the DFT

As described in Section 2, balance and evenness are, respectively, the zeroth and first coefficients of the discrete Fourier transform (DFT) of a rhythm encoded as a circular pattern (i.e., encoded as an Argand vector, as defined in Sec. 2.1). Generally, the DFT is a one-to-one transformation between two representations of the same signal: often a time-domain representation and a frequency-domain

(a) Balance & evenness, high at top.   (b) Balance & entropy, high at top.   (c) Evenness & entropy, high at top.

(d) Balance & evenness, low at top.   (e) Balance & entropy, low at top.   (f) Evenness & entropy, low at top.
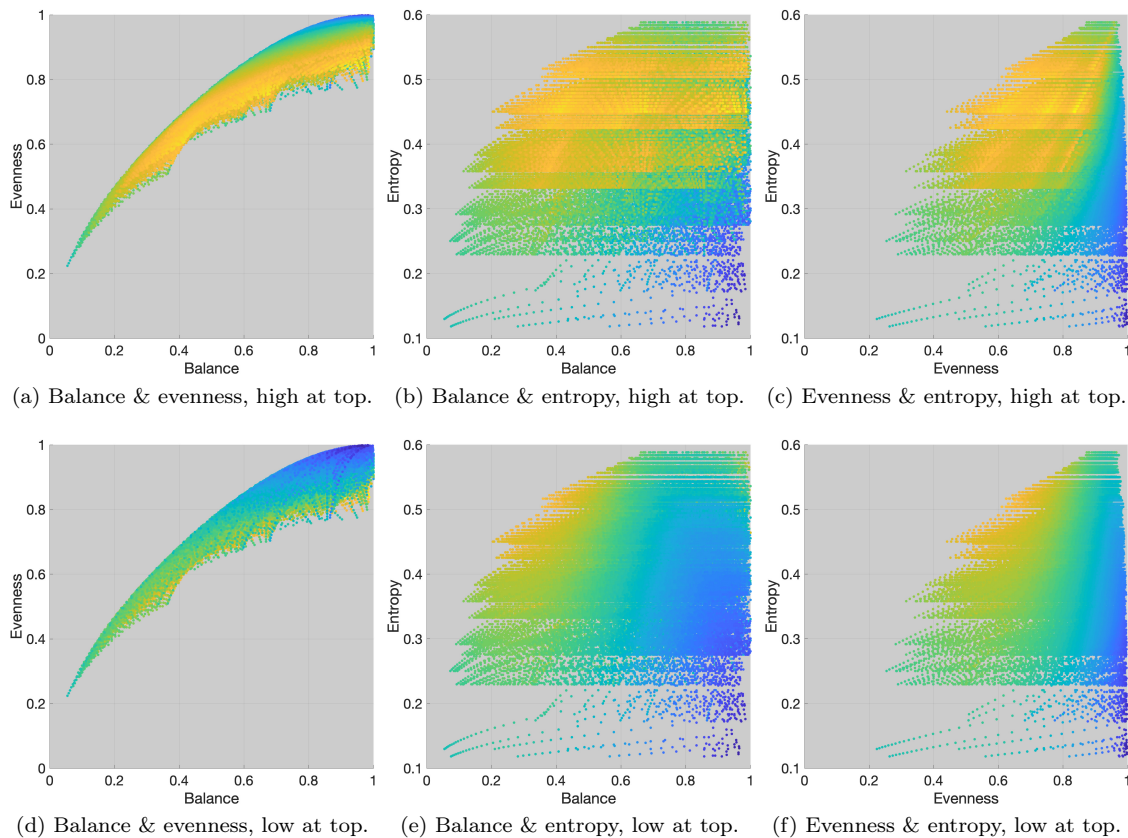
Figure 11: Scatter plots showing the fitted values of *SayOld* (yellow is high, green is intermediate, blue is low) across different pairs of balance, evenness, and entropy values in the 319,763 different rhythms with 6 and 7 beats in 8 to 32 pulses. All other covariates were set to their mean, so these simulations are averaged over old and new rhythms. In the top row, rhythms with higher fitted values of *SayOld* are plotted on top of rhythms with lower fitted values of *SayOld*, which facilitates seeing the rhythms most likely to induce an "old" response. In the bottom row, rhythms with lower fitted values of *SayOld* are plotted on top of rhythms with higher fitted values of *SayOld*, which facilitates seeing the rhythms least likely to induce an "old" response.

representation. Such transformations are of relevance to human perception of the environment. For example, the human cochlea performs a Fourier-like transformation of the incoming sound signal. Indeed, the brain, with its tonotopic structure, even retains a spatial organization that is based on the frequency representation of sound [90]. Furthermore, seminal work by Nozaradan [91] shows that comparison of the DFT of the envelope of a rhythm with the DFT of the amplitude envelope of neural responses can show neural entrainment, which carries perceptual relevance for beat perception.

However, it is important to note that the Argand vector representation of the signal we use here – see Equation (1) – is a different and considerably more parsimonious (sparse) representation of a rhythm than the "traditional" waveform or envelope representations that the cochlear or Nozaradan's DFT, respectively, operate on.

