

Modelling individual differences in chord pleasantness judgments

Joshua Frank and Peter M C Harrison

Centre for Music and Science, Faculty of Music, University of Cambridge

11 West Road, Cambridge, CB3 9DP, United Kingdom

Author Note

Joshua Frank  <https://orcid.org/0009-0007-5741-0456>

Peter M C Harrison  <https://orcid.org/0000-0002-9851-9462>

Experimental data are available via the Open Science Framework (<https://osf.io/6g5aq/>). The full online experiment's code is available via GitHub (<https://github.com/joshfrank95/2024-chord-pleasantness-experiment-PsyNet-code>), as are the full analysis scripts with key data (<https://github.com/joshfrank95/individual-differences-in-chord-pleasantness-2024>).

We are grateful to the UKRI: ESRC Doctoral Training Programme for funding (Grant number ES/P000738/1). The authors have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Joshua Frank, Centre for Music and Science, University of Cambridge, 11 West Road, Cambridge, CB3 9DP, United Kingdom. Email: jbf43@cam.ac.uk

This is the final accepted manuscript. The full article will appear in *Psychology of Aesthetics, Creativity, and the Arts*.

Abstract

A foundational question in empirical music aesthetics concerns how certain note combinations (chords) are perceived to be more pleasant than others. While normative patterns of chord pleasantness judgments are well-studied, relatively little is known about how listeners vary in these judgments. We address this question using a set of variance decomposition techniques: structural equation modelling (SEM) to disentangle preference variance from response noise, and Q-type principal component analysis (Q-PCA) to identify latent dimensions of participant variation. We apply these techniques to a new dataset where 106 online participants rated a set of 68 chords for pleasantness. SEM shows that the chords vary in both preference variance and response noise; while response noise varies relatively unsystematically, preference variance increases with familiarity and decreases with spectral interference (i.e., beating resulting from interactions between neighbouring partials). Q-PCA meanwhile shows that listener preferences vary on two primary dimensions: one corresponding to spectral interference, the other to cardinality (the number of notes in the chord). The former component is associated with musical expertise (more expertise means more disliking of interference), whereas the latter component does not correlate with any measured background variables. These findings confirm the previously established effect of musical expertise on chord preferences, but contrast with previous studies' claims that responses to interference are undifferentiated across participants. Our study provides a potential model for using variance decomposition techniques to study individual differences in other aesthetic domains.

Keywords: individual differences, chords, psychoacoustics, consonance, aesthetics

Background

The question of why certain combinations of tones sound more pleasant than others has been a subject of academic discourse in the West for over two millennia. Systematic empirical research in this area began around the turn of the 20th century, with the work of Hermann von Helmholtz and Carl Stumpf. Helmholtz (1954; first published 1862) considered the pleasantness of simultaneous tone combinations to be a dual function of *roughness* (the periodic fluctuations in amplitude, or *beating*, that occurs when two frequency components within a chord are closely spaced; also called *interference*) and a listener's individual characteristics. Stumpf (1890), on the other hand, proposed perceptual fusion of a sonority's constituent tones, based largely on their degree of match to a harmonic spectrum (*harmonicity*), as the operative mechanism. While Helmholtz is commonly credited with the first scientific investigation of this phenomenon, a relationship between beating effects (roughness) and pleasantness had been noted by scientists and music theorists at least as far back as 1700 (Cazden, 1962). Subsequent research has largely confirmed that psychoacoustic predictors, alongside stimulus familiarity (assessed by prevalence within a culturally relevant corpus), play a dominant role in the determination of average pleasantness ratings of chords, at least within a Western population (Smit *et al.*, 2019; Harrison & Pearce, 2020; Eerola & Lahdelma, 2021; Friedman *et al.*, 2021).

Historically, most research into the aesthetic evaluation of chords has used behavioural experiments to construct normative datasets for the pleasantness of individual chords, where pleasantness is quantified by averaging over all participants in a given group. This approach works very conveniently for stimulus-level computational modelling, where the researcher predicts the pleasantness rating from a collection of stimulus features such as interference and harmonicity. However, aesthetic responses to musical chords are not identical between listeners, with effects of musical training (Arthurs *et al.*, 2018; Popescu *et al.*, 2019; Smit *et al.*, 2019; Lahdelma & Eerola, 2020) and culture (McDermott *et al.*, 2016; Lahdelma *et al.*, 2021; Milne *et al.*, 2023) identified in prior research.

These studies provide evidence of variation between groups of participants with different educational or cultural backgrounds, but it remains possible that aesthetic responses vary further on an

individual level in a manner that is disguised by averaging over all participants in a group. An individual listener will naturally form part of several higher-level groups, be they social, cultural, educational, or otherwise, and their aesthetic preferences will be informed by this particular history. However, they will also have an idiosyncratic history distinct from the trends embodied in these groups; they may listen predominantly to a certain genre of music or even individual songs, enjoy particular kinds of multimedia entertainment, and have certain conscious aesthetic ideals that deviate from the modal ideals of their social/cultural group. Any number of such circumstances may lead to meaningful inter-participant differences in aesthetic appraisals, in ways that may either be common to multiple listeners or entirely unique.

The possibility of fine-grained differentiation in aesthetic responses to chords can be investigated using individual-differences designs. Several possible approaches may be taken in this regard.

Researchers may construct specialised metrics on which to measure listeners' individual levels of sensitivity to specific psychoacoustic parameters, and correlate these patterns of variation with patterns of variation in participants' ratings of musical chords varying along these parameters. McDermott *et al.* (2010) took this approach by measuring sensitivity to interference and harmonicity using custom acoustic stimuli, and then correlating these results with levels of differentiation in participants' judgments of consonant versus dissonant (according to Western music theory) chords, with the conclusion that harmonicity, but not interference, serves as an axis of individual difference in these judgments.

Alternatively, researchers may explore how the relationship between multiple judgments of the same musical chords differs between participants. This approach was taken by Song *et al.* (2022), who investigated correlations between judgments of preference and harmoniousness of musical chords ("preference for harmony") and showed that this measure correlated with an analogous measure of preference for harmony in colour pairs. While the results of their first experiment additionally suggested that listeners' different levels of preference for harmony may be uniquely linked to their preference for stimuli low in interference, their second experiment failed to replicate this finding, suggesting instead a role of preference for familiar chords. The results of a follow-up study (Song *et al.*, 2023) provided

support for the idea that both aversion to interference and preference for familiar stimuli track individual levels of preference for harmony. The authors interpret these findings in terms of a general preference for stimuli that are easier to encode perceptually, with this preference being the core factor varying between individuals.

These previous studies relied on the researchers specifying an *a priori* set of features that they believe listeners might draw on to make aesthetic judgments, such as interference, harmonicity, and familiarity. This approach is reasonable based on the strong evidence for the role of these features in determining pleasantness judgments at the stimulus level, but precludes the possibility of identifying other axes of individual difference arising organically from the data. Studies of group-level differences between cultures use a similar approach: for example, McDermott *et al.* (2016) and Milne *et al.* (2023) used small sets of predefined predictors to investigate differences in aesthetic judgments between cultural groups.

Here, we propose an alternative approach, in which we apply variance decomposition techniques that do not require any measurements other than pleasantness responses to a set of stimulus chords, and that make minimal *a priori* assumptions about relevant stimulus features. We investigate our data from two complementary perspectives. First, we decompose variation between participants *within* chords using variance analysis and structural equation modelling (SEM), to identify which chords show the greatest degree of variation in judgments and how this variation is split between reliable individual differences and response noise. Second, we decompose variation between participants *across* chords using Q-type principal component analysis (Q-PCA), to identify axes along which participants show the greatest variation in their response profiles. Instead of relying on feature sets specified *a priori*, these axes are defined solely in terms of the actual stimuli used in the experiment. This less restrictive approach allows us, in theory, to uncover arbitrary patterns of individual differences without limiting ourselves to *a priori* hypotheses about relevant features, while still allowing us to use such features *post hoc* to help us to interpret the results.

In the visual domain, such methodologies have already been used to identify meaningful dimensions of variation in individuals' aesthetic evaluations in diverse contexts, ranging from simple rectangles (McManus, 1980; McManus *et al.*, 2010) to fractal images (Spehar *et al.*, 2016) to scenes and faces (Bignardi *et al.*, 2024). These studies have typically shown that participant preferences do vary on a small number of dimensions, which can often be related to stimulus and participant background features.

Following this model, we thereby seek to address the following primary research questions:

1. Which chords elicit the greatest disagreement in pleasantness judgments?
2. To what extent does this disagreement reflect differences in latent preferences versus response noise?
3. Can these latent preferences be summarised using a low-dimensional representation?
4. How might these dimensions of individual differences relate to liking/disliking of particular acoustic features, and to participant background variables?

Musical Terminology

Before we proceed to the description of the study, we provide a short review of relevant musical terminology for readers new to the field of consonance modelling.

This paper concerns the perception of musical *chords*. A chord is defined as a collection of *tones* that are played simultaneously. Here we use the term “tone” to refer to a *complex tone*, that is, a complex of individual sine waves. The vast majority of pitched sounds are in fact complex tones; a vibrating body, be it a string, membrane, or air column, vibrates at multiple modes simultaneously when energised, and each of these modes produces oscillations at a given frequency that are carried through the surrounding medium. The composite auditory effect of these oscillations, which have frequencies and amplitudes dependent on the properties of the vibrating body, contributes to the percept of the tone's characteristic timbre, or tone colour.

The most common type of complex tone in the natural world is the *harmonic complex tone*. This is a complex tone whose components all have frequencies that are integer multiples of a shared *fundamental frequency*, i.e., they occupy a portion of the *harmonic series* above some highest common

divisor. The fundamental frequency in such tones is responsible for the perception of pitch; a higher frequency will result in a higher perceived pitch, and vice versa. Where complex tones are *inharmonic*, that is, where they do not embody this integer-multiple property, the perception of a single pitch is not guaranteed, and the auditory effect is noticeably different. The human voice, as well as most Western musical instruments, tend to create harmonic complex tones.

Tones can be identified by their fundamental frequency in Hz, but in musical contexts, it is often helpful to use a logarithmic *pitch* scale instead, because this better reflects perceptual judgments. In particular, we use the standard MIDI scale, where 60 corresponds to *middle C* on the piano (± 261.6 Hz), and moving up or down by integer values reflects movement by semitone intervals. In theory, MIDI notes can take non-integer values, but if we restrict consideration to integer values, this gives us the 12-tone *chromatic scale* that is the basis of most Western music compositions.

Due to the unique nature of the type of tone used in our stimuli, discussed later in this paper, we make minimal reference to *pitches* directly, using instead the related concept of *pitch classes*. A pitch class may be understood intuitively as referring to a single named note of the chromatic scale, e.g., F#. More formally, pitch classes are a representation of pitches in which the notion of *octave invariance* is applied. Octave invariance refers to the phenomenon whereby two notes separated by exactly one octave are perceived as qualitatively similar: an F# played at one position on the piano *sounds like* an F# played at any other position, in a way that it *does not sound like* a G. There are thus only twelve pitch classes, corresponding to the twelve named notes in the chromatic scale. By convention, these are represented as zero-indexed integers, with C = 0, C# = 1, up to B = 11.

Musical chords thus involve the simultaneous sounding of multiple pitch classes, which we can represent as *pitch-class sets*. As in mathematical set theory, elements of a pitch-class set are inherently unordered: [1, 2, 3] is exactly equivalent to [3, 2, 1] or any other permutation of the elements 1, 2, and 3. The major triad C, E, G can thus be represented as [0, 4, 7] or any permutation of these elements. From a musical perspective, representation of a chord as a pitch-class set discards any information regarding the

chord's *inversion*, that is, the identity of the lowest note in the chord. As discussed later in this paper, ignoring inversional information is appropriate in the context of our stimuli.

When analysing the perception of our musical chords, we compute various features from the chord itself. Some of these features refer to further music-theoretic concepts that are useful to define now.

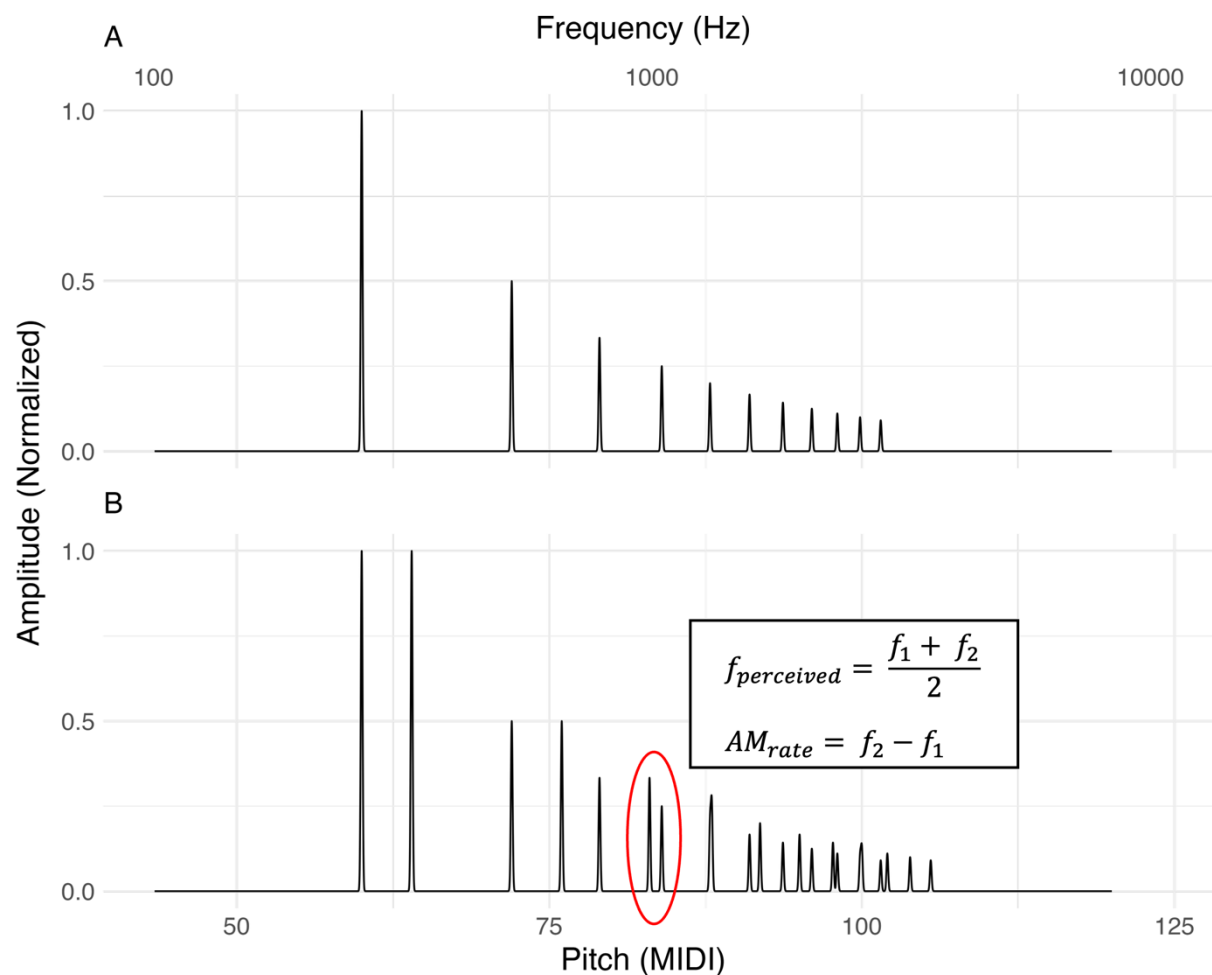
One such concept is the *major scale*. The major scale is a subset of the chromatic scale, comprising seven pitch classes which, when stacked in ascending order, are formed of a characteristic set of stepwise intervals (in semitones): 2, 2, 1, 2, 2, 2. For example, the C major scale comprises the white notes of the piano: C, D, E, F, G, A, B. The major scale is highly privileged in a Western musical context (Johnson-Laird et al., 2012), being by far the most commonly heard scale and viewed by music theorists as holding conceptual and structural priority within the Western tonal vocabulary.

