Reshaping musical consonance with timbral manipulations and massive online experiments

Raja Marjieh[1, 2, *], Peter M. C. Harrison[2, 3, †] (equal contribution)
Harin Lee[2, 4], Fotini Deligiannaki[2, 5], Nori Jacoby[2, ‡]

[1] Department of Psychology, Princeton University, Princeton, New Jersey, USA
[2] Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
[3] Centre for Music and Science, University of Cambridge, Cambridge, UK
[4] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[5] German Aerospace Center (DLR), Institute for AI Safety and Security, Bonn, Germany

* raja.marjieh@princeton.edu
† pmch2@cam.ac.uk
‡ nori.jacoby@ae.mpg.de

Author summary: *"We reveal effects of timbre on consonance perception that motivate a new understanding of the evolution of musical scales."*

The phenomenon of musical 'consonance' is crucial for many musical styles, determining how notes are organized into scales, how scales are tuned, and how chords are constructed from scales. Western music theory assumes that consonance depends solely on frequency ratios between chord tones; however, psychoacoustic theories predict a dependency also on the 'timbre' (tone color) of the underlying sounds. We investigate this possibility with 24 large-scale behavioral experiments (4,666 participants), constructing detailed continuous maps of consonance judgments for different timbres, and simulating these judgments with representative computational models. We find that timbral manipulations can indeed modify consonance judgments, transforming both the magnitude and the location of consonance peaks. We show how these results shed new light into the mechanisms underlying consonance perception as well as the cultural evolution of scale systems. More broadly, this work showcases how large-scale behavioral experiments can inform classical questions in auditory perception.

Introduction

Many musical styles involve multiple performers playing or singing simultaneously (Brown & Jordania, 2013; Mehr et al., 2019; Savage et al., 2015). In Western music, this practice is underwritten by the notion of ‘harmony’, defining how multiple musical tones may be combined together into polyphonic ‘sonorities’ or ‘chords’. To a given listener, certain chords will sound particularly pleasant, or ‘consonant’, while others will sound relatively unpleasant, or ‘dissonant’. This phenomenon has immense importance in many musical styles, determining how musical notes are organized into scales, how these scales are tuned, and how chords are constructed from these scales (Chiba et al., 2019; Gill & Purves, 2009; Hall, 1973; Huron, 1994, 2001). It has consequently drawn sustained attention from many researchers ranging from philosophers (Pythagoras) to mathematicians (Leibniz, Euler), music theorists (Zarlino, Rameau), and modern-day psychologists and ethnomusicologists (Bowling & Purves, 2015; Eerola & Lahdelma, 2021; Euler, 1739; Friedman et al., 2021; McDermott et al., 2016; Rameau, 1722; Sethares, 2005; Smit & Milne, 2021; Stolzenburg, 2015; Tenney, 1988; Terhardt, 1974).

Consonance perception is thought to derive from both psychoacoustic and cultural factors (e.g. Harrison & Pearce, 2020). Several psychoacoustic mechanisms have been proposed over the years, including fusion (Stumpf, 1890, 1898) and combination tones (Krueger, 1904, 1910; Preyer, 1879), but the two main extant theories attribute consonance either to interference between partials (Helmholtz, 1875) or to harmonicity detection (Terhardt, 1974) (see Friedman et al., 2021; Harrison & Pearce, 2020; McDermott et al., 2010 for a similar conclusion). Both theories predict that chords comprising harmonic tones should sound most pleasant when the tones are related by *harmonic* pitch intervals (i.e. those that correspond to simple frequency ratios, e.g. 2:1, 3:2). This would explain why scale systems across the world seem to have developed to favor harmonic pitch intervals (Gill & Purves, 2009; McBride & Tlusty, 2021). Within a given society, cultural familiarity will further contribute to consonance perception, biasing listeners towards preferring sonorities that occur often within a given musical style. Styles based on Western tonality will therefore reinforce preferences for harmonic pitch intervals (Harrison & Pearce, 2020; Johnson-Laird et al., 2012), but styles with different harmonic systems may induce different biases (Ambrazevičius, 2017; McDermott et al., 2016).