In the traditional case, each of the signal vector's entries x_n is the signal's *amplitude at time* n/f_s (where $n = 0, 1, \ldots, N-1$, the number of digital samples is N, and f_s is the sample rate); hence this is a *time-domain* representation. Each of its DFT vector's entries $(\mathcal{F}\boldsymbol{x})_m$ (where $m = 0, 1, \ldots, N-1$), are the signal's *magnitude and phase at frequency* mf_s/N; hence this is a *frequency-domain* representation. Both representations approach the continuous (analog) case as f_s and N approach infinity.

Both the Argand vector representation and its DFT are quite different to the traditional versions above: the phase of each of the Argand vector's entries z_k is the *time of the kth onset* ($k = 0, 1, \ldots, K-1$); hence this is an *onset-domain* representation. The magnitude of each entry of its DFT $|(\mathcal{F}\boldsymbol{z})_j|$ quantifies the similarity of the rhythm's sequence of K onsets to a sequence of isochronous pulses with interonset interval j/K; hence this is an *IOI-sequence-domain* representation. For $j = 0$, this corresponds to the balance of the rhythm; for $j = 1$, the evenness of the
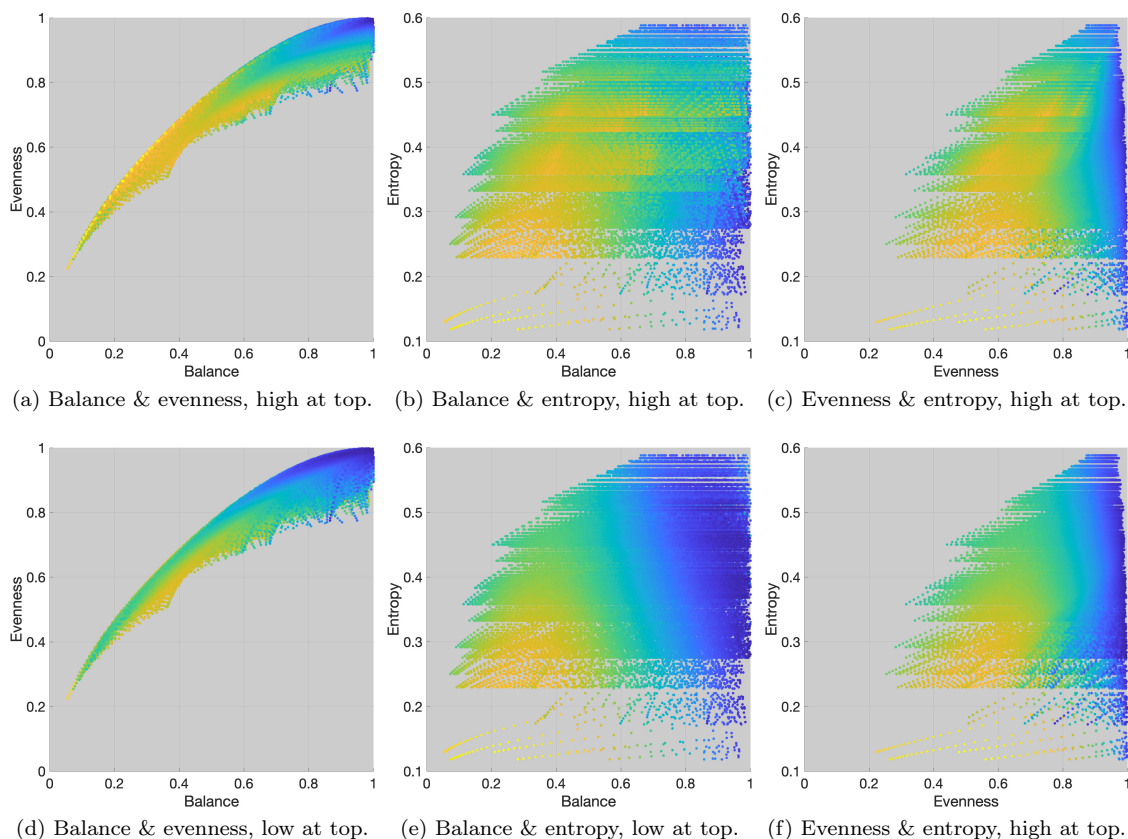
Figure 12: Scatter plots showing the fitted values of *Liking* (yellow is high, green is intermediate, blue is low) across different pairs of balance, evenness, and entropy values in the 319,763 different rhythms with 6 and 7 beats in 8 to 32 pulses. All other covariates were set to their mean, so these simulations are averaged over old and new rhythms. In the top row, rhythms with higher fitted values of *Liking* are plotted on top of rhythms with lower fitted values of *Liking*, which facilitates seeing where the most liked rhythms occur. In the bottom row, rhythms with lower fitted values of *Liking* are plotted on top of rhythms with higher fitted values of *Liking*, which facilitates seeing where the least liked rhythms occur.