Another important concept is the *chord root*. What leads us to call, for example, the chord consisting of the notes C, E, and G not only a major chord (defined by its intervallic content, or *quality*), but specifically a *C major* chord? In the majority of Western music today, and indeed for the past 400 years, listeners perceive chords as having specific tonal functional properties. These properties are highly contingent on each chord being perceived as anchored on a specific, usually singular, pitch class. When we hear the simultaneous notes C, E, and G, regardless of their register or the ordering of notes in pitch height, we identify the note C as holding a unique part of the chord's identity, and refer to it as the chord's root pitch class. Identifying the root of simple triadic chords is usually trivial, but this becomes more complex in the case of structurally uncommon chords, leading to the development of sophisticated algorithmic approaches to the problem (e.g., Parncutt, 1988).

Some of our other features, which we term *psychoacoustic*, are computed with reference to the chords' *frequency spectra*. A frequency spectrum breaks down a sound into its constituent sine-wave components, representing these in a two-dimension frequency-amplitude plane. This representation is useful for computing various features that are determined by the interaction and/or arrangement of individual sine-wave components of the chords. The x-axis may be designed to reflect raw frequency, log-frequency, or MIDI pitch (in which case the representation becomes a *pitch spectrum*).

Two psychoacoustic features of particular importance in the study of aesthetic responses to chords are *harmonicity* and *roughness* (or *interference*). Harmonicity refers, broadly speaking, to the degree to which the frequency- or pitch-amplitude spectrum of a chord resembles that of a single harmonic complex tone. Panel A of Figure 1 shows such a spectrum in idealised form; the key aspects of what makes this spectrum harmonic is that its constituent frequency components are spaced at integer multiples (in Hz) of a common frequency. A single harmonic complex tone will have maximal harmonicity. Panel B of Figure 1, on the other hand, shows a chordal spectrum that clearly deviates from perfect harmonicity. While a highest common divisor may still be found for these frequency components, the harmonic spectrum above that value would contain many frequencies not present in the chord's spectrum. Thus, the harmonicity of this chord will be lower.

Roughness, or interference, is a perceptual property that does not emerge to any meaningful degree in isolated harmonic complex tones, as it arises only when frequency components of a spectrum are closely spaced. An example of such spacing is highlighted in panel B of Figure 1, along with a summary of the resulting perceptual effect. Closely spaced frequency components give rise to amplitude modulation, which above a certain rate manifests in the grating, typically unpleasant sensation of roughness. When calculating the interference value of a chord, the typical procedure involves summation of the roughness effects resulting from all individual pairs of frequency components that are sufficiently closely spaced.

Figure 1*Pitch-Amplitude Spectra of Idealized Harmonic Complex Tones*

Panel A of Figure 1 shows the pitch-amplitude spectrum (with frequency on a log scale along the top axis) of a single idealized harmonic complex tone, specifically the note middle C (MIDI 60). The lowest frequency component acts as the fundamental, and is responsible for the identification of the pitch of the note. The upper frequency components, termed *harmonics*, follow a regular pattern along the pitch (or frequency) axis. Panel B shows the pitch-amplitude spectrum of a chord consisting of the notes MIDI 60 and MIDI 64 (C and E). The frequency components are no longer all multiples of a common fundamental frequency, so we term them *partials* rather than harmonics. Their arrangement is more complex than in the case of only a single tone.

Interactions between frequency components produce distinctive acoustic effects. In particular, if two partials are close in frequency (see circled region), then a beating effect is induced, corresponding to amplitude modulation (i.e., periodic increases and decreases in the sound pressure level) at a rate equal to the frequency difference between the partials (see inset), with a perceived frequency (of the carrier wave) equal to the mean of the interacting frequencies. At low amplitude modulation rates (e.g., 3 Hz), this is heard as periodic fluctuation in loudness; at higher frequencies (e.g., 30 Hz), this is heard as (typically unpleasant) *roughness*.

Methods

Participants

$N = 106$ participants completed the experiment (mean age = 41.1, SD = 12.3, 50 female, 53 male, 3 non-binary). For the correlation analyses in our study, a sample size of 85 was determined to be sufficient for a statistical power of .80 given a moderate effect size ($r = .30$); we recruited a sample slightly larger than this requirement. Participants were recruited via Prolific, using inbuilt prescreening for normal hearing. 93 participants reported no formal training in Western art music, with the remaining 13 reporting between 0.5 and 13 years of training (overall mean = 0.62, SD = 2.1 years). Participants were compensated at a rate of approximately £10.00 per hour.

Ethics statement

This experiment was approved by the University of Cambridge Faculty of Music Research Ethics Committee.

Stimuli

We wished to use a stimulus set that systematically probed the chords available in Western tonal music. However, the space of all possible chords playable on a standard classical piano is vast, far too many for a single study. In order to simplify this space, we adopted two strategies in combination.

The first strategy was to adopt Shepard tones. Shepard tones are a special kind of synthetic tone whose spectral envelope is constructed to hold perceived pitch height constant across the entire playable range (Shepard, 1964; see Parncutt *et al.*, 2023 for a study using tones with a somewhat related

construction). Correspondingly, there are only 12 Shepard tones playable within the standard 12-tone chromatic scale (one for each pitch class, $C = 0$ through $B = 11$), rather than the 88 keys on the piano keyboard, which massively reduces the number of chords to choose from. Chords played with Shepard tones are thus invariant to octave transposition (i.e., they sound the same when transposed up or down an octave), and they are also invariant to inversion (i.e., there is no sense in which the notes in the chord can be reordered in terms of pitch height). The property of inversion invariance means that, when using Shepard tones, there are no intervals greater than a tritone (6 semitones); for example, ascending by 11 semitones is equivalent to descending by 1 semitone. This is because there is no sense in which the two notes are ordered in pitch height: a chord comprising E and F, for example, will be heard as equivalently representing a one-semitone interval E–F and an eleven-semitone interval F–E. These invariance properties dramatically reduce the space of chord types from which we wish to sample. In practice, the perceptual effect of a Shepard-tone chord is somewhat similar to an orchestral rendition of a chord, where the instruments are too numerous to pick out individually and it is hard to determine exactly what pitch heights are present. Figure 2 illustrates the spectral structure of chords composed of Shepard tones.

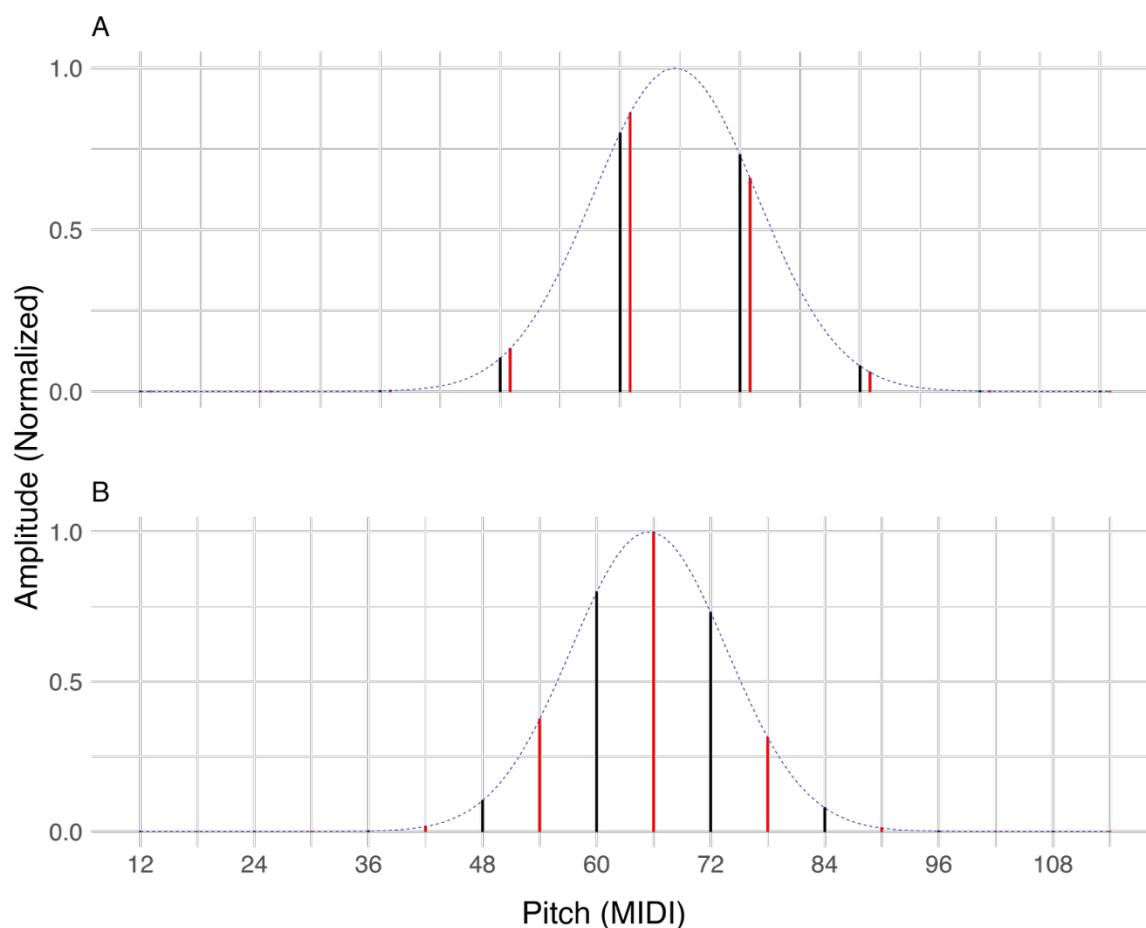
The second strategy was to reduce our stimulus set further by taking into account *transpositional invariance*. Informally, transposing a chord means to move it along the pitch scale; more formally, transposing pitch classes is defined as addition modulo 12 (e.g., transposing pitch class 1 up by 3 semitones gives 4, and transposing pitch class 10 up by 3 semitones gives 1). Transpositional *invariance* describes the way in which, at least for Western listeners, two chords will sound similar if one chord is produced by transposing all notes in the other chord by the same interval. For example, if we denote pitch class sets in the form $[a, b, \dots]$, then the chord $[0, 4, 7]$ would be transpositionally equivalent to the chord $[1, 5, 8]$. Indeed, music theorists would describe both of these chords as ‘major triads’, and would consider them to be equivalently consonant (in the absence of musical context). It would be redundant to test multiple chords that are transpositions of each other; we therefore reduce our stimulus set by discarding all chords that duplicate each other under transposition, just leaving one representative in each case (i.e., we keep $[0, 4, 7]$ and discard $[1, 5, 8]$). After discarding transpositional duplicates, there are still

multiple ways to represent a chord as a pitch class set, corresponding to all possible permutations of the set's constituent elements; we follow standard music-theoretic conventions for determining which representative to keep (Rahn, 1980; Harrison & Pearce, 2018b).

We used 68 chords as stimuli, representing all possible chords of cardinality 2 (6 chords), 3 (19 chords), and 4 (43 chords). Stimuli were rendered using Shepard tones with a total of nine octave-spaced partials mapped onto a Gaussian amplitude curve with central frequency of MIDI 65.5 (359.46 Hz) and standard deviation of 8.2 semitones. Chords were rendered with nominal bass-note pitch class randomly assigned between 0 (C) and 11 (B) for each chord for each participant, and had a 3-second duration with constant amplitude.

Figure 2

Pitch-Amplitude Spectra of Two-Note Shepard-Tone Chords, with Gaussian Amplitude Curve Superimposed



Note. A: Pitch-amplitude spectrum of a Shepard-tone chord comprised of the pitch classes C (black lines) and C# (red lines). This is equivalently perceived as an interval of one semitone (C to C#) or eleven semitones (C# to C). The black and red lines are close together, reflecting the relatively high degree of roughness present in this chord.

B: Pitch-amplitude spectrum of a Shepard-tone chord comprised of the pitch classes C (black lines) and F# (red lines). This is perceived as a tritone, but with ambiguity as to which is the “lower” pitch class. The black and red lines are spaced further apart, reflecting the relatively low degree of roughness present

in this chord. The dashed blue curve in both figures reflects the Gaussian amplitude function onto which the octave-spaced representatives of each pitch class in the chord are mapped.

Procedure

Chord ratings

Participants signed a consent form prior to undertaking the experiment, followed by a free volume adjustment and an automated test to ensure they were using appropriate headphones. They then rated all 68 stimuli for pleasantness (5-point Likert-type scale, ranging from “very unpleasant” to “very pleasant”), hearing each stimulus only once without the option to repeat playback. Participants were allowed to give their answers before the full 3-second duration of each stimulus had elapsed, and were also free to consider their answers for a longer time, with submission of an answer triggering the next stimulus presentation. Stimuli were presented in a randomised order. A 20-second break followed after participants had rated all stimuli, before they rated the full set of chords again, in a new randomised order but with the same pitch classes as in their first rating. This double-rating paradigm was chosen to allow us to examine participants’ reliability, and to achieve a more accurate estimate of systematic variance at the stimulus level through our SEM analyses below. The rating portion of the experiment took an average of just under nine minutes to complete.

Questionnaire

After giving their ratings, participants filled out a questionnaire capturing background information of various kinds. We captured age; gender (free-text response); occupational status; answers to a subset of questions from the Goldsmiths Musical Sophistication Index (GMSI; Müllensiefen *et al.*, 2014) within the categories of Musical Training, Singing Abilities, Perceptual Abilities, and Emotion; linguistic history; years of formal musical training; handedness; and how frequently participants listened to various musical genres and viewed various film genres. The full experiment, including stimulus generation and the questionnaire, is available as a PsyNet script via GitHub (see Data and Code Availability below).

Stimulus feature extraction

We extracted the following features from the stimuli, for use in analysis of the results:

1. **Interference**, using the model of Hutchinson & Knopoff (1978).
2. **Harmonicity**, using the model of Milne (2013). This model was chosen as it can retain the actual spectrum of the stimulus, rather than using idealised spectra. Another model capable of this is that of Harrison & Pearce (2018), which we also tested, but decided not to use due to its high collinearity with cardinality in our dataset.
3. **Familiarity**, assessed as corpus prevalence using the model of Harrison & Pearce (2020) with chord counts aggregated appropriately to reflect the perceptual invariances of Shepard tones, similar to the procedure of Eerola & Lahdelma (2021).
4. **Cardinality** (the number of pitch classes in the chord).
5. **Major-scale compatibility**, measuring the number of distinct ways in which a chord can be formed from the notes of a major scale, as in Parncutt *et al.* (2023), and returning the proportion of all possible combinations accounted for by the chord in question. For example, in a C major scale, the major triad can be formed between the notes C-E-G, F-A-C, and G-B-D, for a total of three appearances. The total number of combinations of the seven notes of a major scale is 127, so the major triad will have a major-scale compatibility value of $3/127 = .0236$.
6. **Major-scale compatibility (binary)**, which takes a value of 1 if major-scale compatibility is greater than 0, and a value of 0 otherwise. This feature is called *Tonal Consonance* in Eerola & Lahdelma (2021).
7. **Interval counts**, one for each possible pitch-class interval (1, 2, 3, 4, 5, 6 semitones), counting the number of times that interval occurs in the chord.
8. **Scale-degree presence**, computed by first establishing the pitch class of the chord root using Parncutt's (1988) root-finding algorithm. Relative to a given root pitch class, scale degrees represent the distance between the pitch classes in the chord and the chord root, so we compute these by subtracting the chord root pitch class from each pitch class in the chord, and assign a

value of 1 to each present scale degree and 0 to each absent scale degree. The chord root itself (scale degree 0) is present in all chords, so we omit this feature from the modelling.

Data pre-processing

We removed all data from participants who did not complete the full experiment, and z-scored ratings within participants.