To this date, a crucial unsolved question has been whether consonance perception depends on the *timbre* of the underlying chord tones. Previous literature provides contradictory perspectives here. Traditional Western music theory implies that consonance should be independent of timbre; it provides just one scheme for categorizing intervals into consonances and dissonances, and this scheme applies equally to all musical instruments (Tenney, 1988). On the other hand, prominent psychoacoustic theories of consonance (in particular, Helmholtz's interference theory) imply that consonance judgments should vary substantially depending on the positions and magnitudes of the tones' upper harmonics (Helmholtz, 1875). In apparent contradiction to Helmholtz's theory, however, recent decades of psychological studies seem to show that timbral manipulations do not qualitatively affect consonance judgments (Friedman et al., 2021; McDermott et al., 2010; McLachlan et al., 2013; Nordmark & Fahlén, 1988).

Here we address this question with a series of 24 large-scale behavioral experiments with 4,666 participants. These experiments have three particularly important features:

Continuous treatment of pitch intervals. Previous consonance research has used stimuli drawn solely from discrete scales, in particular the Western 12-tone chromatic scale (Bowling et al., 2018; Friedman et al., 2021; Johnson-Laird et al., 2012; McDermott et al., 2010; McLachlan et al., 2013; McPherson et al., 2020; Nordmark & Fahlén, 1988). This is problematic because it neglects potentially interesting structure in between the scale degrees of the chromatic scale, and because the resulting paradigm is inherently Western-centric. Here we instead avoid making any assumptions about scale systems, and instead take advantage of novel psychological techniques (dense rating; Gibbs Sampling with People; Harrison et al., 2020) to construct continuous consonance maps directly from behavioral data.

Systematic exploration of timbral features. Several recent consonance studies have included timbral manipulations, but generally only explored a limited number of manipulations (Friedman et al., 2021; McLachlan et al., 2013; Nordmark & Fahlén, 1988) or used manipulations designed to demonstrate generalizability rather than to test particular hypotheses (McDermott et al., 2010). Here we take a more systematic approach. We focus in particular on spectral manipulations, because (as we show later) these yield particularly clear hypothetical effects in computational modeling. In a series of studies, we address the three main ways of manipulating a harmonic spectrum: (a) changing the *frequencies* of the harmonics (Study 2), (b) changing the *amplitudes* of the harmonics (Study 3), and (c) deleting individual harmonics entirely (Study 4).

Concurrent computational modeling. Previous research has developed many computational models operationalizing different theories of consonance perception. Here we use such models to understand what predictions different theories should make for different spectral manipulations. We focus in particular on the two psychoacoustic models that performed best in a recent systematic evaluation of almost all extant consonance models (Harrison & Pearce, 2020): the interference model of Hutchinson and Knopoff (1978), which calculates interference by summing contributions from all pairs of partials in the acoustic spectrum, and the harmonicity model of Harrison and Pearce (2018), which models the precision of a template-based pitch-finding process. We confirm the robustness of the results by running supplementary analyses with a collection of alternative interference models (Sethares, 1993; Vassilakis, 2001) and harmonicity models (Boersma, 1993; Milne, 2013) (*Supplementary Materials*).

Together, these 24 experiments characterize the relationship between timbre and consonance with an unprecedented level of detail, shedding new light on the psychological mechanisms underpinning consonance perception, as well as the close connection between musical instruments and the cultural evolution of musical styles.

Results

Baseline results for harmonic dyads (Study 1)

Consonance perception may be modeled as a function that maps from a given collection of notes (e.g. C4, E5, G5) to a scalar number representing the participant's subjective evaluation (e.g. the number '0.6', corresponding to 'very consonant'). Our task here is to estimate the shape of this function, and to understand how it is moderated by timbre.

Following a long tradition of music theory and music psychology research, we begin by studying the consonance of two-tone chords (*dyads*) as a function of the frequency ratio between those tones (e.g., Bowling et al., 2018; Helmholtz, 1875; Hutchinson & Knopoff, 1978; McDermott et al., 2010; Schwartz et al., 2003). We represent those frequency ratios as *pitch intervals*, where the pitch interval in semitones is calculated as $12 \log_2 \frac{f_2}{f_1}$, where f_1 and f_2 are the two frequencies. In subsequent studies (Study 5, Study 7) we then generalize the approach to three-tone chords (*triads*). Throughout these studies we investigate the moderating role of timbre.