rhythm. The higher coefficients are not so easily characterized by physical analogies like "balance" and "evenness". This does not imply, however, that they are not perceptually relevant. They can be explained in the following way: rhythms with maximal magnitudes for the jth coefficient maximize the number of successive interonset intervals of length j/K, and cluster the remaining onsets between the end and start of that IOI sequence. When the rhythm has additional constraints – such as all onsets falling at some equal division of the period, as in our tested rhythms – the above-described patterns cannot occur; now, maximizing the jth coefficient leads to intriguing and clearly structured patterns (some interesting examples are illustrated in Supplementary Materials, Section 7). For a time-domain representation of a rhythm – such as the indicator vector described in Section 2.1 – the first coefficient of the frequency-domain DFT and the zeroth coefficient of the IOI-domain DFT are identical; that is, $(\mathcal{F}\boldsymbol{x})_1 = (\mathcal{F}\boldsymbol{z})_0$. Hence, the magnitudes of both quantify balance. Crucially, however, there is no single coefficient of the frequency domain DFT, $(\mathcal{F}\boldsymbol{x})_m$ or $|(\mathcal{F}\boldsymbol{x})_m|$, that quantifies evenness like the first coefficient of the IOI-domain DFT, $|(\mathcal{F}\boldsymbol{z})_1|$. Taken together, this means that only the DFT of the Argand vector allows calculation of evenness, whereas the DFT of both representations allow measurement of balance. As we are here interested in both evenness and balance, we consistently use the DFT on the Argand representation of rhythms.

We now describe the full DFT of the Argand vector (not just its zeroth and first coefficients), its relevance to rhythm perception, and we will provide a new theory that motivates investigating the magnitudes of specific coefficients of this DFT; notably, balance and evenness. The full discrete

28

Fourier transform of a K-onset rhythm's Argand vector $\boldsymbol{z}$ is

$$(\mathcal{F}\boldsymbol{z})_j = \sum_{k=0}^{K-1} z_k \, \mathrm{e}^{-2\pi \mathrm{i} j k / K}. \tag{4}$$

The resulting DFT, $\mathcal{F}\boldsymbol{z}$, is another K-entry complex vector. Importantly, $\mathcal{F}\boldsymbol{z}$ can be transformed back to the original vector $\boldsymbol{z}$ with the inverse DFT

$$z_k = \frac{1}{K} \sum_{j=0}^{K-1} (\mathcal{F}\boldsymbol{z})_j \, \mathrm{e}^{2\pi \mathrm{i} j k / K}, \tag{5}$$

which simply involves a sign flip in the exponent and scaling by $1/K$ rather than 1 (different scaling conventions can be used so long as their product is $1/K$).

In this way, the DFT is a one-to-one and invertible mapping between complex vectors – crucially, no information is lost switching between the two representations (the onset-domain Argand vector $\boldsymbol{z}$ and the IOI-domain DFT vector $\mathcal{F}\boldsymbol{z}$); the information is just presented differently. Importantly, *each* entry of the DFT vector $\mathcal{F}\boldsymbol{z}$ is derived from *all* entries of the Argand vector $\boldsymbol{z}$, and vice versa. In essence, *each* entry of the DFT is an informative summary of some unique aspect of the *entire* rhythm's onset times.

A feature's robustness to random jitter   We propose here, for the first time, an interesting theory: the magnitudes of the DFT's lowest three coefficients – particularly the first – are more robust to random temporal perturbations (jitter) of the rhythm, hence they can form more stable and, therefore, more relevant percepts. Temporal jitter is a pertinent consideration because it naturally arises in rhythms from inaccuracies of temporal perception, memory, and motor skills that are engaged in both the perception and production of rhythm. From a perceptual standpoint, it makes sense to be able to generalize over such random jitter – to be not distracted by it, and to be able to maintain a stable and coherent percept in spite of these inevitable random variations. In summary, features that show robustness against random jitter are likely candidates for perceptual relevance.

The mathematical proposition that underlies this argument, can be stated more formally:

Proposition 5.1. *Let the Argand vector $\boldsymbol{z}$ be a primitive K-onset rhythm, such that $3 \leq K \leq 16$. When the rhythm's onset times are perturbed by K independent and identical random normal distributions (modulo the rhythm's period), the expected coefficient of variation of the magnitudes of the coefficients of $\mathcal{F}\boldsymbol{z}$ is smallest for its first coefficient $(\mathcal{F}\boldsymbol{z})_1$, second smallest for its zeroth coefficient $(\mathcal{F}\boldsymbol{z})_0$, and third smallest for its second coefficient $(\mathcal{F}\boldsymbol{z})_2$.*

An analytical proof of this proposition eludes us (and, even if there were a formula relating the variance of the DFT components to variance in the time domain, it is likely complicated and dependent on the structure of the rhythm itself). Fortunately, we can verify this proposition for $3 \leq K \leq 16$ through numerical simulations. Sixteen is an arbitrary cut-off for K, simply being the highest we have tested, and it seems plausible that the $1, 0, 2$ pattern will continue indefinitely as K increases. The simulations actually suggest the coefficient of variation for the magnitudes of the higher coefficients also form a regular pattern, including that the $(K-1)$th coefficient is often the fourth smallest. It is worth highlighting how different this is to the properties of the DFT of an indicator (time-domain) vector whose most robust coefficient is the zeroth, then the first, then the second, and so on, up to the Nyquist frequency (at which point the pattern reflects) – this being a natural consequence of this DFT's domain being frequency.