Results

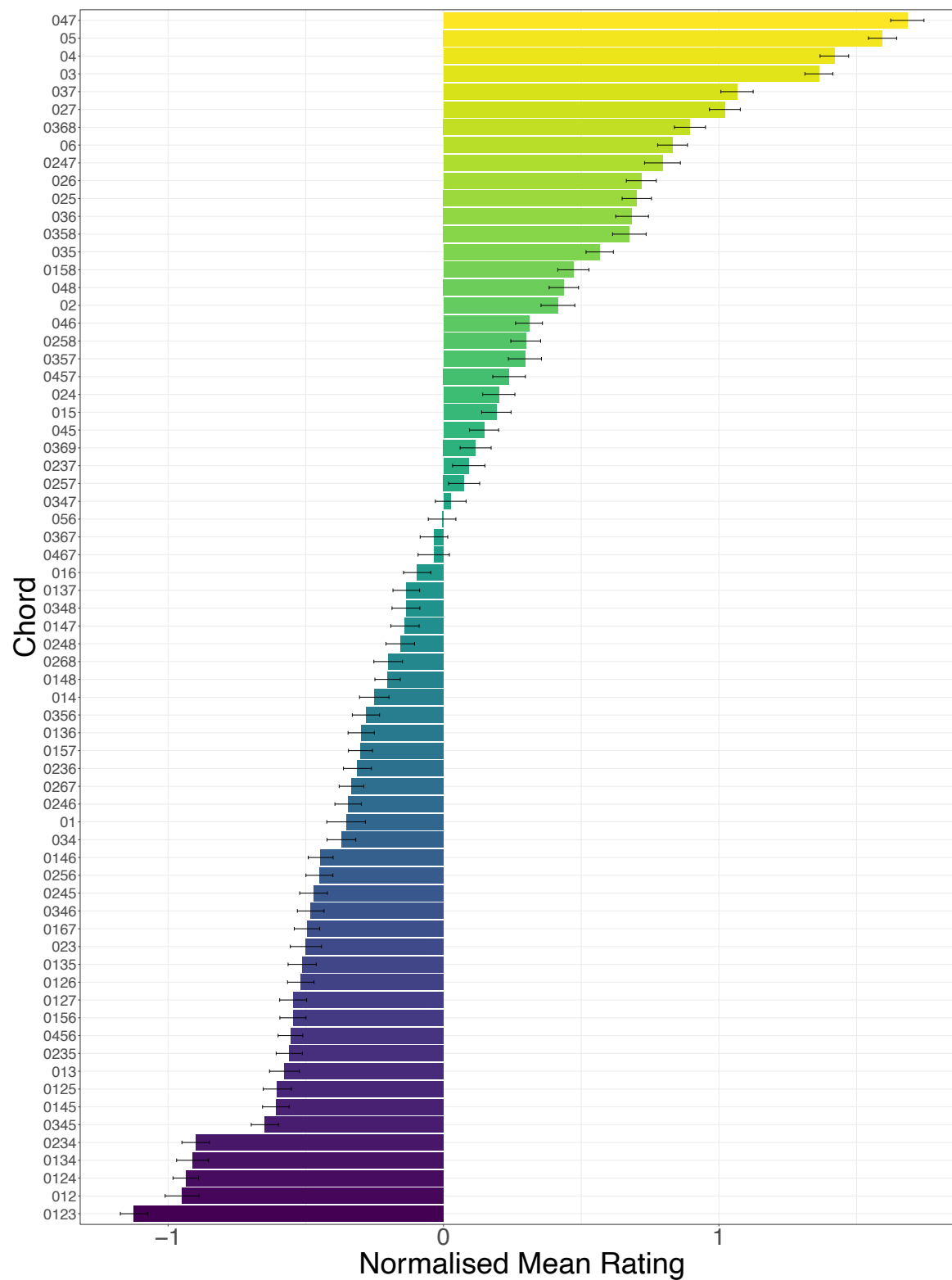
Predicting average ratings

Population-level ratings are highly reliable and consistent with previous work (Bowling et al., 2018)

Individual participants gave moderately reliable ratings on a trial-by-trial basis (mean test-retest reliability of .53). Once averaged over participants to give a single score for each chord in each test phase, the correlation between test and retest scores is .99, indicating that the chord-level pleasantness ratings are highly reliable. Average ratings per chord are shown in Figure 3 below.

Figure 3

Normalised Average Ratings with Standard Errors for Each Chord in the Dataset



Given that the Shepard tone has not been used in this area of research before, we wished to check that our rating data were broadly in line with those of prior research that has used more traditional timbres. In part this can be seen through inspection of Figure 3, in which, as expected, the higher ratings are dominated by chords traditionally considered consonant, such as the major and minor triads and the major third/minor sixth, minor third/major sixth, and perfect fourth/fifth dyads, while the lower ratings are dominated by traditionally dissonant chords containing semitone and whole-tone clusters. To explore this more systematically, we compared our population-level ratings to published data from Bowling *et al.* (2018). This dataset provides mean attractiveness ratings for 298 different chords played using a synthesised piano. To make the two datasets directly comparable, we averaged the latter dataset's ratings over all inversions of a given chord. We performed our correlations separately for each level of chord cardinality, because Bowling *et al.* used a blocked design that makes ratings not directly comparable across chords of different cardinalities. The correlations are very strong (Table 1), indicating that the two experiments deliver highly consistent results. In a follow-up analysis (Supplementary Materials), we explored these correlations further, examining what might predict the small differences in the ratings; we found in particular that our dataset elicits slightly reduced pleasantness for intervals of a second/seventh, and slightly increased pleasantness for the minor third/major sixth and the tritone. This is consistent with the fact that Shepard tone spectra include energy at the octave (which would enhance interference for seconds/sevenths) but do not include energy at the major third or perfect fifth (which would enhance interference for the minor third/major sixth and the tritone).

Table 1

Correlations Between Our Results and the Results of Bowling et al. (2018)

Cardinality	Correlation (Pearson's r)
Dyads	.94 ($p < .01$)
Triads	.95 ($p < .001$)
Tetrads	.91 ($p < .001$)

Psychoacoustic features reliably predict mean pleasantness judgments

Individual differences are best interpreted with respect to their population baselines. We therefore precede our individual-differences analyses with a set of population-level analyses, replicating and extending previous studies that use computational models to predict chord judgments as a function of stimulus features. This has the additional advantage of adding the Shepard tone to the set of timbres that have been utilised within this modelling tradition.

We constructed six regression models to predict average ratings of our stimuli. Each model uses a different set of features as predictors, designed to capture different aspects of the stimuli and/or different theoretical perspectives on consonance (Table 2). These can be split into models utilising notation-based features (interval counts, scale degrees, and their combination as the Notation model); a well-established model shown to have high explanatory power in prior research (Classic model; Harrison & Pearce, 2020); an extension of this model including the less well studied measures of major-scale compatibility (Classic +); and a model utilising all possible extracted features. Whilst all models achieve statistical significance (see Supplementary Materials for more details), AIC analyses (Table 3) indicate that the Classic + model is the best-performing, with an adjusted R^2 of .90 (95% CI: [.85, .92]), only marginally lower than the model including all predictors (.93, 95% CI: [.89, .93]; confidence intervals generated using the bias-corrected accelerated bootstrap, 10,000 replicates). Interference provides the strongest contribution to the model, followed by major-scale compatibility, and then familiarity (Table 4). Harmonicity, cardinality, and major-scale compatibility (binary) are not statistically significant. An analogous analysis using robust regression delivers equivalent results (Supplementary Materials).

Table 2*Description of Feature Sets Used as Predictors in Linear Models*

Model	Interference	Harmonicity	Familiarity	Cardinality	Major-scale compatibility (binary)	Major-scale compatibility	Interval counts	Scale-degree presence
Full	✓	✓	✓	✓	✓	✓	✓	✓
Classic +	✓	✓	✓	✓	✓	✓		
Classic	✓	✓	✓	✓				
Interval							✓	
Scale degree								✓
Notation							✓	✓

Table 3*Model AIC Comparisons*

	K	AICc	Delta AIC	AICc weight	Cumulative AICc weight	Log-likelihood	Adjusted R^2
Psychoacoustic	8	48.44	0.00	0.99	0.99	-15.00	.8985
Classic	6	57.94	9.50	0.01	1.00	-22.28	.8783
Full	24	59.04	10.61	0.00	1.00	8.43	.931
Notation	19	96.63	48.19	0.00	1.00	-21.40	.8506
Interval	8	98.52	50.08	0.00	1.00	-40.04	.7881
Scale degree	13	144.65	96.21	0.00	1.00	-55.95	.6314

Table 4*Parameters of the Psychoacoustic Features Model*

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE (B)	<i>t</i>	<i>p</i>
Intercept	N/A	N/A	N/A	0.300	[-0.487, 1.08]	0.379	0.791	.432
Interference	-0.665	[-0.758, -0.583]	0.048	-2.041	[-2.44, -1.72]	0.148	-13.758	< .0001
Harmonicity	0.088	[-0.015, 0.197]	0.057	0.823	[-0.149, 1.76]	0.534	1.540	.129
Familiarity	0.180	[0.077, 0.284]	0.041	1.766	[1.18, 6.00]	0.405	4.356	< .0001
Cardinality	-0.057	[-0.183, 0.064]	0.049	-0.055	[-0.180, 0.063]	0.047	-1.159	.251
Major-scale compatibility (binary)	-0.029	[-0.130, 0.092]	0.057	-0.037	[-0.169, 0.114]	0.074	-0.502	.617
Major-scale compatibility	0.246	[0.092, 0.402]	0.071	13.241	[4.70, 22.3]	3.847	3.442	.001

Note. Adjusted $R^2 = .8985$ (95% CI: [.845, .924]). 95% confidence intervals were generated with bootstrapping (10,000 replicates).

Individual differences

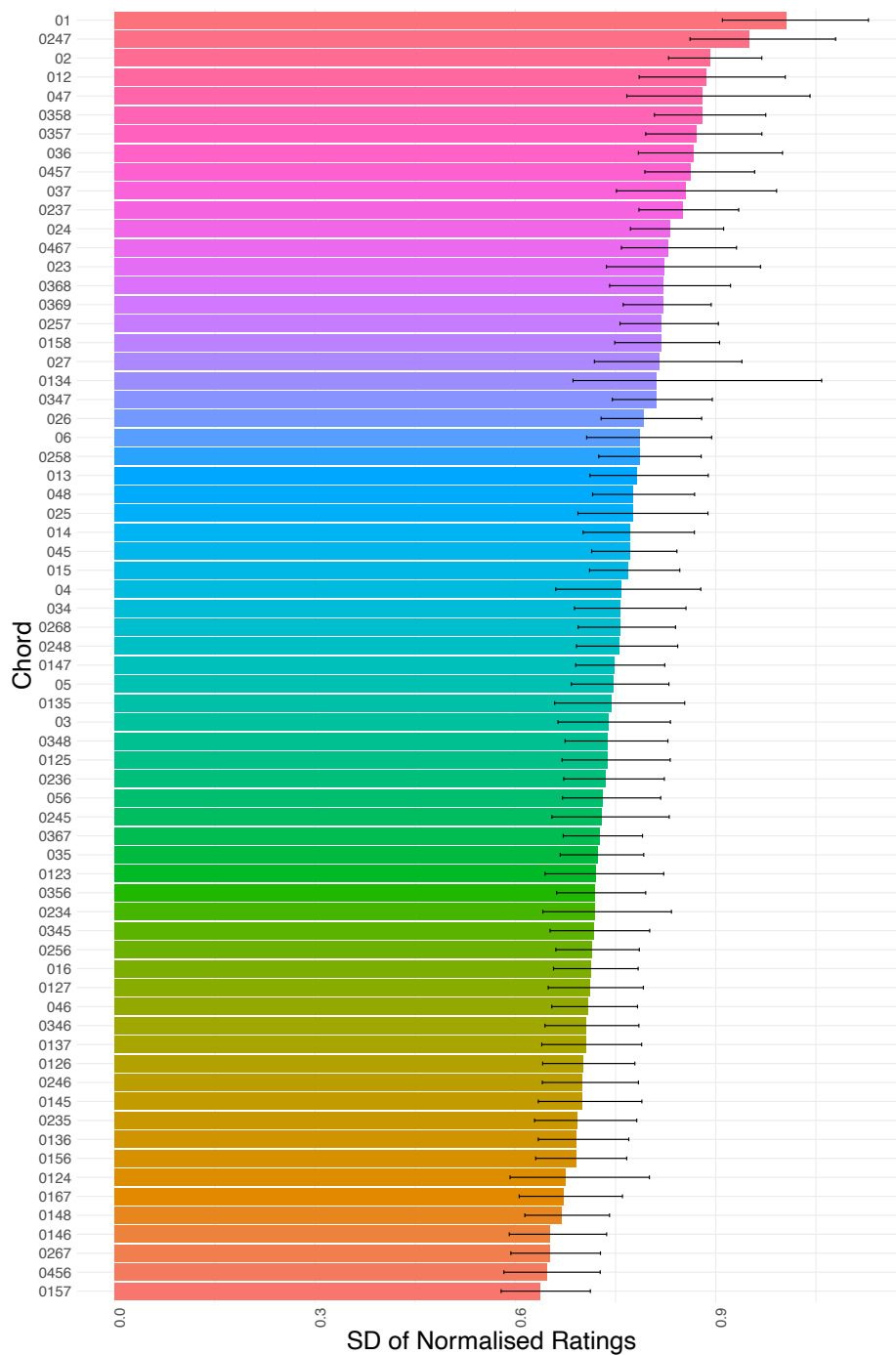
Smooth and familiar chords support a greater degree of individual difference

As a first exploration of individual differences in our data, we assessed whether the standard deviations of ratings differ between chords, visualised in Figure 4. This acts as a coarse approximation of whether certain chords show higher levels of disagreement between participants than others. On average, the SD is about 0.77, but there is some variation. The chord with the highest SD (1.01) is [0, 1]; the lowest (0.64) is [0, 1, 5, 7]. Interestingly, both chords are relatively rough. However, the former chord is a common component of many higher-cardinality chords, including many musically common chords in its inverted form (equivalent in Shepard timbre) of [0, 11]. The semitone/major seventh is the roughest dyad,

but is not particularly rough in comparison to many higher-cardinality chords, with an interference value falling just below the third quartile of the overall dataset. The [0, 1, 5, 7] chord also falls between the mean and third quartile, with a value slightly lower than the semitone/major seventh dyad, reflecting the Hutchinson & Knopoff algorithm's normalisation by overall amplitude spectrum. Looking at the full dataset, we find that correlations between SD and chord features are small, but reliable. Interference ($r = -.28, p = .02$) and cardinality ($r = -.35, p = .003$) are negatively correlated with SD, indicating that rough chords tend to display less variation in ratings; all else being equal, chords with higher cardinality also tend to be rougher. Familiarity ($r = .28, p = .019$), major-scale compatibility ($r = .42, p = .0003$), and major-scale compatibility (binary; $r = .33, p = .005$) show positive correlations with SD; these metrics all capture aspects of a chord's common-practice usage, indicating that more well-known chords are more differentiated in individual pleasantness ratings.

Figure 4

Standard Deviations (SD) of Ratings (Averaged Over Test and Retest Ratings) For Each Chord, With Bootstrapped 95% Confidence Intervals (10,000 Replicates)



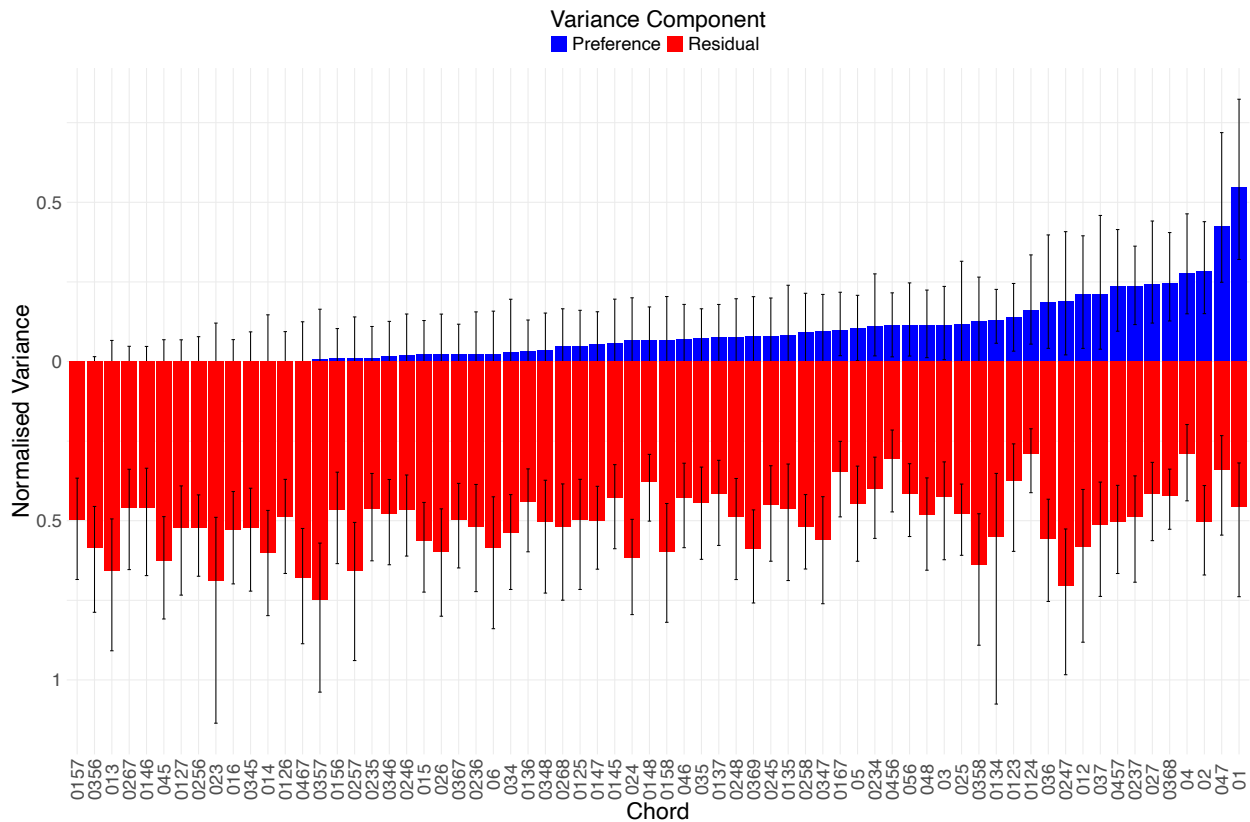
Variation in chord ratings, as captured by SD, can come from two sources. One is variance in underlying preferences: some participants like the chord more than other participants. The other source is

rating noise: for example, participants may be less sure of their ratings for certain chords, leading to higher variability that is not due to reliable differences between participants.