Study 1A characterizes dyadic consonance perception for synthetic *harmonic complex tones*. These tones are constructed by combining pure tones (i.e., simple sinusoids) whose frequencies are all integer multiples (i.e., *harmonics*) of a common *fundamental frequency*. Such tones have long been used as idealized approximations of the pitched sounds produced by the human voice and by common musical instruments (e.g. Hutchinson & Knopoff, 1978; Kameoka & Kuriyagawa, 1969; Plomp & Levelt, 1965).

In each trial of the experiment, we played US participants (*N* = 198) dyads comprising two harmonic complex tones (10 harmonics per tone, 3 dB/octave spectral roll-off, 1.3 s in duration), sampling the pitch interval between the two tones from a uniform distribution over the range of 0-15 semitones, and sampling the pitch of the lower tone from a uniform distribution over the range G3-F4. The participant was then asked to rate the dyad's 'pleasantness' on a scale from 1 ('completely disagree') to 7 ('completely agree') (Figure 1A). Following previous research, we use pleasantness as a convenient synonym for consonance that is understood well by nonmusicians (e.g. Lahdelma & Eerola, 2020; McDermott et al., 2010, 2016). We collected many ratings from many participants for many dyads, and summarized the results using a Gaussian kernel smoother (bandwidth: 0.2 semitones), finding consonance peaks using a peak-picking algorithm constructing 95% confidence intervals using nonparametric bootstrapping (see *Methods*).

Figure 1B summarizes the results from Study 1A (see *Supplementary Materials* for a video version). Despite making no assumptions about discrete scale systems in the experiment's design, we see the Western discrete scale system emerging from the data, as well as traditional hierarchies of consonance and dissonance. In particular, we find eight clear peaks in the pleasantness judgments; these peaks are located close to the integer semitones that make up the Western 12-tone chromatic scale (average distance of 0.05, 95% CI: [0.03, 0.08] semitones; random chance would give 0.25 semitones). The relative heights of these peaks replicate traditional music-theoretic classifications of intervals into 'consonant' (blue) and 'dissonant' (red) categories (Figure 1C; mean difference between consonant/dissonant intervals is 0.38, 95% CI: [0.29, 0.46]). The results also correlate very well with the results of previous behavioral experiments studying the relative consonance of different Western intervals, including aggregated results from seven laboratory studies from the late 19th/early 20th centuries (Schwartz et al., 2003) (r = .96, 95% CI: [.85, .99], ρ = .94), a recent laboratory study by Bowling et al. (2018) (r = .91, 95% CI: [.71, .98], ρ = .95), and a recent online study by McPherson et al. (2020) (r = .94, 95% CI: [.84, .98], ρ = .94). Lastly, a Monte Carlo split-half correlation analysis showed an excellent internal reliability (r = .87, 95% CI: [.74, .94], 1,000 permutations). Together, these results give us confidence in the reliability and validity of our experimental methods.