Figure 13 shows the expected (mean) *coefficients of variation* (CV) of the magnitudes of the seven DFT coefficients that result from randomly jittered rhythms with seven onsets. (The coefficient of variation is standard deviation divided by mean, and so shows the extent of variability in relation to the mean; it is a dimensionless unit that is useful when comparing items that may be on different scales). To calculate these CVs, a total of 100,000 different rhythms were generated under a uniform random distribution over the length of the rhythmic period, then every event in each such rhythm was independently and randomly jittered with a normal distribution (modulo the period) 5,000 times. For every such perturbation, the magnitudes of the seven DFT coefficients were calculated. The CVs of these seven magnitudes were then calculated from these 5,000 different perturbations. The same process was repeated for all 100,000 different rhythms. The

resulting 100,000 different sets of seven CVs were then averaged to obtain an estimate of the mean CVs for the seven coefficients' magnitudes across all seven-onset rhythms. Repeated iterations of this procedure with perturbations of different standard deviations showed these estimates to be highly stable. The same process was also done for rhythms comprising 2 to 16 onsets (illustrations of these are provided in Supplementary Materials, Sec. 9, along with the MATLAB routines used to generate these simulations). In all cases the same pattern holds – the mean CV is smallest for the first coefficient (evenness), second smallest for the zeroth coefficient (balance), and third smallest for the second coefficient, as stated in 5.1.
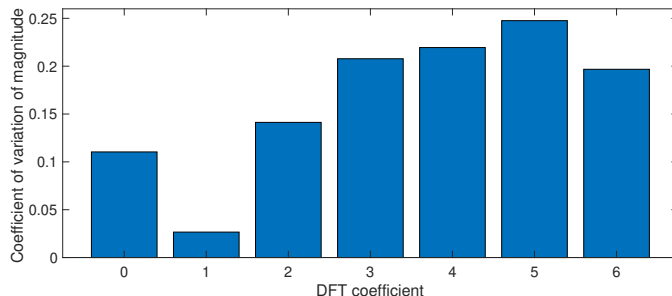
Figure 13: The mean values, across 100,000 different seven-onset rhythms, for the coefficients of variation (standard deviation divided by mean) of the magnitudes of the seven DFT coefficients resulting from (4) when the rhythms' onset times are independently and identically perturbed with a random normal distribution with a standard deviation of 1/60 of the period.

Taken together, these simulations demonstrate that percepts based on the first and zeroth coefficients of the DFT would be more robust to random temporal perturbations of the rhythm. This provides a teleological explanation for their perceptual relevance: balance and evenness allow for robust characterizations of human-perceived and human-produced – hence noisy – rhythmic patterns. Eventually, we hope to deepen our understanding of the DFT in perception to discover the role of some of the higher coefficients – the second coefficient, being the third most robust, certainly invites investigation; as does the last coefficient, which is usually the fourth most robust. The methods and theory used here are also directly applicable to any domain with discrete one-dimensional periodic patterns, such as spatial frieze patterns; it would be interesting to see whether these robust coefficients are also perceptually relevant in the visual domain.

It may also be possible to extend these methods to two-dimensional periodic patterns. In the visual domain, this would make the resulting features applicable to wallpaper patterns; in the musical domain, they would become applicable to the relationships between pitch class intervals as well as to the relationship between types of triads (three-tone chords) because every triad type (e.g., major, minor, or augmented) can be identified by just two directed pitch-class intervals. It would also be interesting to compare the predictive power of the coefficients of Argand vector's DFT with the coefficients of indicator vector's DFT (i.e., the circular moments of the rhythm's onset distribution).

5.2.2 Entropy

Predictive accuracy is beneficial to the survival of an organism, such as a human, in a complex and dynamically changing environment. Interonset interval entropy is one measure of the un-predictability of rhythms, hence it summarizes an important aspect of perception and cognitive processing. Due to the temporal nature of sound, predictability plays an important role in the auditory domain; notably in the form of expectancies that shape the music listening experience [92]. It is often postulated (particularly in the context of music, e.g., [93]) that if stimuli are too predictable (low in entropy), they are boring; if they are too unpredictable (high in entropy), they are unrewarding because predictions are unlikely to be correct and hence rewarded under psychological mechanisms such as Huron's proposed ITPRA [94]. An æsthetically optimal entropy is, therefore, often supposed to fall in a "Goldilocks zone" that is, for a given observer, neither too small nor too large [95]. Considering the importance of predictability in the auditory domain, it is surprising that this feature has been largely investigated in isolation. Here, we found that entropy interacts with other features. It seems that, for the rhythms tested here, rhythms with low IOI entropy can be particularly appealing when they also have low evenness and balance –

the complexities of the latter perhaps balancing out the simplicity of the former. Interestingly, however, our model comparisons showed that a squared entropy predictor adds little (its inclusion was not supported by cross-validation and its estimated effect size was negligible); hence, we find little evidence of a Goldilocks effects for IOI entropy alone. This suggests that future research using entropy as a measure of the complexity or predictability of periodic stimuli should consider deploying a more holistic approach and also consider its interactions with other predictors.