We conducted some structural equation modelling (SEM) analyses to explore these possibilities. SEM estimates a decomposition of overall variance in a dataset into portions attributable to a latent variable and to residual variance. We created one SEM for each chord in the dataset, with test and retest ratings as inputs, and the latent variable (which captures the covariance between test and retest ratings) acting as a measure of stable individual differences. The resulting distributions of variances over all chords are presented in Figure 5.

Figure 5

Variance Decomposition Estimated by Structural Equation Models Over All Chords, With Bootstrapped 95% Confidence Intervals (10,000 Replicates)



Note. Residual variances are the mean of test and retest residuals, calculated per iteration during bootstrapping. Factor loadings in all models were constrained to equality.

On a chord-wise basis, the variance structure appears to be dominated by measurement noise; however, this is not the case over all chords equally. Correlating the two types of variance estimated by the SEMs with chord features, we find that similarly to the SD analysis above, interference ($\rho = -.27, p = .028$) and cardinality ($\rho = -.25, p = .044$) are negatively correlated with preference variance, while familiarity ($\rho = .49, p < .0001$) and major-scale compatibility ($\rho = .25, p = .036$) are positively correlated with preference variance. No reliable correlations are present between chord features and measurement noise. This pattern can be understood as particularly rough chords eliciting more reliably negative responses across participants, leading to lower differentiation in response profiles, and familiar chords in general having more stable individual aesthetic associations, possibly engendered through exposure; low-interference chords also tend to be more familiar, leaving more possibility for exposure to chords with lower levels of interference.

The semitone/major seventh dyad and the major triad stand out in this analysis, having by far the highest variance in preferences between participants. The beating effect of interference in the semitone/major seventh dyad is a clear and isolated percept, whereas in chords of higher cardinality this effect is often both much more unpleasant and possibly obscured by other features. This type of beating may be an interesting auditory phenomenon to which some listeners are drawn and by which others are put off. This effect may not manifest under all timbral conditions, and can neither be systematically confirmed nor generalised outside our sample group, but it may be an interesting stimulus feature for further study. As noted previously, despite the prominence of this beating effect, the [0, 1] chord is not particularly rough in comparison to many higher-cardinality chords; this may allow it to avoid being universally unpleasant, which would allow the pronounced beating effect to engender different responses across participants. The major triad having a high degree of individual differences in ratings would make sense given its ubiquity in popular music, as people may form different associations with various songs or genres prominently featuring this chord. It is also quite smooth, making universal negative responses unlikely. It should be noted that the major triad is responsible for a large positive skew in the distribution

of familiarity values, as it is 4.5 times as familiar as the next-most-familiar chord in the stimulus set; this may play an outsized role in driving the observed correlation between familiarity and preference variance.

Individual differences in pleasantness judgments are dominated by varying preferences for interference and cardinality

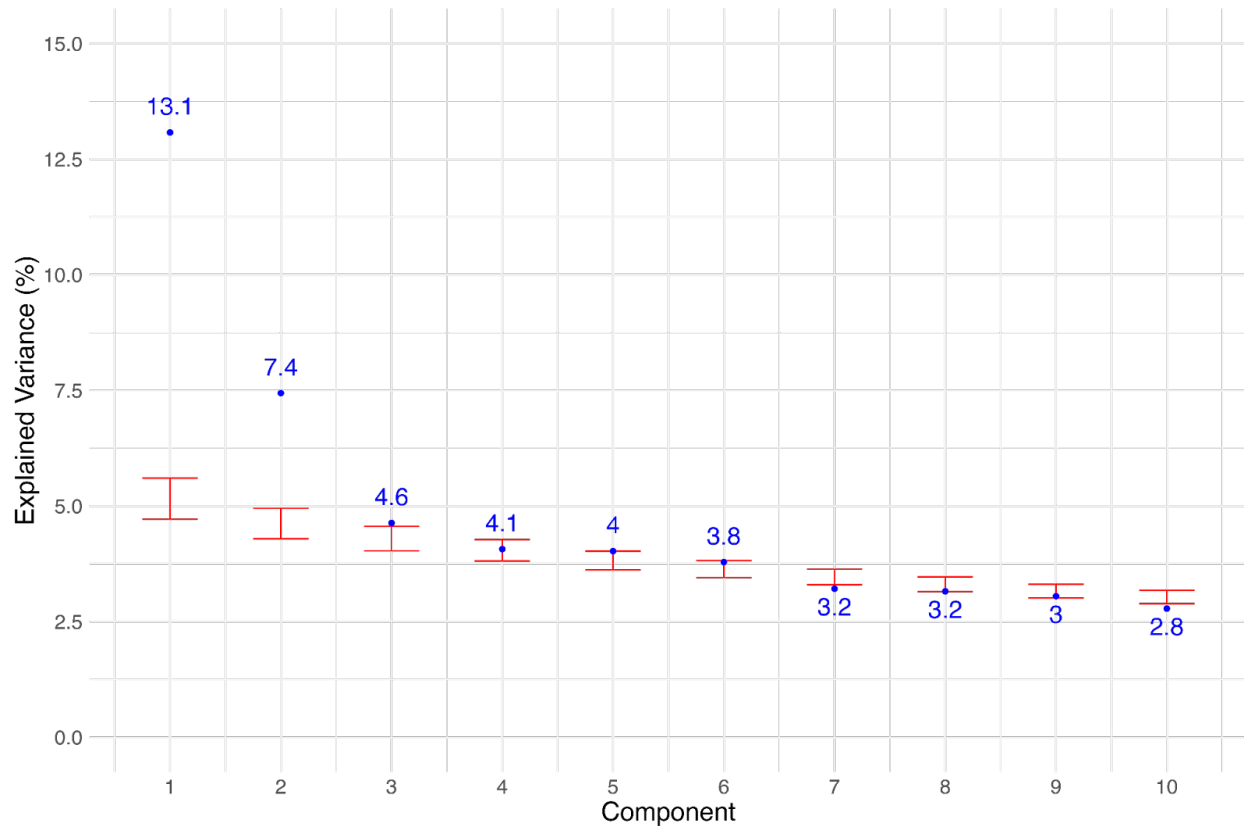
The previous analyses showed that chords vary in the consistency of their pleasantness judgments. We now ask a complementary question: how do listeners vary in their individual preferences for chords? To answer this question, we use Q-type principal component analysis (Q-PCA) to find a low-dimensional representation of listener individual differences.

We first constructed a matrix in which each row corresponds to a participant, and each column to a chord, with each cell containing one participant's average rating of one chord. Conducting PCA on this matrix recovers linear combinations of columns which explain the greatest degree of between-row variance, i.e., dimensions which capture the ways in which participants varied most in their ratings. In comparison to the SEM approach above, in which one model was fitted for each chord, Q-PCA involves only a single model and uses averaged ratings, reducing the degree of noise present in the analysis. Regardless, we also expect to see a reasonable degree of correspondence between chords with strong influence on the resulting principal components and chords which showed high degrees of between-participant variation in the SEM analysis.

Scree plotting of the resulting components' explanatory power (Figure 6, blue points) does not reveal a particularly pronounced elbow, so we used permutation testing (10,000 replicates) to test the statistical significance of the components (Figure 6, red error bars). According to this method, the first three components are statistically significant at the .05 level; however, the explanatory power of the third component is small, and only exceeds the upper 95% range of the permuted data by less than 0.1%, so we decided to restrict our interpretation to the first two components.

Figure 6

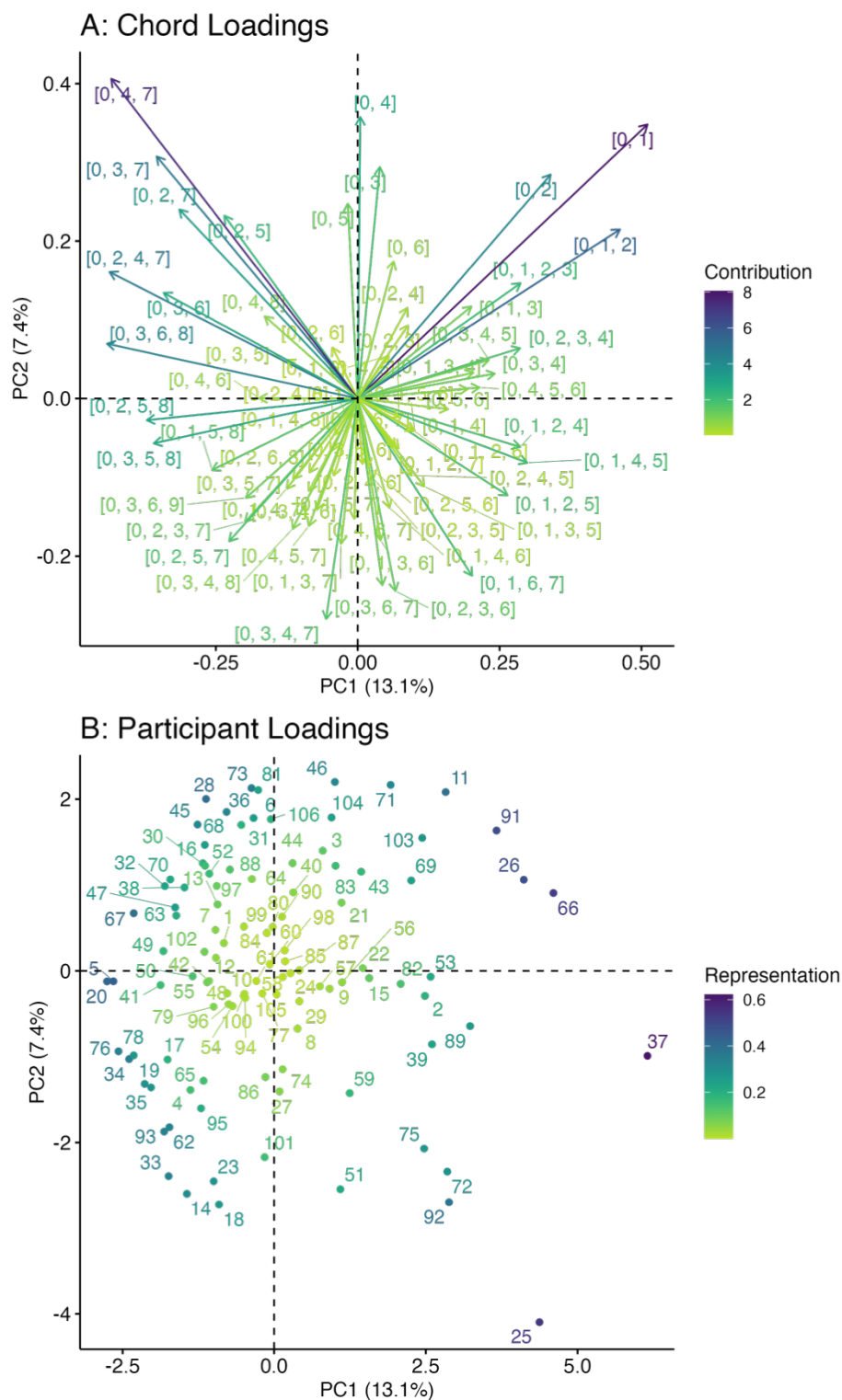
Explained Variances (Blue) and 95% Permutation Ranges (Red, 10,000 Replicates) of the First Ten Principal Components



This PCA thereby scores participants on two dimensions of aesthetic preferences, accounting for 20.5% of total variance in participant judgments. This is a strong performance when considered in the context of the participant-level test-retest correlations (mean $r = .53$, $R^2 = .28$); most of the remaining variance presumably corresponds to response noise. To assess the reliability of the two PCA dimensions, we computed two sets of PCA scores for each participant, one from their first set of ratings and one from their second set of ratings, and then correlated them. The resulting correlations are .83 (PC1) and .74 (PC2), indicating that these scores are relatively reliable. In what follows, however, we analyse PCA scores computed from the averages of those two sets of ratings.

Figure 7

Biplots of (A) Chords' and (B) Individuals' Loadings on the First Two Principal Components (PCs)

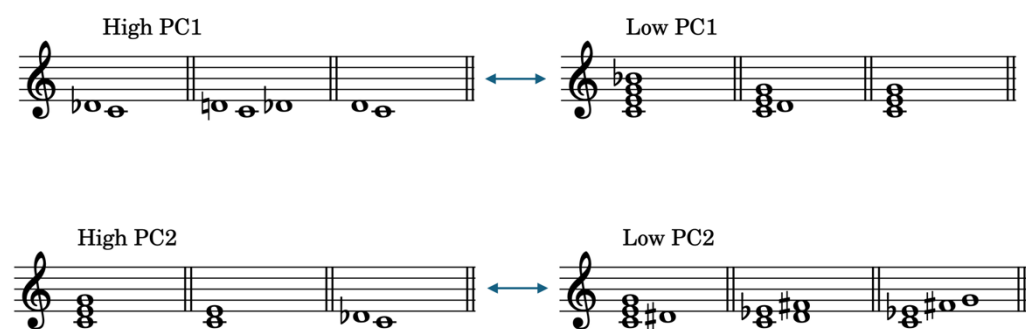


Note. Numbers in brackets are the percentage of total variance explained by each PC.

PCA scores for chords and participants are plotted in Figure 7. The three highest-scoring chords on PC1 are [0, 1], [0, 1, 2], and [0, 2], all of which contain exclusively some combination of semitones and whole tones, i.e., intervals causing a high degree of interference. The three lowest-scoring chords on this dimension are [0, 3, 6, 8], [0, 2, 4, 7], and [0, 4, 7]. While the first and second of these do contain whole tone intervals, the chords are dominated by major and minor thirds, which cause less beating. The three highest-scoring chords on PC2 are [0, 4, 7], [0, 4], and [0, 1], and the three lowest-scoring are [0, 3, 4, 7], [0, 2, 3, 6], and [0, 3, 6, 7]. No intervallic differences stand out here, but it can be noted that the highest-scoring chords are all of cardinality 2 or 3, while the lowest-scoring chords are all of cardinality 4. Notated versions of these chords are presented in Figure 8.

Figure 8

Notated Versions of Chords With Most Extreme Loadings on PC1 and PC2



Note. Inversions are arbitrary in this figure, as Shepard tones do not produce a percept of inversion in chords.

We systematically assessed the relationship between chord features and their loadings on the two PCs in two complementary ways. First, we correlated chord features with chords' coordinates on each principal component. Correlations are given in Table 5 below. PC1 coordinates correlate most highly with interference and semitone count (both $r = .67$), while PC2 coordinates are most highly correlated with cardinality ($r = -.78$).