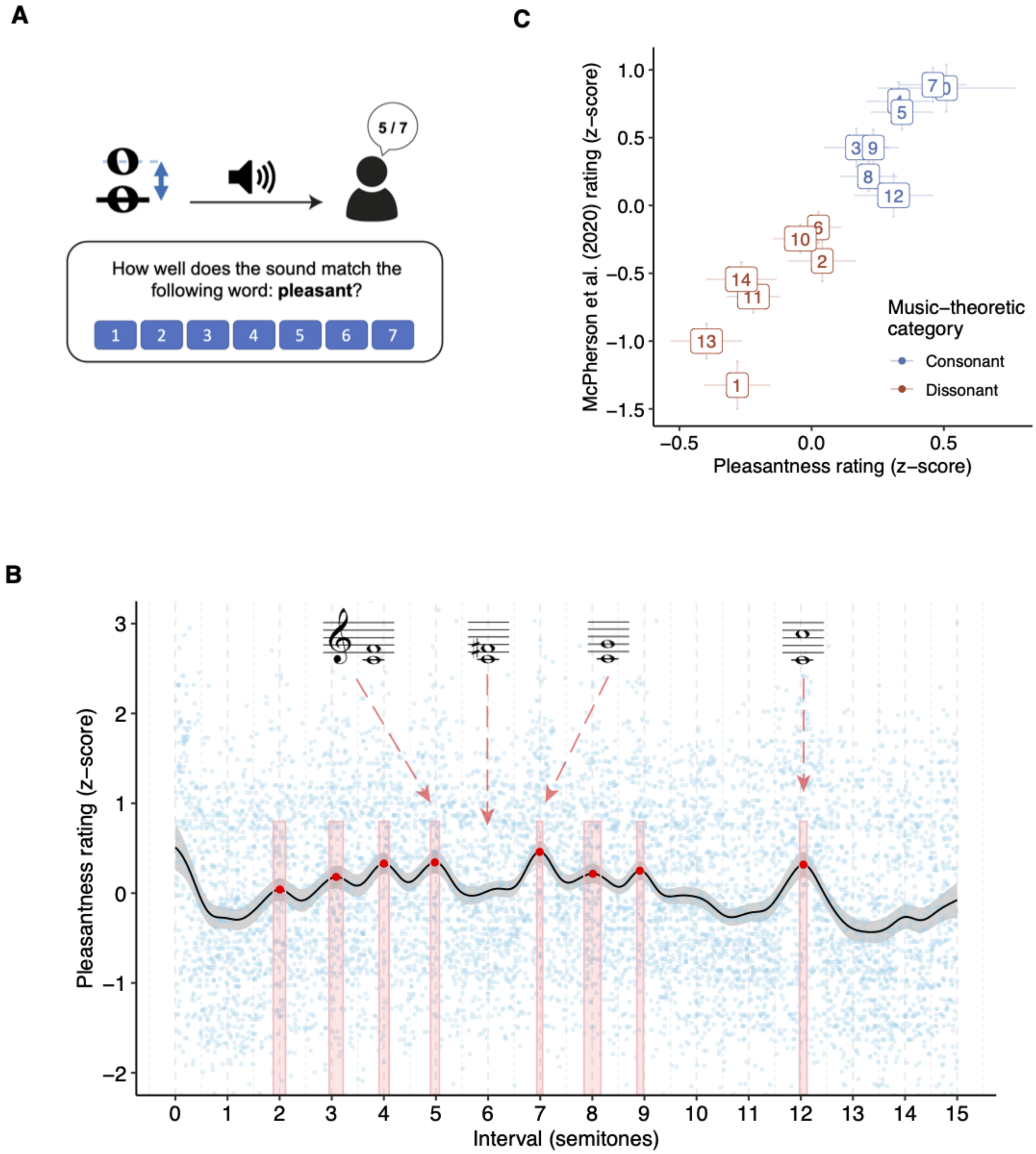


Figure 1. Dyadic consonance for harmonic complex tones (Study 1A, 198 participants). (**A**) Schematic illustration of the rating task. (**B**) Consonance profile (black line) derived through kernel smoothing (*z*-scored, Gaussian kernel, bandwidth 0.2 semitones, 95% confidence interval), superposed on raw data (blue points). Peaks estimated by a peak-picking algorithm are marked in red with 95% confidence intervals (bootstrapped). (**C**) Comparing smoothed ratings at integer intervals to traditional music-theoretic classifications and to data from McPherson et al. (2020) (*z*-scored over participants, 95% bootstrapped confidence intervals).

Synthetic harmonic complex tones are traditionally intended as approximations to the kinds of complex tones produced by ‘real’ musical instruments. To verify that they elicit similar consonance profiles, we conducted three follow-up experiments repeating Study 1A but with tones from three synthetic musical instruments: flute, guitar, and piano (Study 1B, 602 participants). We find that these instruments indeed produce broadly comparable consonance profiles to the idealized harmonic complex tones (Figure 2; mean r = .56, mean ρ = .62) though with certain peaks falling either side of the 95% statistical significance cutoff for different instruments (e.g. the minor 7th peak, 10 semitones, is only statistically significant for the guitar).