Testing the perceptual relevance of entropy is of broad theoretical interest. At the heart of many information processing theories of cognition lies the assumption that human cognition effectively operates on computational principles equipped to deal with situations of uncertainty (see [96] for a review). As such, entropy plays a pivotal role; furthermore, information processing theories in neuroscience often suggest that the brain continuously attempts to minimize entropy during information processing [97]. However, rather than exploring the computational principles underlying perception, other research has explored whether entropy itself is a perceptually relevant feature [6, 7]. Indeed, following this approach it was shown that the predictive coding perspective on appreciation of art is able to account for previously inexplicable violations of "good Gestalts" in great art works [98, 99]. Taken together, this previous research suggests that entropy may carry perceptual relevance in both evaluation and memory contexts. Here, we specifically explored interonset interval entropy, which captures the information content of one important aspect of temporal variability. We observed that it is not only perceptually relevant by itself, but also interacts with balance and evenness, and thereby modifies the perception of these features. From an information processing perspective, it makes sense that predictability modifies the perception of another feature. For example, a feature may have clear perceptual relevance as demonstrated in increased liking or enhanced memory. However, some of the possible values that this feature can assume may force the entire object into a very unpredictable form; this also needs to be considered. As a result, future work investigating specific stimulus features may want to consider including an interaction term of the feature of interest with entropy.

5.3   Framework for Perceptual Relevance

In addition to exploring the perceptual relevance of balance, evenness, and entropy, we were also concerned with providing a theoretical framework that allows us to operationalize perceptual relevance. The framework is detailed in section 3, and it implies that any experimental task that can show systematic effects of stimulus features on memory and/or evaluation has the potential to establish the perceptual relevance of those features.

Much published work on individual stimulus features implicitly operates under the same principle: a feature that systematically effects memory or evaluation must be relevant to perception (e.g., [37, 39]). However, we think it is important to provide a theoretical basis for this principle. We hope that this unified framework will help to bridge gaps between impactful work on the perceptual relevance of stimulus features across different modalities. As well as facilitating the design of future experiments for perceptual relevance, the framework could also be used to examine published research that may have unintentionally established perceptual relevance of stimulus features in experiments that were primarily interested in something else; for example, memory.

6   Conclusion

We provided a theoretical framework to test the perceptual relevance of stimulus features and deployed it within the domain of musical rhythm perception. We presented a wide variety of different periodic rhythms, twice each, to 177 participants. For each trial, we obtained a binary recognition response (have you heard the rhythm before in the experiment?) and an æsthetic evaluation on a seven-point liking scale. Using three high-level mathematical descriptions of rhythmic structure (balance, evenness, and interonset interval entropy) and two additional covariates (trial number and an indicator for whether or not the rhythm had been played twice), we were able to explain a meaningful proportion of the variance in recognition, memory bias, and liking responses. The results show that balance, evenness and IOI entropy are perceptually relevant features, and that they play distinct roles in the perception of periodic rhythms.

We also proposed a new theory for why evenness and balance are perceptually relevant features: each is a magnitude that characterizes a distinct aspect of an entire periodic rhythm, whilst both are also relatively robust to the small random temporal perturbations that are typical of the

rhythms perceived and produced by humans. This theory can be tested by undertaking further investigation of the perceptual relevance of the magnitudes of other coefficients of the discrete Fourier transform of a rhythm's Argand vector – notably, quantifying the relative importance of the DFT's second and last coefficients.

References

[1] S. Brown and J. Jordania, "Universals in the world's musics," *Psychology of Music*, vol. 41, no. 2, pp. 229–248, 2011.

[2] A. J. Milne and K. N. Olsen, "Pitch, timbre, and rhythm," in *The Science and Psychology of Music: From Beethoven at the Office to Beyoncé at the Gym* (W. F. Thompson and K. N. Olsen, eds.), ABC-CLIO, 2019.

[3] A. J. Milne, D. Bulger, S. A. Herff, and W. A. Sethares, "Perfect balance: A novel principle for the construction of musical scales and meters," in *Mathematics and Computation in Music* (T. Collins, D. Meredith, and A. Volk, eds.), vol. 9110 of *LNAI*, (Berlin), pp. 97–108, Springer, 2015.

[4] A. J. Milne, D. Bulger, and S. A. Herff, "Exploring the space of perfectly balanced rhythms and scales," *Journal of Mathematics and Music*, vol. 11, no. 2–3, pp. 101–133, 2017.

[5] A. J. Milne and R. T. Dean, "Computational creation and morphing of multilevel rhythms by control of evenness," *Computer Music Journal*, vol. 40, no. 1, pp. 35–53, 2016.

[6] E. Van Geert and J. Wagemans, "Order, complexity, and aesthetic preferences for neatly organized compositions," *Psychology of Aesthetics, Creativity, and the Arts*, 2019.

[7] M. T. Pearce, "Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation," *Annals of the New York Academy of Sciences*, 2018.

[8] A. J. Milne, S. A. Herff, D. Bulger, W. A. Sethares, and R. T. Dean, "XronoMorph: Algorithmic generation of perfectly balanced and well-formed rhythms," in *Proceedings of the 2016 International Conference on New Interfaces for Musical Expression (NIME 2016)*, (Brisbane, Australia), pp. 388–393, Griffith University, 2016.

[9] J. Yust, "Harmonic qualities in Debussy's 'Les sons et les parfums tournent dans l'air du soir'," *Journal of Mathematics and Music*, vol. 11, no. 2–3, pp. 155–173, 2017.