Table 5*Correlations Between Chord Features and Coordinates on Principal Components (PCs)*

	PC1 coordinates	PC2 coordinates
Interference	.67	-.29
Harmonicity	-.23	.39
Familiarity	-.39	.36
Cardinality	-.13	-.78
Major-scale compatibility	-.32	.52
Major-scale compatibility (binary)	-.37	.31
Interval 1 count	.67	-.30
Interval 2 count	.03	-.18
Interval 3 count	-.30	-.33
Interval 4 count	-.18	-.31
Interval 5 count	-.37	-.35
Interval 6 count	-.10	-.39
Scale degree 1 presence	.38	-.09
Scale degree 2 presence	-.07	-.03
Scale degree 3 presence	-.14	-.09
Scale degree 4 presence	-.21	-.08
Scale degree 5 presence	-.13	-.19
Scale degree 6 presence	-.06	-.23
Scale degree 7 presence	-.32	-.31
Scale degree 8 presence	.08	-.18
Scale degree 9 presence	-.11	-.10
Scale degree 10 presence	.13	-.05
Scale degree 11 presence	.29	.01

Second, we constructed two regression models, taking the chords' loadings on PC1 and PC2 as the respective outputs. Interval counts and scale degrees were excluded from this analysis, as a model with all predictors suffers from high multicollinearity. Parameters of these models are presented in Tables 6 and 7.

Table 6

Parameters of Multiple Regression Model Predicting Chords' Loadings on PC1 From Psychoacoustic Feature Values

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE	<i>t</i>	<i>p</i>
Intercept	N/A	N/A	N/A	0.326	[-0.058, 0.713]	0.182	1.794	.078
Interference	0.820	[0.701, 0.995]	0.070	0.838	[0.681, 1.02]	0.071	11.782	< .0001
Harmonicity	0.106	[-0.038, 0.309]	0.083	0.327	[-0.124, 0.888]	0.256	1.279	.206
Familiarity	-0.195	[-0.280, -0.083]	0.060	-0.638	[-2.56 -0.394]	0.194	-3.284	.002
Cardinality	-0.660	[-0.906, -0.477]	0.070	-0.213	[-0.267, -0.166]	0.023	-9.376	< .0001
Major-scale compatibility (binary)	-0.127	[-0.280, 0.040]	0.083	-0.055	[-0.123, 0.016]	0.036	-1.530	.131
Major-scale compatibility	-0.217	[-0.451, 0.006]	0.103	-3.903	[-8.21, 0.206]	1.845	-2.116	.039

Note. Adjusted $R^2 = .7899$ (95% CI: [.685, .835]). 95% confidence intervals were calculated using bootstrapping (10,000 replicates).

Table 7

Parameters of Multiple Regression Model Predicting Chords' Loadings on PC2 From Psychoacoustic Feature Values

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE	<i>t</i>	<i>p</i>
Intercept	N/A	N/A	N/A	0.793	[0.483, 1.06]	0.158	5.033	< .0001
Interference	0.218	[0.032, 0.422]	0.080	0.168	[0.019, 0.281]	0.062	2.729	.008
Harmonicity	-0.136	[-0.314, 0.035]	0.095	-0.318	[-0.715, 0.093]	0.222	-1.434	.157
Familiarity	0.337	[0.119, 0.522]	0.068	0.829	[0.550, 2.88]	0.169	4.920	< .0001
Cardinality	-0.801	[-0.989, -0.577]	0.081	-0.195	[-0.231, -0.144]	0.020	-9.900	< .0001
Major-scale compatibility (binary)	-0.081	[-0.235, 0.094]	0.095	-0.026	[-0.073, 0.031]	0.031	-0.853	.397
Major-scale compatibility	0.276	[0.037, 0.527]	0.118	3.734	[0.623, 7.21]	1.600	2.335	.023

Note. Adjusted $R^2 = .7224$ (95% CI: [.535, .792]). 95% confidence intervals were calculated using bootstrapping (10,000 replicates).

The parametric versions of both models indicate that interference, familiarity, cardinality, and major-scale compatibility are all significant predictors. Both non-parametric bootstrapping and robust regression (see Supplementary Material) give comparable results for the first three predictors; both methods exclude major-scale compatibility in the PC1 model, and it is partially excluded by bootstrapping and excluded by robust regression in the PC2 model. However, the beta coefficients for interference and cardinality in the PC1 model, and for cardinality in the PC2 model, are much higher than those of the other significant predictors.

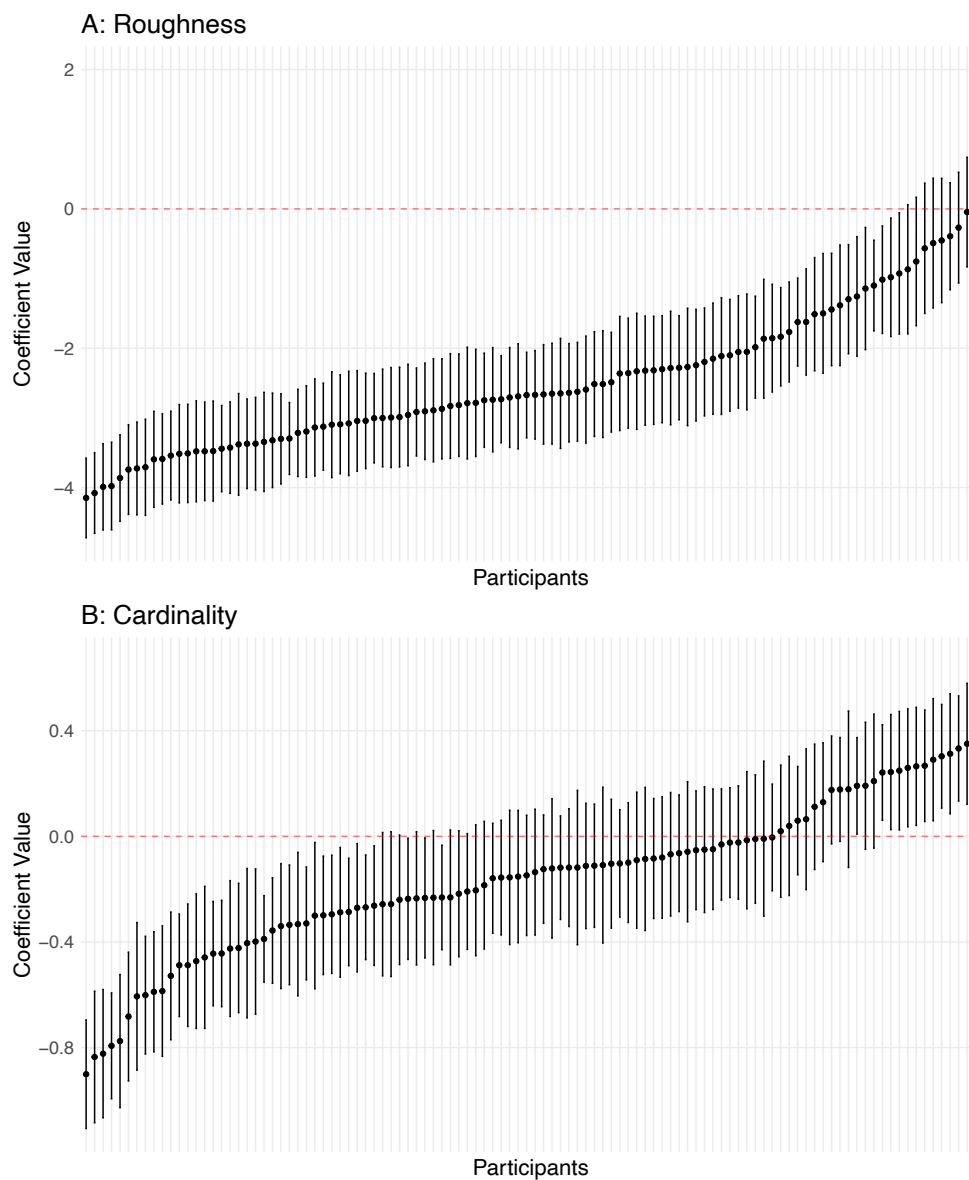
Taking this in conjunction with the correlations in Table 5 above, we make the following suggestions. PC1 is most closely related to interference: interference has the highest beta coefficient in the regression model, and joint-highest correlation value. The other joint-highest correlation is with semitone count, which may be taken as an alternative index of interference, as semitones are the interval with highest interference within any chord, an effect that is particularly pronounced in Shepard-tone chords. Interestingly, PC1 is also negatively predicted by cardinality in the regression model, despite having a minimal pairwise correlation with cardinality. This could mean that PC1 reflects both a lesser disliking for interference and a greater disliking for cardinality. However, it could also be an artifact of the interference modelling process. In particular, the Hutchinson & Knopoff model calculates perceived “roughness”, and includes a step where this perceived roughness is discounted for chords containing more notes. If this discounting process is insufficient, and a greater discount would more accurately capture perceptual judgments, then we would expect to see this negative regression coefficient for cardinality, which would augment the model’s own incomplete cardinality discount. There is also some evidence that familiarity makes an independent contribution to PC1: it has a small but significant beta coefficient in the parametric regression (Table 7), and a large beta coefficient in robust regression (see Supplementary Materials). This could be explained by the presence of an underlying factor that drives both effects, for example a familiarity with Western music (in which chords tend to be smooth). However, given the disagreement between models regarding familiarity and the tentative nature of our explanation of cardinality’s effect, we focus on interpreting PC1 in terms of interference, which has the strongest effect in all models and correlations, in what follows. Meanwhile, PC2 seems to be more straightforwardly related to cardinality, which possesses both the greatest correlation and the highest beta coefficient by some margin.

To obtain a complementary perspective on interference and cardinality judgments, we fit a multiple regression model to each participant’s dataset (averaging first across repeated stimuli), and studied the resulting distributions of regression coefficients (Figure 9). As expected, the vast majority of participants have a statistically significant negative response to interference. In contrast, cardinality

displays a more even spread; many participants do not have a statistically significant response to cardinality, though quite a few have a reliable negative response, and a few have a reliable positive response.

Figure 9

Coefficient Values and Asymptotic 95% Confidence Intervals of Multiple Regressions Conducted per Participant, With Interference and Cardinality Predicting Ratings



Note. Participants are ordered by coefficient values separately within each panel.

Musical experience impacts aesthetic responses to interference

To assess whether systematic differences in listeners' backgrounds are related to their rating strategies, we correlated the background variables gathered in the questionnaire with participants' coordinates in PC-space. We also created two data-derived indices for use as participant features. The first is taste-typicality (Chen *et al.*, 2022), which captures how closely each individual participant's ratings accord with the population average. This is calculated as the Pearson correlation between a participant's ratings over all chords and the mean scores over all chords excluding this participant. The second is reliability, calculated as the Pearson correlation between a participant's ratings over all chords in the first and second phases of the experiment. The resulting correlations between these features and participant coordinates per PC are given in Table 8.

The largest correlations on PC1 are with taste-typicality and reliability. Both correlations are negative, indicating that participants with more typical tastes and more reliability in their responses tend to give higher ratings to chords falling toward the negative pole of PC1, i.e., smoother chords. Further significant correlations with PC1 are observed for several related metrics of musical experience and ability: self-reported musical training and the GMSI measures of Musical Training, Perceptual Abilities, and Singing Abilities all correlate negatively with PC1, albeit more weakly than taste-typicality and reliability. Taste-typicality and reliability are also strongly positively correlated ($r = .77, p < .0001$), and the abovementioned GMSI metrics of musical experience and ability are significantly positively correlated with both taste-typicality and reliability, again more weakly ($.20 \leq r \leq .25$, all $p < .05$). Gender also correlates significantly with PC1; however, this may simply be a result of its small but significant correlation with Musical Training ($r = .24, p = .017$); i.e., in our sample, participants who identified as women tend to have a higher level of GMSI-measured musical training, which may be responsible for their placement towards the negative pole of PC1. Some categories of film genre viewing also show significant correlations with PC1, but these do not form an immediately apparent pattern and are not considered further here.

We interpret this as follows. The archetypical response profile, as seen in our population-level modelling (see Table 4), involves an aversion to interference as the prime determinant of ratings. Participants who embody this profile most explicitly, as reflected in their taste-typicality scores, naturally show this trend in enhanced form through their placement along PC1, reacting negatively to interference and positively to smoothness to a degree greater than that shown by participants falling further from the norm. Participants who do not respond to roughness (or respond positively, as seen in only one participant) are therefore atypical; this is supported by the fact that all such participants (whose upper confidence intervals for roughness in Figure 9A meet or exceed zero) fall well below the first quartile of taste-typicality scores. Taste-typical participants also tend to show greater consistency in their ratings, and tend to have higher levels of music experience/ability.

Regarding PC2, the only significant correlations are with taste-typicality and reliability. These are both positive, suggesting that participants with higher scores on these measures tend to give higher ratings to chords with fewer notes. Cardinality was not a significant predictor of population-level ratings in a multiple regression context (Table 4), but does have a significant negative correlation with these ratings ($r = -.55, p < .0001$). It therefore makes sense that greater taste-typicality also reflects preference for lower cardinality, but to a lesser degree than seen for interference; both the explanatory power of PC2 and the correlation between taste-typicality and this component are lower than for PC1.

Table 8

Correlations Between Participant Background Variables and Principal Component Coordinates

	PC 1	PC 2
Data-derived metrics		
Taste-typicality	-.82 ($p < .0001$)	.49 ($p < .0001$)
Reliability	-.72 ($p < .0001$)	.40 ($p < .0001$)
Self-report metrics		
Age	.01 ($p = .901$)	.10 ($p = .291$)
Gender	-.29 ($p = .004$)	-.10 ($p = .316$)

Language	.01 ($p = .888$)	.16 ($p = .097$)
Formal music training	-.27 ($p = .005$)	.00 ($p = .967$)
Handedness	.00 ($p = .994$)	.02 ($p = .844$)
GMSI summary metrics		
Music Training	-.27 ($p = .005$)	.06 ($p = .566$)
Perceptual Abilities	-.26 ($p = .008$)	.01 ($p = .933$)
Singing Abilities	-.24 ($p = .014$)	.02 ($p = .843$)
Emotion	-.05 ($p = .631$)	-.07 ($p = .507$)
Music genre listening		
Popular	.15 ($p = .134$)	.06 ($p = .546$)
Rock	-.17 ($p = .074$)	-.09 ($p = .335$)
Hip-hop/Rap	-.03 ($p = .725$)	-.11 ($p = .268$)
Metal/Hard rock	-.13 ($p = .201$)	-.03 ($p = .797$)
Jazz/Blues	-.09 ($p = .384$)	-.09 ($p = .36$)
Folk/Ethnic/World	-.08 ($p = .407$)	-.09 ($p = .339$)
Electronic	.00 ($p = .985$)	.09 ($p = .36$)
Classical	-.08 ($p = .432$)	.11 ($p = .268$)
Soundtrack	-.08 ($p = .402$)	.00 ($p = .981$)
Film genre viewing		
Horror	.03 ($p = .757$)	-.08 ($p = .428$)
Thriller	.04 ($p = .705$)	.02 ($p = .833$)
Romance/Drama	.08 ($p = .414$)	-.03 ($p = .772$)
Action	-.25 ($p = .01$)	-.02 ($p = .878$)
Western animated	-.23 ($p = .02$)	.04 ($p = .706$)
Eastern animated	-.27 ($p = .005$)	.04 ($p = .685$)
Musical	-.02 ($p = .875$)	.06 ($p = .545$)
Science fiction	-.28 ($p = .003$)	-.07 ($p = .458$)
Fantasy	-.16 ($p = .098$)	-.07 ($p = .475$)

Comedy	.02 ($p = .862$)	-.03 ($p = .754$)
--------	--------------------	---------------------

Discussion

We used a set of variance decomposition techniques to address various questions about individual differences in chord preferences. We found that chords with high corpus prevalence and low roughness elicited the greatest disagreement in pleasantness judgments; SEM showed that this effect was driven by variance in latent preferences rather than variance in response noise. Q-PCA then showed that listener preferences are captured well by a two-component structure. The first component corresponds to a heightened aversion to interference, and is positively associated with musical expertise. The second component corresponds to an aversion to cardinality (i.e., chords containing more notes), and does not correlate with measured background variables. Both aversions correlate positively with participant-level reliability and taste-typicality.