The differences we see between these instruments are potentially interesting but difficult to interpret definitively, because the instruments vary on multiple factors, including both temporal and spectral features. In the rest of this paper, we therefore focus on *artificial* harmonic complex tones, manipulated in very precise and interpretable ways, so that we can better distinguish the underlying causal factors that affect consonance perception.

Previous research has documented how consonance judgments can vary as a function of musical experience (e.g., Plomp & Levelt, 1965; Popescu et al., 2019). In all of our experiments, we therefore asked our participants to report their years of musical experience, allowing us to estimate the sensitivity of our results to musical expertise. Figure 3 plots results for Study 1A differentiated by musical expertise (median split: 2.0 years of musical experience; the *Supplementary Materials* provide analogous plots for all other dyadic consonance experiments). We found, in general, that participants with different levels of musical experience gave qualitatively similar results. Participants with more musical experience (> 2 years, ‘musicians’) tended to give more differentiated judgments than participants with less musical experience (≤ 2 years, ‘nonmusicians’) (mean *SD* of *z*-scored ‘musician’ profiles = 0.23 [0.18, 0.28], mean *SD* of *z*-scored ‘nonmusician’ profiles = 0.15 [0.11, 0.18], mean difference = 0.08 [0.07, 0.10], bootstrapped over experiments), but in general the judgments correlated quite highly across both groups (mean ρ of .68 [.57, .80], bootstrapped over experiments). This consistency may be due to universal psychoacoustic processes, but it may also be due to the sophisticated implicit musical knowledge that listeners are known to develop even in the absence of formal musical training (Bigand & Poulin-Charronnat, 2006; Müllensiefen et al., 2014). In the following studies, we focus on analyzing data aggregated over all musical experience levels.