[10] S. Arom, *African Polyphony and Polyrhythm: Musical Structure and Methodology*. Cambridge, UK: Cambridge University Press, 1991.

[11] T. Fiore and C. Callender, "Perfect balance and the discrete Fourier transform [Special Issue]," *Journal of Mathematics and Music*, vol. 11, no. 2–3, 2017.

[12] G. T. Toussaint, *The Geometry of Musical Rhythm: What Makes a "Good" Rhythm Good?* CRC Press, 2013.

[13] I. Newton, *Opticks: or, a treatise of the reflexions, refractions, inflexions and colours of light. Also two treatises of the species and magnitude of curvilinear figures.* London, UK: William Innys, fourth ed., 1730.

[14] M. Meyer, "On the attributes of the sensations," *Psychological Review*, vol. 11, pp. 83–103, 1904.

[15] G. Révész, *Zur Grundlegung der Tonspsychologie*. Leipzig: Veit, 1913.

[16] A. Bachem, "Tone height and tone chroma as two different pitch qualities," *Acta Psychologica*, vol. 7, pp. 80–88, 1950.

[17] R. N. Shepard, "Circularity in judgements of relative pitch," *The Journal of the Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, 1964.

[18] J. London, *Hearing in Time: Psychological Aspects of Musical Meter*. Oxford: Oxford University Press, 2004.

[19] D. Deutsch, K. Dooley, and T. Henthorn, "Pitch circularity from tones comprising full harmonic series," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 589–597, 2008.

[20] J. A. Grahn, "See what I hear? Beat perception in auditory and visual rhythms," *Experimental Brain Research*, vol. 220, pp. 51–61, 2012.

[21] Wikipedia contributors, "List of musical symbols — Wikipedia, the free encyclopedia." https://en.wikipedia.org/wiki/List_of_musical_symbols#Notes_and_rests, 2019.

[22] A. Friberg and J. Sundberg, "Time discrimination in a monotonic, isochronous sequence," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2524–2531, 1995.

[23] G. L. Collier and J. L. Collier, "Microrhythms in jazz: A review of papers," *Annual Review of Jazz Studies*, pp. 117–139, 1996.

[24] A. Friberg and A. Sundström, "Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern," *Music Perception: An Interdisciplinary Journal*, vol. 19, no. 3, pp. 333–349, 2002.

[25] J. Frühauf, R. Kopiez, and F. Platz, "Music on the timing grid: The influence of microtiming on the perceived groove quality of a simple drum pattern performance," *Musicae Scientiae*, vol. 17, no. 2, pp. 246–260, 2013.

[26] L. Kilchenmann and O. Senn, "Microtiming in Swing and Funk affects the body movement behavior of music expert listeners," *Frontiers in Psychology*, vol. 6, no. 1232, 2015.

[27] O. Senn, L. Kilchenmann, R. von Georgi, and C. Bullerjahn, "The effect of expert performance microtiming on listeners' experience of groove in swing or funk music," *Frontiers in Psychology*, vol. 7, p. 1487, 2016.

[28] C. M. Conway and M. H. Christiansen, "Statistical learning within and between modalities," *Psychological Science*, vol. 17, no. 10, pp. 905–912, 2006.

[29] N. B. Turk-Browne and B. J. Scholl, "Flexible visual statistical learning: Transfer across space and time," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 35, no. 1, pp. 195–202, 2009.

[30] M. J. Cortese, M. M. Khanna, and S. Hacker, "Recognition memory for 2,578 monosyllabic words," *Memory*, vol. 18, pp. 595–609, 2010.

[31] M. Rohrmeier and P. Rebuschat, "Implicit learning and acquisition of music," *Topics in Cognitive Science*, vol. 4, pp. 525–553, 2012.

[32] N. Chater and P. Vitányi, "Simplicity: A unifying principle in cognitive science?," *Trends in Cognitive Sciences*, vol. 7, no. 1, pp. 19–22, 2003.

[33] J. Feldman, "The simplicity principle in perception and cognition," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 7, pp. 330–340, 2016.

[34] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt, "A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization.," *Psychological Bulletin*, vol. 138, no. 6, pp. 1172–1217, 2012.

[35] J. Wagemans, J. Feldman, S. Gepshtein, R. Kimchi, J. R. Pomerantz, P. A. van der Helm, and C. van Leeuwen, "A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations.," *Psychological Bulletin*, vol. 138, no. 6, pp. 1218–1252, 2012.

[36] K. Siedenburg, "Specifying the perceptual relevance of onset transients for musical instrument identification," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 1078–1087, 2019.

[37] S. C. Van Hedger, H. C. Nusbaum, S. L. Heald, A. Huang, H. P. Kotabe, and M. G. Berman, "The aesthetic preference for nature sounds depends on sound object recognition," *Cognitive Science*, vol. 43, no. 5, p. e12734, 2019.

[38] K. Schwabe, C. Menzel, C. Mullin, J. Wagemans, and C. Redies, "Gist perception of image composition in abstract artworks," *i-Perception*, vol. 9, no. 3, p. 2041669518780797, 2018.

[39] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?," in *CVPR 2011*, pp. 145–152, IEEE, 2011.