Why might chords with high corpus prevalence and low roughness elicit the greatest disagreement in pleasantness judgments? Chords with high corpus prevalence will be heard more by listeners, giving them a greater potential to engender individual aesthetic associations. This effect would also explain why chords with high interference elicit less disagreement: chords with high interference tend to have low corpus prevalence (Harrison & Pearce, 2020) and hence have little opportunity to engender learned associations.

Previous work has suggested that listener chord preferences vary along a single dimension that is strongly associated with familiarity with Western music (McDermott *et al.*, 2010; Arthurs *et al.*, 2018; Popescu *et al.*, 2019; Smit *et al.*, 2019; Lahdelma & Eerola, 2020). According to these studies, experience with Western music causes listeners to display stronger preferences for music-theoretically consonant over dissonant chords. Our two-component model of listener preferences replicates and extends this finding; in particular, our first component is reliably associated with musical training, and strongly correlated with taste-typicality. In other words, our more musically trained participants produced judgments that better resemble the population's average pleasantness judgments, which themselves

correspond closely to the music-theoretic consonance hierarchy. Our results differ, however, in the psychoacoustic interpretation of this component. McDermott *et al.* (2010, 2016) argue that these differences in consonance preferences are driven by varying preferences for harmonicity. In contrast, we find that our component is much better explained by interference.

Why might these two approaches have yielded such contrasting results? Since the two methodologies are so different, there are many possible explanations. One explanation might be the different stimulus sets used; our experiment exhaustively samples the set of available 2-, 3-, and 4-note chords, whereas McDermott *et al.*'s paradigm uses a much smaller set of chords that all occur frequently in Western music (even if some occur more than others). Perhaps individual differences manifest differently in these two contexts. Alternatively, the discrepancy might come down to how McDermott *et al.* operationalise roughness and harmonicity preferences. For example, McDermott *et al.* operationalise roughness preferences by comparing preferences for dichotic (different ears) versus diotic (same ears) presentations of beating dyads. The rationale is that roughness should primarily occur in the diotic condition, where the two tones can acoustically interfere; if a participant strongly dislikes the diotic presentation versus the dichotic presentation, the participant is considered to strongly dislike roughness. This operationalisation is clever, but somewhat indirect, and it is possible that other individual differences are at play here, such as preferences for binaural beats. Ideally the measure should be validated against other plausible roughness preference measures, for example having listeners evaluate amplitude-modulated sinusoids. There are clearly many avenues for future work here.

The claim that aversion to interference is a component of individual differences in aesthetic responses to chords is not entirely novel; Song *et al.* (2022, 2023) found evidence that listeners' degree of aversion to interference was related to their degree of preference for harmony, with the latter being suggested as a cross-modal axis of individual difference. Our work extends this finding by providing further evidence not only that levels of aversion to interference differ meaningfully between individuals, but that this difference accounts for the largest proportion of systematic between-listener variation in chord ratings.

It is worth noting that, while our results indicate that response to interference in chords is the primary axis of systematic individual differences in our participants, this does not mean that some listeners are likely to respond *positively* to increasing levels of interference. The regression coefficients in Figure 9A show that only one of our 106 participants exhibits a reliable (albeit weak) preference for increasing interference, while all other participants show either reliable aversion or no response to increasing interference. This finding is in line with the results of Milne *et al.* (2023), who found that increasing interference was reliably associated with increased ratings of instability across different cultural groups. Their results also indicated a stronger link between interference and instability in musicians than non-musicians, matching our findings of a reliable correlation between levels of musical expertise/ability and aversion to interference as embodied in our first principal component.

Our second component to chord preferences corresponds to aversion to cardinality. Interestingly, this component does not correlate with any of our participant background variables, making its origins harder to explain. While cardinality has been manipulated in a few previous studies (e.g., Lahdelma *et al.*, 2022a; McDermott *et al.*, 2010), none have found cardinality to contribute significantly to population-level preferences, and none have tested for individual differences in such preferences. However, cardinality may be interpreted as a form of complexity, and individual differences in complexity preferences have been identified both in the musical domain (Clemente *et al.*, 2022; Güçlütürk & van Lier, 2019) and the visual domain (Corradi *et al.*, 2020; Güçlütürk *et al.*, 2016). Future studies could investigate this further by assessing relationships between preferences for chord cardinality and other forms of musical complexity, and by capturing a wider range of participant background variables.

The two data-derived participant-level metrics included in our analyses, taste-typicality and reliability, were both quite strongly correlated with both components of our PCA. Taste-typicality has recently been proposed as a major correlate of between-participant differences in aesthetic preferences (Chen *et al.*, 2022), though the degree to which this is the case may differ between cultural groups (Roth *et al.*, 2025). Intra-participant reliability, on the other hand, is seldom treated in itself as a measure of relevance to individual differences. Our results support taste-typicality as a major correlate of individual

differences. However, we believe this is best seen in the context of a combination of factors, wherein participants who are more typical also tend to be more reliable and have a higher level of musical training or ability. This may be seen as a sort of “cultural task expertise” effect, with certain individuals holding pronounced and stable embodiments of the normative response profile, possibly engendered through strong enculturation in the form of explicit training. The relationship between taste-typicality, reliability, and specialised training deserves further study.

Our results also contribute to a long series of studies of normative, population-level chord preferences (e.g., Smit *et al.*, 2019; Harrison & Pearce, 2020; Eerola & Lahdelma, 2021; Friedman *et al.*, 2021). In particular, we compared the predictive power of a variety of feature sets, and found in favour of a model with interference, familiarity, and major-scale compatibility as significant predictors (Classic +). Interestingly, harmonicity is a strong predictor when major-scale compatibility is not present (see Supplementary Materials), but major-scale compatibility subsumes harmonicity’s predictive power once added. These two measures are strongly positively correlated in our dataset ($r = .72$). This is consistent with the suggestion of Lahdelma *et al.* (2022b) that harmonicity may act as a proxy for familiarity with the statistical regularities of tonal music. It suggests that listeners are not necessarily judging chords for their levels of harmonicity *per se*, but are simply recognising their level of compatibility with the highly privileged (in a Western tonal context; Johnson-Laird *et al.*, 2012) major scale. This possibility deserves further investigation.

However, several essential limitations should be acknowledged of our experimental data. First, regarding sampling, we only tested Western listeners (specifically UK and US), so our findings cannot be considered globally representative. Moreover, because we wanted our variance analyses to accurately reflect this general Western population, we did not select for musical training or other potentially musically relevant covariates (e.g., genre preferences). This limits our statistical power for testing the relationship between such covariates and our retrieved axes of individual differences (though note that the correlation with musical training was nonetheless statistically significant). Future work could test these associations more systematically by recruiting participants with more variation in these covariates.

Regarding experimental procedure, we employed just a single timbre (Shepard tones), and while we did find high consistency with previous research at the population level, we cannot be sure that individual differences might not manifest differently with other timbres. Additionally, our experimental paradigm is unlike naturalistic music listening in several respects (the use of Shepard tones, providing conscious scoring of music events, listening to chords in isolation), and hence our findings may not neatly generalise to broader musical contexts.

Our analysis methods also have important limitations. We rely heavily on computational models, in particular of interference and harmonicity; while these models have been substantially validated in previous research (e.g., Harrison & Pearce, 2020; Eerola & Lahdelma, 2021), they will inevitably be imperfect approximations to the underlying psychoacoustic processes. Moreover, our variance decomposition analyses (SEM, Q-PCA) make various assumptions in order to make the modelling tractable, such as considering test and retest ratings to be exchangeable, and considering participant preferences to be linear in the space of chords. Other analysis methods make different trade-offs between interpretability and predictive power, for example, collaborative filtering (Celikors & Field, 2025). It would be interesting to collect larger datasets with which to compare and possibly combine these methods.

Nonetheless, we think that this analysis approach has great potential to deliver insights about individual differences in many aesthetic domains, inside and outside of music. The application of SEM to disentangle latent preference variation from response noise is, to our knowledge, novel; it would be interesting to apply this analysis to other aesthetic domains and see whether there are scenarios where response noise, not just latent preference variation, can be reliably explained by measurable stimulus properties. The application of Q-PCA and analogous methods to summarise individual preferences has multiple precedents in empirical aesthetics research (McManus, 1980; McManus *et al.*, 2010; Spehar *et al.*, 2016; Bignardi *et al.*, 2024), but has not yet been applied in music aesthetics research; fruitful insights could be obtained by applying it to other musical domains, for example melody and timbre.

One practical application of these techniques could be to develop efficient tests of individual preferences. One could use SEM to identify a subset of stimuli that maximise latent preference variance while minimising response noise; one could also use Q-PCA to identify stimuli that have maximal loadings on the principal components of interest. More generally, such research could benefit from the rich psychometric literature on survey design, which often works by taking a large set of candidate items and using statistical methods to identify optimal subsets that can form a subsequent short-form survey.

These techniques also have clear practical applications in personalised recommendation. For example, personalised music recommendation is central to many current music playback platforms (e.g., Spotify, Apple Music), and improved understanding of individual aesthetic taste is key to increasing the accuracy of these algorithms at a person-to-person level. Similarly, this approach may increase the ability to tailor music-therapeutic interventions to individual patients, as the individual's unique reception style can be taken into consideration in selecting materials intended to support a desired effect. These applications provide a clear motivation to scale up the present research to study the full range of musical parameters that underpin naturalistic music listening.

Data and code availability

Raw data in anonymised form are available as .csv files on OSF via this link:

<https://osf.io/6g5aq/>

Experiment code is available on GitHub via this link:

<https://github.com/joshfrank95/2024-chord-pleasantness-experiment-PsyNet-code>

Analysis code and key data are available on GitHub via this link:

<https://github.com/joshfrank95/2025-modelling-individual-differences-in-chord-pleasantness-analysis>

References

- Arthurs, Y., Beeston, A. V., & Timmers, R. (2018). Perception of isolated chords: Examining frequency of occurrence, instrumental timbre, acoustic descriptors and musical training. *Psychology of Music*, 46(5): 662–681. <https://doi.org/10.1177/0305735617720834>
- Bignardi, G., Smit, D. J. A., Vessel, E. A., Trupp, M. D., Ticino, L. F., Fisher, S. E., & Polderman, T. J. C. (2024). Genetic effects on variability in visual aesthetic evaluations are partially shared across visual domains. *Communications Biology*, 7: 55. <https://doi.org/10.1038/s42003-023-05710-4>
- Bowling, D. L., Purves, D., & Gill, K. Z. (2018). Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences*, 115(1): 216–221. <https://doi.org/10.1073/pnas.1713206115>
- Cazden, N. (1962). Sensory Theories of Musical Consonance. *The Journal of Aesthetics and Art Criticism*, 20(3): 301–319. <https://doi.org/10.2307/427328>
- Celikors, E., & Field, D. J. (2025). Beauty is in the eye of your cohort: Structured individual differences allow predictions of individualized aesthetic ratings of images. *Cognition*, 256: 106036. <https://doi.org/10.1016/j.cognition.2024.106036>
- Chen, Y., Chang, A., Rosenberg, M. D., Feng, D., Scholl, B. J., & Trainor, L. J. (2022). “Taste typicality” is a foundational and multi-modal dimension of ordinary aesthetic experience. *Current Biology*, 32(8): 1837–1842. <https://doi.org/10.1016/j.cub.2022.02.039>
- Clemente, A., Pearce, M. T., & Nadal, M. (2022). Musical Aesthetic Sensitivity. *Psychology of Aesthetics, Creativity, and the Arts*, 16(1): 58–73. <https://doi.org/10.1037/aca0000381>
- Corradi, G., Chuquichambi, E. G., Barrada, J. R., Clemente, A., & Nadal, M. (2020). A new conception of visual aesthetic sensitivity. *British Journal of Psychology*, 111: 630–658. <https://doi.org/10.1111/bjop.12427>
- Eerola, T., & Lahdelma, I. (2021). The Anatomy of Consonance/ Dissonance: Evaluating Acoustic and Cultural Predictors Across Multiple Datasets with Chords. *Music & Science*, 4: 1–19. <https://doi.org/10.1177/20592043211030471>

- Friedman, R. S., Kowalewski, D. A., Vuvan, D. T., & Neill, W. T. (2021). Consonance Preferences Within an Unconventional Tuning System. *Music Perception*, 38(3): 313–330.
<https://doi.org/10.1525/mp.2021.38.3.313>
- Güçlütürk, Y., Jacobs, R. H. A. H., & van Lier, R. (2016). Liking versus Complexity: Decomposing the Inverted U-curve. *Frontiers in Human Neuroscience*, 10: 112.
<https://doi.org/10.3389/fnhum.2016.00112>
- Güçlütürk, Y., & van Lier, R. (2019). Decomposing Complexity Preferences for Music. *Frontiers in Psychology*, 10: 674. <https://doi.org/10.3389/fpsyg.2019.00674>
- Harrison, P. M. C., & Pearce, M. T. (2018). An Energy-Based Generative Sequence Model For Testing Sensory Theories of Western Harmony. *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 160–167. <https://doi.org/10.48550/arXiv.1807.00790>
- Harrison, P. M. C., & Pearce, M. T. (2020). Simultaneous Consonance in Music Perception and Composition. *Psychological Review*, 127(2): 216–244. <https://doi.org/10.1037/rev0000169>
- Harrison, P. M. C., & Pearce, M. T. Representing Harmony in Computation Music Cognition. PsyArXiv Preprint. <https://doi.org/10.31234/osf.io/xswp4>
- von Helmholtz, H. L. F. (1954). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. A. J. Ellis (trans.). Mineola, NY: Dover.
- Hutchinson, W., & Knopoff, L. (1978). The acoustic component of western consonance. *Journal of New Music Research*, 7(1): 1–29. <https://doi.org/10.1080/09298217808570246>
- Johnson-Laird, P. N., Kang, O. E., & Leong, Y. C. (2012). On Musical Dissonance. *Music Perception*, 30(1): 19–35. <https://doi.org/10.1525/mp.2012.30.1.19>
- Lahdelma, I., Athanasopoulos, G., & Eerola, T. (2021). Sweetness is in the ear of the beholder: chord preference across United Kingdom and Pakistani listeners. *Annals of the New York Academy of Sciences*, 1502: 72–84. <https://doi.org/10.1111/nyas.14655>

- Lahdelma, I., & Eerola, T. (2020). Cultural familiarity and musical expertise impact the pleasantness of consonance/dissonance but not its perceived tension. *Scientific Reports*, 10: 8693. <https://doi.org/10.1038/s41598-020-65615-8>
- Lahdelma, I., Armitage, J., & Eerola, T. (2022a). Affective priming with musical chords is influenced by pitch numerosity. *Musicae Scientiae*, 26(1): 208–217. <https://doi.org/10.1177/1029864920911127>
- Lahdelma, I., Eerola, T., & Armitage, J. (2022b). Is Harmonicity a Misnomer for Cultural Familiarity in Consonance Preferences? *Frontiers in Psychology*, 13: 802385. <https://doi.org/10.3389/fpsyg.2022.802385>
- McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2010). Individual Differences Reveal the Basis of Consonance. *Current Biology*, 20(11): 1035–1041. <https://doi.org/10.1016/j.cub.2010.04.019>
- McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature*, 535(7613): 547–550. <https://doi.org/10.1038/nature18635>
- McManus, I. C. (1980). The aesthetics of simple figures. *British Journal of Psychology*, 71: 505–524. <https://doi.org/10.1111/j.2044-8295.1980.tb01763.x>
- McManus, I. C., Cook, R., & Hunt, A. (2010). Beyond the Golden Section and Normative Aesthetics: Why Do Individuals Differ so Much in Their Aesthetic Preferences for Rectangles? *Psychology of Aesthetics, Creativity, and the Arts*, 4(2): 113–126. <https://doi.org/10.1037/a0017316>
- Milne, A. J. (2013). A computational model of the cognition of tonality (PhD thesis). The Open University, Milton Keynes, England.
- Milne, A. J., Smit, E. A., Sarvasy, H. S., & Dean, R. T. (2023). Evidence for a universal association of auditory roughness with musical stability. *PLoS ONE*, 18(9): e0291642. <https://doi.org/10.1371/journal.pone.0291642>

- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE*, 9(2): e89642. <https://doi.org/10.1371/journal.pone.0089642>
- Parncutt, R. (1988). Revision of Terhardt's Psychoacoustical Model of the Root(s) of a Musical Chord. *Music Perception*, 6(1): 65–93. <https://doi.org/10.2307/40285416>
- Parncutt, R., Engel, I., & Radovanovic, L. (2023). Consonance and Dissonance of Simultaneous Trichords in Western Music: A Listening Experiment to Test Models of Harmonicity and Roughness. *Psychological Topics*, 32(1): 13–34. <https://doi.org/10.31820/pt.32.1.2>
- Popescu, T., Neuser, M. P., Neuwirth, N., Bravo, F., Mende, W., Boneh, O., Moss, F. C., & Rohrmeier, M. (2019). The pleasantness of sensory dissonance is mediated by musical style and expertise. *Scientific Reports*, 9: 1070. <https://doi.org/10.1038/s41598-018-35873-8>
- Rahn, J. (1980). *Basic atonal theory*. New York: Schirmer.
- Roth, L. H. O., Kawabata, H., & Leder, H. (2025). Universal, Cultural, or Individual? An Intercultural Assessment of Shared Proportions of Private Taste in Aesthetic Appreciation. *Empirical Studies of the Arts*. <https://doi.org/10.1177/02762374251334962>
- Shepard, R. N. (1964). Circularity in Judgments of Relative Pitch. *The Journal of the Acoustical Society of America*, 36: 2346–2353. <https://doi.org/10.1121/1.1919362>
- Smit, E. A., Milne, A. J., Dean, R. T., & Weidemann, G. (2019). Perception of affect in unfamiliar musical chords. *PLoS ONE*, 14(6): e0218570. <https://doi.org/10.1371/journal.pone.0218570>
- Song, S. E., Kowalewski, D. A., & Friedman, R. S. (2022). Preference for Harmony: A Link Between Aesthetic Responses to Combinations of Colors and Musical Tones. *Psychomusicology: Music, Mind, and Brain*, 32(1-2): 33–45. <https://doi.org/10.1037/pmu0000290>
- Song, S. E., Kowalewski, D. A., & Friedman, R. S. (2023). Preference for Harmony: A Preference for Structural Simplicity, Familiarity, or Both? *Empirical Studies of the Arts*, 43(1): 209–228. <https://doi.org/10.1177/02762374231216033>

Spehar, B., Walker, N., & Taylor, R. P. (2016). Taxonomy of Individual Variations in Aesthetic Responses to Fractal Patterns. *Frontiers in Human Neuroscience*, 10: 350.