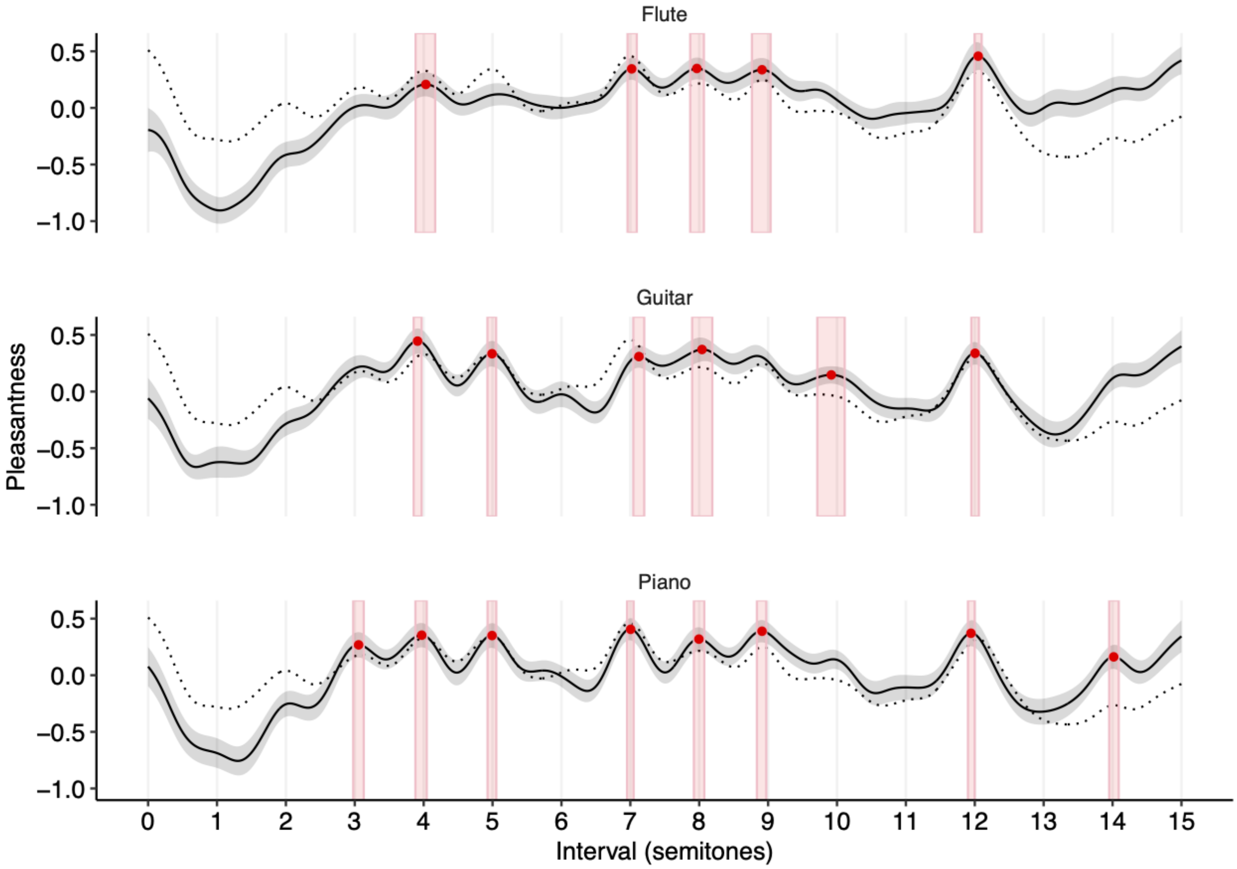


Figure 2. Dyadic consonance for synthesized Western instruments (Study 1B). Consonance profiles for the Western instruments (*z*-scored) are plotted with 95% bootstrapped confidence intervals (bandwidth of 0.2 semitones). The consonance profile for harmonic complex tones (Study 1A) is plotted as a reference dotted line.