[40] L. Goetschalckx, P. Moors, S. Vanmarcke, and J. Wagemans, "Get the picture? Goodness of image organization contributes to image memorability," *Journal of Cognition*, vol. 2, no. 1, 2019.

[41] S. A. Herff, K. N. Olsen, A. Anic, and N. K. Schaal, "Investigating cumulative disruptive interference in memory for melodies, words, and pictures," *New Ideas in Psychology*, vol. 55, pp. 68–77, 2019.

[42] S. A. Herff, K. N. Olsen, and R. T. Dean, "Resilient memories for melodies: The number of intervening melodies does not influence novel melody recognition," *Quarterly Journal of Experimental Psychology*, vol. 71, no. 5, pp. 1150–1171, 2018.

[43] D. Deutsch, "Auditory pattern recognition," in *Handbook of perception and human performance* (K. R. Boff, L. Kaufman, and J. P. Thomas, eds.), vol. 2 Cognitive processes and performance, ch. 32, pp. 1–49, New York, NY, USA: John Wiley & Sons, 1986.

[44] J. A. Deutsch and D. Deutsch, "Attention: Some theoretical considerations," *Psychological Review*, vol. 70, no. 1, pp. 80–90, 1963.

[45] J. S. Snyder, C. M. Schwiedrzik, A. D. Vitela, and L. Melloni, "How previous experience shapes perception in different sensory modalities," *Frontiers in Human Neuroscience*, vol. 9, p. 594, 2015.

[46] R. N. Shepard and M. Teghtsoonian, "Retention of information under conditions approaching a steady-state," *Journal of Experimental Psychology*, vol. 62, no. 3, pp. 302–9, 1961.

[47] S. A. Herff, K. N. Olsen, R. T. Dean, and J. Prince, "Memory for melodies in unfamiliar tuning systems: Investigating effects of recency and number of intervening items," *Quarterly Journal of Experimental Psychology*, vol. 71, no. 6, pp. 1–45, 2018.

[48] S. A. Herff, R. T. Dean, and K. N. Olsen, "Inter-rater agreement in memory for melody as a measure of listeners' similarity in music perception," *Psychomusicology: Music, Mind, and Brain*, vol. 27, no. 4, pp. 297–311, 2017. accepted 10/16/2017.

[49] S. A. Herff, R. T. Dean, and N. K. Schaal, "Context effects of background babbling on memory for melodies," *Musicae Scientiae*, 2018.

[50] S. A. Herff, K. N. Olsen, J. Prince, and R. T. Dean, "Interference in memory for pitch-only and rhythm-only sequences," *Musicae Scientiae*, vol. 22, no. 3, pp. 344–361, 2018.

[51] S. A. Herff and D. Czernochowski, "The role of divided attention and expertise in melody recognition," *Musicae Scientiae*, vol. 23, no. 1, pp. 69–86, 2019.

[52] W. J. Dowling, "Tonal strength and melody recognition after long and short delays," *Perception and Psychophysics*, vol. 50, no. 4, pp. 305–313, 1991.

[53] S. Berman, D. Friedman, and M. Cramer, "ERPs during continuous recognition memory for words and pictures," *Bulletin of the Psychonomic Society*, vol. 29, no. 2, pp. 113–116, 1991.

[54] W. Donaldson and B. B. Murdock, "Criterion change in continuous recognition memory," *Journal of Experimental Psychology*, vol. 76, no. 3, Pt1, pp. 325–330, 1968.

[55] G. M. Olson, "Learning and retention in a continuous recognition task," *Journal of Experimental Psychology*, vol. 81, no. 2, pp. 381–384, 1969.

[56] N. A. Macmillan and C. D. Creelman, "Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes," *Psychological Bulletin*, vol. 107, no. 3, pp. 401–413, 1990.

[57] R. B. Zajonc, "Feeling and thinking: Preferences need no inferences," *American Psychologist*, vol. 35, no. 2, pp. 151–175, 1980.

[58] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tilyc, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.

[59] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[60] D. Müllensiefen and A. R. Halpern, "The role of features and context in recognition of novel melodies," *Music Perception: An Interdisciplinary Journal*, vol. 31, no. 5, pp. 418–435, 2014.

[61] D. Müllensiefen, B. Gringras, J. Musil, and L. Stewart, "The musicality of non-musicians: An index for assessing musical sophistication in the general population," *PLOS ONE*, vol. 9, no. 2, p. e89642, 2014.

[62] D. Moelants, "Hype vs. natural tempo: a long-term study of dance music tempi," in *The 10th International Conference on Music Perception and Cognition*, 2008.

[63] Applied Acoustic Systems DVM Inc., "AAS Chromaphone 2 acoustic object synthesizer plug-in VST AU AAX RTAS," 2018. Retrieved from https://www.applied-acoustics.com/chromaphone-2/.

[64] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[65] P.-C. Bürkner, "Advanced Bayesian multilevel modeling with the R package brms," *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018.

[66] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software*, vol. 76, no. 1, p. 10.18637/jss.v076.i01, 2017.

[67] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research, Cambridge University Press, 2007.

[68] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. Boca Raton, FL, USA: CRC Press, third ed., 2014.

[69] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, vol. 2, no. 4, pp. 1360–1383, 2008.