<https://doi.org/10.3389/fnhum.2016.00350>

Stumpf, C. (1890). *Tonpsychologie*. Leipzig: Verlag S. Hirzel.

SUPPLEMENTARY MATERIALS

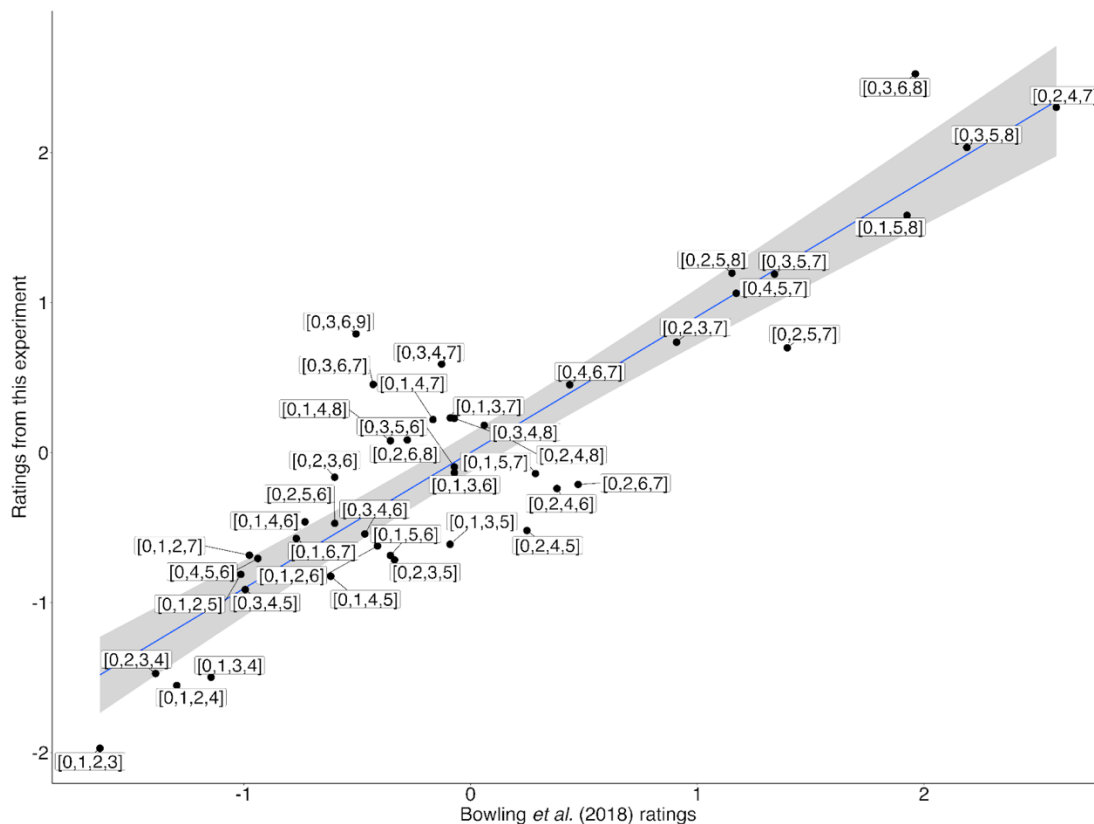
Timbral Effects Explain Systematic Differences Between Experiments

We performed some analyses to investigate the small differences between the ratings in our experiment and the experiment of Bowling *et al.* (2018), which used a harmonic (piano) timbre for stimulus presentation. Bowling *et al.* tested each chord in different inversions, and so we averaged these inversions to produce a dataset comparable to our own; for modelling purposes, we approximated the chords' intervals as being equal-tempered, despite Bowling *et al.* using just intonation.

Figure S1 plots mean ratings from Bowling *et al.* (horizontal axis) against mean ratings from our experiment (vertical axis), as well as a fitted regression line, for tetrads. Points above the line represent chords rated more highly in our study; points below the line represent chords rated less highly.

Figure S1

Scatter Plot of Normalised Average Ratings From Bowling et al. (2018) vs Those From Our Experiment, For Tetrads; a Regression Line With 95% CI is Included



The two studies differ in various design parameters (e.g., instructions, tuning, participant population), but timbre differences are particularly important. Our Shepard tones and Bowling *et al.*'s piano tones differ in several ways, but perhaps the most notable is that Shepard tones do not contain any non-octave harmonics, such as the compound perfect fifth or the compound major third. After Marjeh *et al.* (2024; see also Sethares, 1993 and Plomp & Levelt, 1965), we might expect this to systematically affect pleasantness judgments due to the different levels of roughness and harmonicity across chords resulting from the change in timbre.

We explore this possibility by creating regression models to predict the differences between our ratings and the ratings collected in Bowling *et al.* The procedure below was conducted only for tetrads, due to the low number of data points for dyads and triads. We began by constructing a regression model predicting our ratings from Bowling *et al.*'s ratings, and then extracted the chord-level residuals from this model. A positive residual means that a chord produced higher pleasantness ratings in our experiment than Bowling *et al.*'s experiment. We then constructed regression models to try and predict these residuals.

We considered two possible ways of predicting these residuals: (i) using psychoacoustic features, and (ii) using interval counts. For the psychoacoustic model, we included relative roughness and relative harmonicity as predictors. Relative roughness and harmonicity features were calculated by applying roughness (Hutchinson & Knopoff, 1978) and harmonicity (Milne, 2013) models to both stimulus sets, taking into account the different timbres used, and subtracting the Bowling *et al.* outputs from our own outputs. Amplitudes of partials are known exactly for the stimuli used in our experiment; we approximated the piano timbre as a harmonic spectrum with amplitudes 1, 0.508024615648808, 0.0933536943448577, 0.222457796845064, 0.102246797048138, 0.0606420093604131, 0.104786291832625, 0.0843553191337145, 0.0445420508366155, 0.0307915719139996, derived by performing a Fourier analysis on a reference piano tone. For the interval counts model, we included the six interval features as predictors.

The psychoacoustic model achieved a relatively low adjusted R^2 of .14 (Table S1.1). Only relative roughness achieved significance, with the beta coefficient having the expected sign (negative, i.e., greater roughness leads to lower pleasantness ratings). This result was supported by modelling using robust regression (Table S1.2). The interval model achieved a higher adjusted R^2 of .56, with the minor second, major second, minor third, and tritone achieving significance (Table S2.1), a pattern mirrored by robust regression (Table S2.2).

Table S1.1

Parameters of Regression Model Predicting Residuals Between Our Data and Those of Bowling et al. (2018) for Tetrads, with Differences in Roughness and Harmonicity as Predictors

	Beta	95% CI (Beta)	SE (Beta)	<i>t</i>	<i>p</i>
Intercept	N/A	N/A	N/A	N/A	N/A
Roughness	-0.434	[-0.650; -0.159]	0.147	-2.946	.005
Harmonicity	-0.195	[-0.503; 0.098]	0.147	-1.325	.193

Note. Adjusted R^2 = .1438 (95% CI: [-.026; .330]). 95% confidence intervals were generated with bootstrapping (10,000 replicates).

Table S1.2

Parameters of Robust Regression Model with MM Estimation Predicting Residuals Between Our Data and Those of Bowling et al. (2018) for Tetrads, with Differences in Roughness and Harmonicity as Predictors

	Beta	SE (Beta)	statistic	<i>p</i>
Intercept	N/A	N/A	N/A	N/A
Roughness	-0.404	0.146	-2.76	.008
Harmonicity	-0.090	0.146	-0.619	.546

Table S2.1

Parameters of a Regression Model Predicting Residuals Between Our Ratings and Those of Bowling et al. (2018) for Tetrads, with Interval Counts as Predictors, and Intercept Set at Zero

	Estimate (B)	95% CI (B)	SE	<i>t</i>	<i>p</i>
Int.1/11	-0.047	[-0.096; -0.006]	0.021	-2.260	.030
Int. 2/10	-0.104	[-0.144; -0.060]	0.021	-4.983	< .0001
Int. 3/9	0.077	[0.032; 0.118]	0.019	4.091	< .0005
Int. 4/8	0.040	[-0.014; 0.069]	0.021	1.894	.066
Int. 5/7	-0.024	[-0.065; 0.022]	0.021	-1.154	.256
Int. 6	0.099	[0.042; 0.154]	0.028	3.508	.001

Note. Adjusted $R^2 = .5554$ (95% CI: [.273; .731]). 95% confidence intervals were generated with bootstrapping (10,000 replicates).

Table S2.2

Parameters of a Robust Regression Model with MM Estimation Predicting Residuals Between Our Ratings and Those of Bowling et al. (2018) for Tetrads, with Interval Counts as Predictors, and Intercept Set at Zero

	Estimate (B)	SE	statistic	<i>p</i>
Int.1/11	-0.047	0.023	-2.09	.044
Int. 2/10	-0.104	0.023	-4.62	< .0001
Int. 3/9	0.078	0.020	3.85	.0005
Int. 4/8	0.042	0.023	1.88	.064
Int. 5/7	-0.027	0.023	-1.19	.242
Int. 6	0.097	0.030	3.22	.002

These results suggest that differences in roughness, driven by the different timbral structures of the stimuli, play a systematic role in the small rating differences observed between our study and the study of Bowling *et al.* This is demonstrated directly, albeit weakly, in the psychoacoustic feature model,

in which roughness was a significant predictor of rating differences (residuals). It is also demonstrated indirectly through the patterns observed in the interval model; the absence of non-octave partials in our stimuli reduces the unpleasant beating components of the minor third and tritone, a pattern which may be hinted towards in the major third as well, despite its failure to reach significance as a predictor. The roughness of the minor and major second is primarily caused by the first partials, which are not lost in Shepard tones; these intervals' negative coefficients may be due to the enhancement of the perceptual salience of these partials by the absence of all others. The much higher explanatory power of the interval model may be due to the differences in tuning between our stimuli and those of Bowling *et al.*; degrees of temperament relative to 12-tone equal temperament differ per interval, a feature which a model using intervals as predictors may inadvertently capture alongside features relating to roughness and harmonicity. Accounting for this not only improves model performance, but appears to be a necessity, as indicated by the lower 95% confidence interval of the non-interval model's adjusted R^2 being negative, indicating very poor reliability of the model.

Parameters of All Linear Models Tested For Predicting Average Pleasantness Ratings

All confidence intervals were generated with bootstrapping (10,000 replicates). Robust linear model (RLM) parameters are given in a separate table below the OLS parameters. P -values for RLMs are estimated using `repmo::rob.pvals` in R, which assumes the test statistic follows an F -distribution. All RLMs used MM-estimation. There are instances in which OLS- and RLM-derived p -values and confidence intervals misalign; this is not the case for the psychoacoustic model, on which we focus in the main text. These models are presented for completeness and for the interest of anyone involved in population-level modelling of chord pleasantness. Parameters beginning with “SD” are scale degree numbers.

Full model - OLS

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE (B)	<i>t</i>	<i>p</i>
Intercept	N/A	N/A	N/A	3.08	[-1.33, 6.91]	1.529	2.016	.0498
Roughness	-1.343	[-2.13, -0.691]	0.221	-4.116	[-6.20, -2.02]	0.677	-6.084	< .0001
Harmonicity	-0.181	[-0.610, 0.258]	0.161	-1.687	[-5.63, 2.40]	1.501	-1.124	.267
Cardinality	-0.430	[-1.44, 0.746]	0.419	-0.417	[-1.41, 0.734]	0.407	-1.026	.310
Familiarity	0.132	[0.041, 0.239]	0.037	1.296	[-1.32, 4.22]	0.366	3.542	< .001
Major-scale compatibility	0.197	[-0.076, 0.470]	0.109	10.629	[-4.24, 24.9]	5.847	1.818	.076
Major-scale compatibility (binary)	0.009	[-0.104, 0.163]	0.057	0.012	[-0.132, 0.207]	0.073	0.163	.872
Int. 1/11	0.748	[0.178, 1.39]	0.220	0.623	[0.136, 1.10]	0.183	3.400	.001
Int. 2/10	0.403	[0.009, 0.834]	0.166	0.336	[0.006, 0.679]	0.138	2.434	.019
Int. 3/9	0.235	[-0.182, 0.652]	0.160	0.181	[-0.143, 0.482]	0.123	1.469	.149
Int. 4/8	-0.076	[-0.530, 0.301]	0.162	-0.061	[-0.407, 0.245]	0.131	-0.469	.641
Int. 5/7	0.104	[-0.542, 0.519]	0.188	0.087	[-0.415, 0.437]	0.156	0.556	.581
Int. 6	0.108	[-0.253, 0.477]	0.135	0.118	[-0.263, 0.513]	0.147	0.799	.428
SD 1	-0.215	[-0.394, -0.068]	0.066	-0.327	[-0.595, -0.099]	0.101	-3.227	.002
SD 2	-0.061	[-0.280, 0.134]	0.073	-0.105	[-0.454, 0.230]	0.126	-0.833	.409

SD 3	-0.116	[-0.300, 0.040]	0.066	-0.177	[-0.458, 0.063]	0.101	-1.751	.087
SD 4	0.078	[-0.143, 0.329]	0.085	0.101	[-0.182, 0.418]	0.109	0.926	.360
SD 5	-0.084	[-0.310, 0.033]	0.050	-0.313	[-1.13, 0.121]	0.189	-1.657	.104
SD 6	-0.247	[-0.437, -0.098]	0.071	-0.397	[-0.688, -0.151]	0.115	-3.469	.001
SD 7	0.136	[-0.340, 0.667]	0.167	0.176	[-0.436, 0.833]	0.216	0.816	.419
SD 8	-0.113	[-0.260, -0.022]	0.048	-0.211	[-0.462, -0.027]	0.089	-2.374	.022
SD 9	-0.167	[-0.356, 0.009]	0.069	-0.313	[-0.652, 0.036]	0.129	-2.422	.020
SD 10	-0.166	[-0.404, 0.051]	0.089	-0.228	[-0.544, 0.069]	0.122	-1.869	.068

Note. Adjusted $R^2 = .931$ (95% CI: [.893, 0.933]). Int. = interval count. SD = presence of scale degree. SD 11 is excluded in this model due to aliasing.