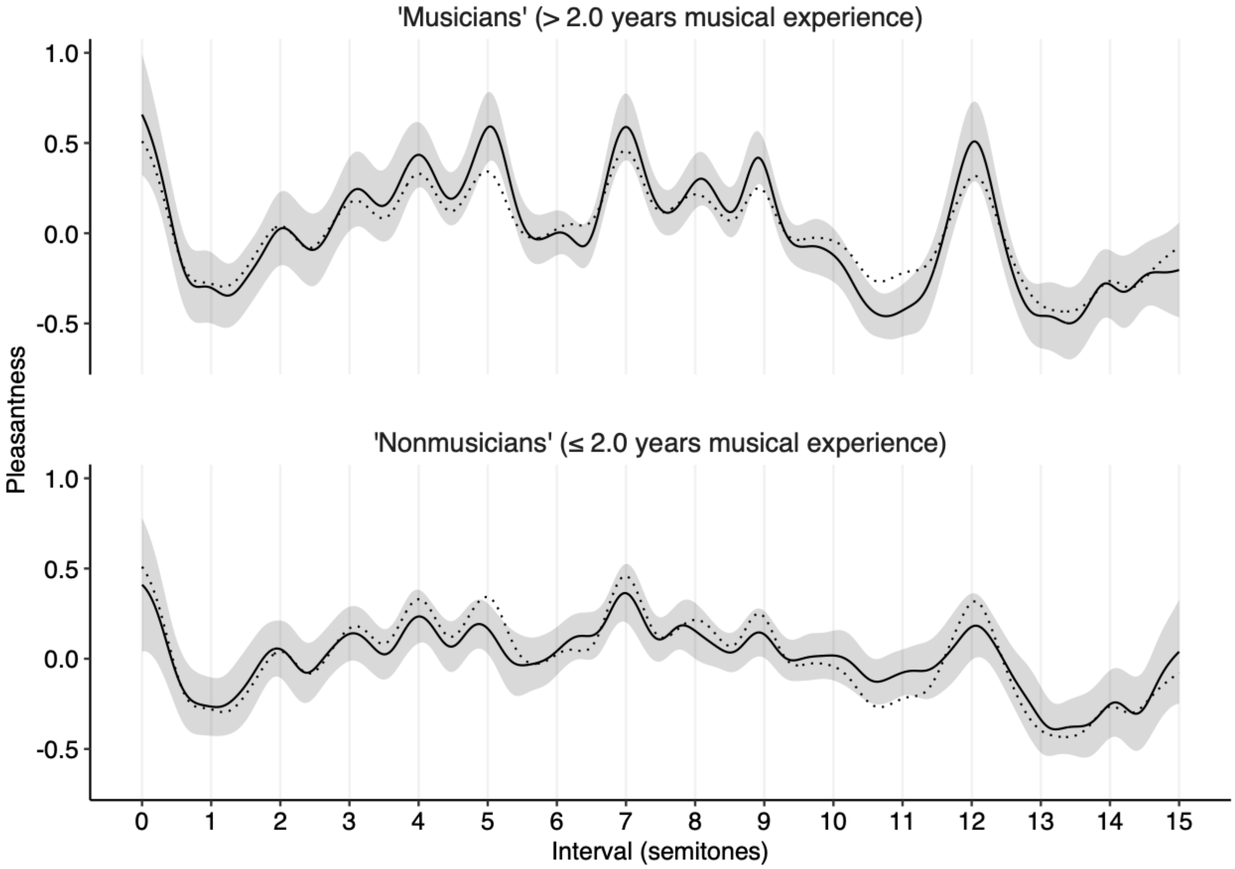


Figure 3. Dyadic consonance for harmonic complex tones as a function of musicianship (Study 1A). Consonance profiles (*z*-scored) are plotted with 95% bootstrapped confidence intervals (bandwidth of 0.2 semitones). The musicianship threshold (2.0 years) corresponds to a median split of the participant group. The reference dotted line corresponds to the consonance profile derived from the full participant group.

Changing harmonic frequencies (Study 2)

We begin by considering how consonance judgments may be altered if we change the frequencies of the harmonics that make up the complex tone. Such effects have been previously hypothesized in previous work (Sethares, 2005) but not yet empirically tested.

We first consider a ‘stretching’ manipulation proposed by Sethares (2005), where we manipulate the spacing between the harmonics in the complex tone (Study 2A). Similar kinds of stretching are present to small degrees in string instruments as a consequence of string stiffness, causing slight inharmonicity (Young, 1952). We define the frequency of the ith partial as $f_i = f_0 \gamma^{\log_2(i+1)}$, where f_0 is the fundamental frequency and γ is the stretching parameter: $\gamma = 1.9$ then defines a ‘compressed’ tone, $\gamma = 2$ defines a standard harmonic tone, and $\gamma = 2.1$ defines a ‘stretched’ tone (Figure 4A).

Interference models predict that this spectral stretching/compression manipulation should yield analogous stretching/compression in consonance profiles (e.g., Figure 4B, red lines; see *Supplementary Materials* for equivalent results from alternative models). Intuitively, this can be understood from the observation that interference is minimized when partials from different tones align neatly with each other; if we then stretch each tone’s spectrum, we must also stretch the intervals between the tones to maintain this alignment.

Interestingly, harmonicity models do not predict such an effect; instead, they generally predict that these manipulations will largely eliminate pleasantness variation, and any residual variation will still be located at harmonic intervals (e.g., Figure 4B, blue lines; see also *Supplementary Materials*). Once the individual tones become inharmonic, the overall chord also becomes inharmonic, irrespective of the intervals between the tones.

We conducted a pair of experiments to construct consonance profiles for stretched (Study 2A(i), 194 participants) and compressed (Study 2A(ii), 202 participants) tones, and compared these to baseline profiles for harmonic tones (Study 1A, 198 participants). As predicted by the interference account, we find that we can indeed induce preferences for stretched and compressed intervals, in line with the corresponding spectral manipulations (Figure 4B; see *Supplementary Materials* for video versions). For example, for dyads comprising stretched tones, we clearly see preferences for stretched octaves (peak at 12.78 [12.68, 12.88] semitones; an unstretched octave would be 12.00 semitones) contrasting with the results from harmonic tones (peak at 12.04 [11.97, 12.11] semitones) (Figure 4B). We see similar stretching/compression for other consonant intervals, though in some cases the peaks lose clarity for the inharmonic tones. These effects are consistent with the predictions of the interference models, but not with the predictions of the harmonicity models; the results therefore provide important evidence that interference between partials is an important contributor to consonance perception, in contrast to recent claims in the literature that interference is irrelevant to consonance perception (Bowling et al., 2018; Cousineau et al., 2012; McDermott et al., 2010, 2016).