[70] A. Gelman, D. Simpson, and M. Betancourt, "The prior can often only be understood in the context of the likelihood," *Entropy*, vol. 19, no. 10, p. 555, 2017.

[71] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[72] R. D. McKelvey and W. Zavoina, "A statistical model for the analysis of ordinal level dependent variables," *Journal of Mathematical Sociology*, vol. 4, pp. 103–120, 1975.

[73] M. R. Veall and K. F. Zimmermann, "Pseudo-R^2's in the ordinal probit model," *Journal of Mathematical Sociology*, vol. 16, no. 4, pp. 333–342, 1992.

[74] T. M. Hagle and G. E. Mitchell, "Goodness-of-fit measures for probit and logit," *American Journal of Political Science*, vol. 36, no. 3, pp. 762–784, 1992.

[75] M. R. Veall and K. F. Zimmermann, "Evaluating Pseudo-R^2's for binary probit models," *Quality & Quantity*, vol. 28, pp. 151–164, 1994.

[76] P. D. Allison, R. Williams, and E. Moral-Benito, "Maximum likelihood for cross-lagged panel models with fixed effects," *Socius: Sociological Research for a Dynamic World*, vol. 3, pp. 1–7, 2017.

[77] P. Fraisse, *Les Structures Rhythmiques*. Louvain, Belgium: Publications Universitaires, 1956.

[78] B. H. Repp, "Sensorimotor synchronization: A review of the tapping literature," *Psychonomic Bulletin and Review*, vol. 12, no. 6, pp. 969–992, 2005.

[79] G. Aschersleben, "Temporal control of movements in sensorimotor synchronization," *Brain and Cognition*, vol. 48, no. 1, pp. 66 – 79, 2002.

[80] B. H. Repp, J. London, and P. E. Keller, "Production and synchronization of uneven rhythms at fast tempi," *Music Perception: An Interdisciplinary Journal*, vol. 23, no. 1, pp. 61–78, 2005.

[81] B. H. Repp, J. London, and P. Keller, "Distortions in reproduction of two-interval rhythms: When the "attractor ratio" is not exactly 1:2," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 2, pp. 205–223, 2012.

[82] B. H. Repp, J. London, and P. E. Keller, "Systematic distortions in musicians' reproduction of cyclic three-interval rhythms," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 3, pp. 291–305, 2013.

[83] S. Nozaradan, I. Peretz, M. Missal, and A. Mouraux, "Tagging the neuronal entrainment to beat and meter," *Journal of Neuroscience*, vol. 31, no. 28, pp. 10234–10240, 2011.

[84] S. Nozaradan, I. Peretz, and A. Mouraux, "Selective neuronal entrainment to the beat and meter embedded in a musical rhythm," *The Journal of Neuroscience*, vol. 32, no. 49, pp. 17572–17581, 2012.

[85] J. A. Grahn, "Neural mechanisms of rhythm perception: Current findings and future perspectives," *Topics in Cognitive Science*, vol. 4, no. 4, pp. 585–606, 2012.

[86] P. Janata, S. T. Tomic, and J. M. Haberman, "Sensorimotor coupling in music and the psychology of the groove," *Journal of Experimental Psychology: General*, vol. 141, no. 1, pp. 54–75, 2012.

[87] M. Davies, G. Madison, P. Silva, and F. Gouyon, "The effect of microtiming deviations on the perception of groove in short rhythms," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 5, pp. 498–511, 2013.

[88] M. A. G. Witek, E. F. Clarke, M. Wallentin, M. L. Kringelbach, and P. Vuust, "Syncopation, body-movement and pleasure in groove music," *PLOS ONE*, vol. 10, no. 9, p. e0139409, 2014.

[89] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL, USA: CRC Press, 2015.

[90] M. Saenz and D. R. Langers, "Tonotopic mapping of human auditory cortex," *Hearing research*, vol. 307, pp. 42–52, 2014.

[91] S. Nozaradan, "Exploring how musical rhythm entrains brain activity with electroencephalogram frequency-tagging," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 369, no. 1658, 2014.

[92] M. Pearce and G. A. Wiggins, "Auditory expectation: The information dynamics of music perception and cognition," *Topics in Cognitive Science*, vol. 4, pp. 625–652, 2012.

[93] G. Loy, *Musimathics*, vol. 1. Cambridge, MA, USA: MIT Press, 2006.

[94] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA, USA: MIT Press, 2006.

[95] D. E. Berlyne, "Novelty, complexity and hedonic value," *Perception and Psychophysics*, vol. 8, no. 5, pp. 279–286, 1970.

[96] J. Fan, "An information theory account of cognitive control," *Frontiers in Human Neuroscience*, vol. 8, p. 680, 2014.

[97] K. J. Friston and K. E. Stephan, "Free-energy and the brain," *Synthese*, vol. 159, no. 3, pp. 417–458, 2007.

[98] S. Van de Cruys and J. Wagemans, "Gestalts as predictions: Some reflections and an application to art," *Gestalt Theory*, vol. 33, no. 3, pp. 325–344, 2011.

[99] S. Van de Cruys and J. Wagemans, "Putting reward in art: A tentative prediction error account of visual art," *i-Perception*, vol. 2, no. 9, pp. 1035–1062, 2011.