Full model - RLM

	Beta	SE (Beta)	B	SE (B)	statistic	p
Intercept	N/A	N/A	3.26	1.59	2.05	.046
Roughness	-1.32	0.229	-4.03	0.703	-5.74	< .0001
Harmonicity	-0.204	0.168	-1.90	1.56	-1.22	.229
Cardinality	-0.455	0.435	-0.442	0.422	-1.05	.298
Familiarity	0.129	0.039	1.26	0.380	3.32	.002
Major-scale compatibility	0.206	0.113	11.1	6.07	1.83	.072
Major-scale compatibility (binary)	0.002	0.059	0.003	0.076	0.040	.968
Int. 1/11	0.724	0.228	0.603	0.190	3.17	.003
Int. 2/10	0.382	0.172	0.318	0.143	2.22	.031
Int. 3/9	0.221	0.166	0.170	0.128	1.33	.189

Int. 4/8	-0.075	0.168	-0.060	0.136	-0.444	.656
Int. 5/7	0.109	0.195	0.091	0.162	0.558	.576
Int. 6	0.113	0.141	0.123	0.153	0.805	.423
SD 1	-0.212	0.069	-0.323	0.105	-3.07	.004
SD 2	-0.051	0.076	-0.088	0.131	-0.676	.505
SD 3	-0.101	0.069	-0.154	0.105	-1.46	.154
SD 4	0.094	0.088	0.122	0.114	1.08	.295
SD 5	-0.080	0.052	-0.298	0.196	-1.52	.144
SD 6	-0.245	0.074	-0.394	0.119	-3.31	.002
SD 7	0.164	0.173	0.213	0.224	0.949	.352
SD 8	-0.112	0.049	-0.208	0.092	-2.26	.029
SD 9	-0.156	0.072	-0.291	0.134	-2.17	.037
SD 10	-0.156	0.093	-0.214	0.127	-1.69	.100

Psychoacoustic model - OLS (included in main text)

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE (B)	<i>t</i>	<i>p</i>
Intercept	N/A	N/A	N/A	0.300	[-0.487, 1.08]	0.379	0.791	.432
Roughness	-0.665	[-0.758, -0.583]	0.048	-2.041	[-2.44, -1.72]	0.148	-13.758	< .0001
Harmonicity	0.088	[-0.015, 0.197]	0.057	0.823	[-0.149, 1.76]	0.534	1.540	.129
Familiarity	0.180	[0.077, 0.284]	0.041	1.766	[1.18, 6.00]	0.405	4.356	< .0001
Cardinality	-0.057	[-0.183, 0.064]	0.049	-0.055	[-0.180, 0.063]	0.047	-1.159	.251
Major-scale compatibility (binary)	-0.029	[-0.130, 0.092]	0.057	-0.037	[-0.169, 0.114]	0.074	-0.502	.617
Major-scale compatibility	0.246	[0.092, 0.402]	0.071	13.241	[4.70, 22.3]	3.847	3.442	.001

Note. Adjusted $R^2 = .8985$ (95% CI: [.845, .924])

Psychoacoustic model - RLM

	Beta	SE (Beta)	B	SE (B)	statistic	p
Intercept	N/A	N/A	0.199	0.384	0.519	.610
Roughness	-0.637	0.049	-1.95	0.150	-13.0	< .0001
Harmonicity	0.103	0.058	0.958	0.541	1.77	.078
Familiarity	0.441	0.042	4.32	0.410	10.5	.003
Cardinality	-0.063	0.050	-0.062	0.048	-1.28	.229
Major-scale compatibility (binary)	-0.040	0.058	-0.051	0.075	-0.681	.492
Major-scale compatibility	0.225	0.072	12.1	3.89	3.12	.003

Classic model - OLS

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE (B)	t	p
Intercept	N/A	N/A	N/A	-0.095	[-0.998, 0.746]	0.399	-0.238	.812
Roughness	-0.693	[-0.785, -0.614]	0.052	-2.124	[-2.55, -1.77]	0.158	-13.420	< .0001
Harmonicity	0.212	[0.111, 0.329]	0.051	1.975	[1.08, 2.96]	0.479	4.124	.0001
Familiarity	0.182	[0.089, 0.262]	0.045	1.784	[1.11, 6.62]	0.444	4.020	.0002
Cardinality	-0.107	[-0.243, 0.018]	0.052	-0.103	[-0.237, 0.017]	0.050	-2.065	.043

Note. Adjusted $R^2 = .8783$ (95% CI: [.818, .909])

Classic model - RLM

	Beta	SE (Beta)	B	SE (B)	statistic	p
Intercept	N/A	N/A	0.269	0.341	0.787	.457

Roughness	-0.610	0.044	-1.87	0.135	-13.8	< .0001
Harmonicity	0.140	0.044	1.30	0.410	3.18	.004
Familiarity	2.51	0.039	24.6	0.380	64.8	< .0001
Cardinality	-0.140	0.044	-0.136	0.043	-3.18	.004

Notation model - OLS

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE (B)	<i>t</i>	<i>p</i>
Intercept	N/A	N/A	N/A	1.76	[0.628, 2.65]	0.352	4.989	< .0001
Int. 1/11	-0.364	[-0.894, 0.137]	0.204	-0.303	[-0.711, 0.110]	0.170	-1.784	.081
Int. 2/10	0.245	[-0.282, 0.755]	0.203	0.204	[-0.229, 0.607]	0.169	1.210	.232
Int. 3/9	0.456	[-0.093, 1.00]	0.214	0.351	[-0.075, 0.732]	0.165	2.131	.038
Int. 4/8	0.108	[-0.471, 0.632]	0.211	0.087	[-0.361, 0.501]	0.170	0.513	.610
Int. 5/7	0.479	[-0.095, 0.971]	0.211	0.399	[-0.078, 0.787]	0.176	2.267	.028
Int. 6	0.316	[-0.097, 0.691]	0.162	0.344	[-0.097, 0.728]	0.176	1.956	.056
SD 1	-0.819	[-1.52, -0.078]	0.291	-1.25	[-2.24, -0.076]	0.443	-2.816	.007
SD 2	-0.658	[-1.36, 0.025]	0.255	-1.13	[-2.17, 0.078]	0.439	-2.576	.013
SD 3	-0.694	[-1.49, 0.078]	0.288	-1.06	[-2.15, 0.138]	0.439	-2.409	.020
SD 4	-0.560	[-1.44, 0.395]	0.331	-0.724	[-1.82, 0.484]	0.428	-1.692	.097
SD 5	-0.372	[-0.881, -0.071]	0.130	-1.39	[-2.57, -0.088]	0.488	-2.850	.006
SD 6	-0.807	[-1.54, -0.057]	0.276	-1.30	[-2.31, -0.061]	0.444	-2.927	.005
SD 7	-0.588	[-1.45, 0.398]	0.334	-0.760	[-1.84, 0.482]	0.432	-1.760	.085
SD 8	-0.542	[-1.25, 0.061]	0.238	-1.01	[-2.07, 0.189]	0.444	-2.276	.027
SD 9	-0.725	[-1.41, -0.146]	0.235	-1.35	[-2.38, -0.184]	0.438	-3.088	.003
SD 10	-0.927	[-1.72, -0.043]	0.315	-1.27	[-2.28, -0.066]	0.432	-2.939	.005
SD 11	-0.459	[-1.17, 0.200]	0.248	-0.819	[-1.90, 0.390]	0.443	-1.849	.070

Note. Adjusted $R^2 = .8506$ (95% CI: [.776, .870]).

Notation model - RLM

	Beta	SE (Beta)	B	SE (B)	statistic	<i>p</i>
Intercept	N/A	N/A	2.00	0.332	6.03	< .0001
Int. 1/11	-0.258	0.192	-0.215	0.160	-1.34	.197
Int. 2/10	0.296	0.191	0.247	0.159	1.55	.138
Int. 3/9	0.544	0.202	0.419	0.155	2.70	.011
Int. 4/8	0.255	0.199	0.205	0.161	1.28	.222
Int. 5/7	0.584	0.199	0.487	0.166	2.93	.006
Int. 6	0.415	0.153	0.452	0.166	2.72	.010
SD 1	-0.978	0.274	-1.49	0.418	-3.57	.001
SD 2	-0.777	0.241	-1.34	0.414	-3.23	.003
SD 3	-0.853	0.272	-1.30	0.415	-3.14	.004
SD 4	-0.764	0.312	-0.988	0.404	-2.45	.023
SD 5	-0.423	0.123	-1.58	0.461	-3.44	.001
SD 6	-0.980	0.260	-1.58	0.419	-3.77	.001
SD 7	-0.786	0.315	-1.02	0.407	-2.50	.021
SD 8	-0.696	0.224	-1.30	0.419	-3.10	.004
SD 9	-0.827	0.221	-1.54	0.413	-3.74	.001
SD 10	-1.10	0.298	-1.51	0.407	-3.70	.001
SD 11	-0.640	0.234	-1.14	0.418	-2.74	.011

Interval-count model - OLS

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE (B)	<i>t</i>	<i>p</i>
Intercept	N/A	N/A	N/A	0.921	[0.641, 1.19]	0.102	9.020	< .0001
Int. 1/11	-0.811	[-0.895, -0.714]	0.056	-0.676	[-0.793, -0.567]	0.047	-14.377	< .0001
Int. 2/10	-0.338	[-0.470, -0.214]	0.056	-0.282	[-0.386, -0.180]	0.047	-6.013	< .0001
Int. 3/9	-0.087	[-0.190, 0.011]	0.057	-0.067	[-0.139, 0.010]	0.044	-1.535	.130
Int. 4/8	-0.105	[-0.216, 0.009]	0.057	-0.085	[-0.168, 0.008]	0.046	-1.863	.067

Int. 5/7	0.140	[0.010, 0.264]	0.056	0.117	[0.004, 0.217]	0.047	2.486	.016
Int. 6	-0.165	[-0.263, -0.065]	0.057	-0.180	[-0.297, -0.071]	0.062	-2.928	.005

Note. Adjusted $R^2 = .788$ (95% CI: [.691, .829]).

Interval-count model - RLM

	Beta	SE (Beta)	B	SE (B)	statistic	p
Intercept	N/A	N/A	0.833	0.0995	8.37	< .0001
Int. 1/11	-0.784	0.055	-0.653	0.046	-14.3	< .0001
Int. 2/10	-0.308	0.055	-0.257	0.046	-5.62	< .0001
Int. 3/9	-0.081	0.055	-0.063	0.043	-1.48	.139
Int. 4/8	-0.104	0.055	-0.084	0.045	-1.88	.062
Int. 5/7	0.159	0.055	0.132	0.046	2.89	.006
Int. 6	-0.147	0.055	-0.161	0.060	-2.68	.009

Scale-degree presence model - OLS

	Beta	95% CI (Beta)	SE (Beta)	B	95% CI (B)	SE (B)	t	p
Intercept	N/A	N/A	N/A	1.36	[0.891, 1.74]	0.192	7.060	< .0001
SD 1	-0.742	[-0.928, -0.543]	0.086	-1.13	[-1.36, -0.875]	0.132	-8.596	< .0001
SD 2	-0.426	[-0.622, -0.254]	0.083	-0.732	[-1.01, -0.444]	0.143	-5.125	< .0001
SD 3	-0.377	[-0.562, -0.164]	0.087	-0.576	[-0.843, -0.243]	0.133	-4.333	< .0001
SD 4	-0.203	[-0.398, -0.003]	0.090	-0.262	[-0.521, -0.008]	0.116	-2.262	.028
SD 5	-0.194	[-0.390, -0.093]	0.079	-0.725	[-1.08, -0.357]	0.297	-2.443	.018
SD 6	-0.481	[-0.689, -0.255]	0.088	-0.774	[-1.05, -0.398]	0.141	-5.490	< .0001
SD 7	0.006	[-0.210, 0.225]	0.090	0.008	[-0.263, 0.281]	0.117	0.069	.946
SD 8	-0.517	[-0.733, -0.331]	0.084	-0.965	[-1.22, -0.665]	0.157	-6.129	< .0001
SD 9	-0.394	[-0.609, -0.184]	0.084	-0.736	[-1.06, -0.292]	0.157	-4.702	< .0001
SD 10	-0.513	[-0.679, -0.340]	0.089	-0.702	[-0.938, -0.457]	0.122	-5.735	< .0001

SD 11	-0.571	[-0.773, -0.386]	0.084	-1.02	[-1.30, -0.760]	0.150	-6.810	< .0001
-------	--------	------------------	-------	-------	-----------------	-------	--------	---------

Note. Adjusted $R^2 = 0.631$ (95% CI: [0.361, 0.726]).

Scale-degree presence model - RLM

	Beta	SE (Beta)	B	SE (B)	statistic	p
Intercept	N/A	N/A	1.39	0.201	6.91	< .0001
SD 1	-0.758	0.091	-1.16	0.138	-8.38	< .0001
SD 2	-0.437	0.087	-0.751	0.150	-5.01	< .0001
SD 3	-0.390	0.091	-0.595	0.139	-4.27	< .0001
SD 4	-0.192	0.094	-0.249	0.121	-2.05	.047
SD 5	-0.206	0.083	-0.773	0.311	-2.48	.014
SD 6	-0.522	0.092	-0.840	0.148	-5.68	< .0001
SD 7	0.016	0.095	0.020	0.122	0.165	.872
SD 8	-0.537	0.088	-1.00	0.165	-6.07	< .0001
SD 9	-0.430	0.088	-0.802	0.164	-4.89	< .0001
SD 10	-0.523	0.094	-0.716	0.128	-5.58	< .0001
SD 11	-0.580	0.088	-1.04	0.157	-6.60	< .0001

**Parameters of Robust Regression Models Predicting Chords' Loadings on Principal Components
From Feature Scores**

All models use MM-estimation.

PC1 model

	Beta	SE (Beta)	B	SE (B)	statistic	<i>p</i>
Intercept	N/A	N/A	0.304	0.186	1.63	.110
Interference	0.775	0.071	0.792	0.072	11.0	< .0001
Harmonicity	0.089	0.084	0.276	0.262	1.05	.292
Familiarity	-0.613	0.060	-2.00	0.197	-10.2	.002
Cardinality	-0.600	0.071	-0.194	0.023	-8.41	< .0001
Major-scale compatibility (binary)	-0.107	0.084	-0.046	0.036	-1.28	.108
Major-scale compatibility	-0.173	0.104	-3.10	1.87	-1.66	.203

PC2 model

	Beta	SE (Beta)	B	SE (B)	statistic	<i>p</i>
Intercept	N/A	N/A	-0.863	0.163	-5.30	< .0001
Interference	-0.226	0.0818	-0.175	0.0631	-2.77	.007
Harmonicity	0.156	0.0974	0.368	0.229	1.61	.108
Familiarity	-0.330	0.0699	-0.813	0.172	-4.72	< .0001
Cardinality	0.844	0.0827	0.206	0.0202	10.2	< .0001
Major-scale compatibility (binary)	0.100	0.097	0.032	0.032	1.02	.303
Major-scale compatibility	-0.235	0.121	-3.19	1.63	-1.95	.056

References

- Bowling, D. L., Purves, D., & Gill, K. Z. (2018). Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences*, 115(1): 216–221. <https://doi.org/10.1073/pnas.1713206115>
- Hutchinson, W., & Knopoff, L. (1978). The acoustic component of western consonance. *Journal of New Music Research*, 7(1): 1–29. <https://doi.org/10.1080/09298217808570246>
- Marjeh, R., Harrison, P. M. C., Lee, H., Deligiannaki, F., & Jacobi, N. (2024). Timbral effects of consonance disentangle psychoacoustic mechanisms and suggest perceptual origins for musical scales. *Nature Communications*, 15: 1482. <https://doi.org/10.1038/s41467-024-45812-z>
- Milne, A. J. (2013). A computational model of the cognition of tonality (PhD thesis). The Open University, Milton Keynes, England.
- Plomp, R., & Levelt, W. J. M. (1965). Tonal Consonance and Critical Bandwidth. *The Journal of the Acoustical Society of America*, 38: 548–560. <https://doi.org/10.1121/1.1909741>
- Sethares, W. A. (1993). Local consonance and the relationship between timbre and scale. *The Journal of the Acoustical Society of America*, 94: 1218–1228. <https://doi.org/10.1121/1.408175>