Other kinds of inharmonicity can be produced by certain percussion instruments, for example the metallophones of Indonesian gamelans and the xylophone-like *renats* used in Thai classical music (Sethares, 2005). Each instrument will have an idiosyncratic spectrum that reflects its particular physical construction, with potentially interesting implications for consonance perception.

In Study 2B we investigate an inharmonic tone inspired by one such instrument, the *bonang*. The bonang is an instrument from the Javanese gamelan comprising a collection of small gongs. In order to achieve arbitrary microtonal pitches, we use a synthetic approximation to the bonang proposed by Sethares (2005) on the basis of field measurements, corresponding to four equally weighted harmonics with frequencies of f_0, $1.52f_0$, $3.46f_0$, and $3.92f_0$ (Figure 5A). Following Sethares (2005), we play dyads where the upper tone corresponds to this idealized bonang, and the lower tone corresponds to a standard harmonic tone with four equally weighted harmonics. This combination is intended to reflect a common kind of texture in Javanese gamelan music, where the inharmonic bonang is played alongside a harmonic instrument or voice.

The results from the corresponding dyad rating experiment are plotted in Figure 5B (Study 2B, 170 participants; see *Supplementary Materials* for a video version). As with the harmonic tones, we see a clear pleasantness peak at the octave, 11.98 semitones [11.88, 12.05]. However, the other peaks previously seen at harmonic intervals are now either missing or displaced to inharmonic locations. In particular, we see clear peaks at 2.60 [2.51, 2.67] semitones and 4.80 [4.70-4.95] semitones, neither of which are harmonic intervals. Conversely, we do not see any peaks at the major third (no peak detected in 95% of bootstrap samples within the interval [3.5, 4.5]) or the perfect fifth (no peak detected in 76% of bootstrap samples within the interval [6.5, 7.5]. These peaks are each predicted by the interference model and the harmonicity model; however, the harmonicity model also predicts several additional peaks that do not manifest clearly in the behavioral data (see also *Supplementary Materials* for additional models).

Sethares (2005) made an interesting claim that the two main scales of Javanese gamelan music (the *slendro* scale and the *pelog* scale) reflect the consonance profiles of its instruments (see also Li, 2006; Schneider, 2001). In particular, he proposed that the inharmonic *slendro* scale might be explained in terms of the consonance profile produced by combining a harmonic complex tone with a bonang tone, as in our own experiment. We have correspondingly annotated Figure 5B with the locations of the slendro scale degrees, approximating the scale as 5-tone equal temperament (Rahn, 1996; Sethares, 2005). As predicted by Sethares (2005), we do find that the empirical consonance curve aligns neatly with the slendro scale, even though our Western participants are likely to have no or little exposure to Javanese gamelan music; in particular, each observed peak (2.6, 4.8, and 12.0 semitones) is located close to a slendro scale degree (2.4, 4.8, and 12.0 semitones). Interestingly, while the two remaining slendro scale degrees (7.2 and 9.6 semitones) do not have corresponding behavioral peaks, they do have corresponding peaks in the harmonicity curve.

To summarize, we have found that manipulating the frequencies of the harmonics can induce inharmonic consonance profiles. In particular, stretching/compressing the harmonic series leads to stretched/compressed consonance profiles (Study 2A), whereas replacing the upper dyad tone with a synthetic bonang tone yields an idiosyncratic consonance profile that aligns with the slendro scale from the Javanese gamelan (Study 2B), even for participants with little or no prior experience with this scale. The stretching/compressing manipulation is particularly interesting from a modeling perspective, because it clearly dissociates the predictions of the interference and the harmonicity models, and shows that only the former are compatible with the behavioral data. The latter manipulation is particularly interesting from a cultural evolution perspective, because it supports the hypothesis that the slendro scale developed in part as a specific consequence of the acoustic properties of Javanese gamelan instruments (Li, 2006; Schneider, 2001; Sethares, 2005).

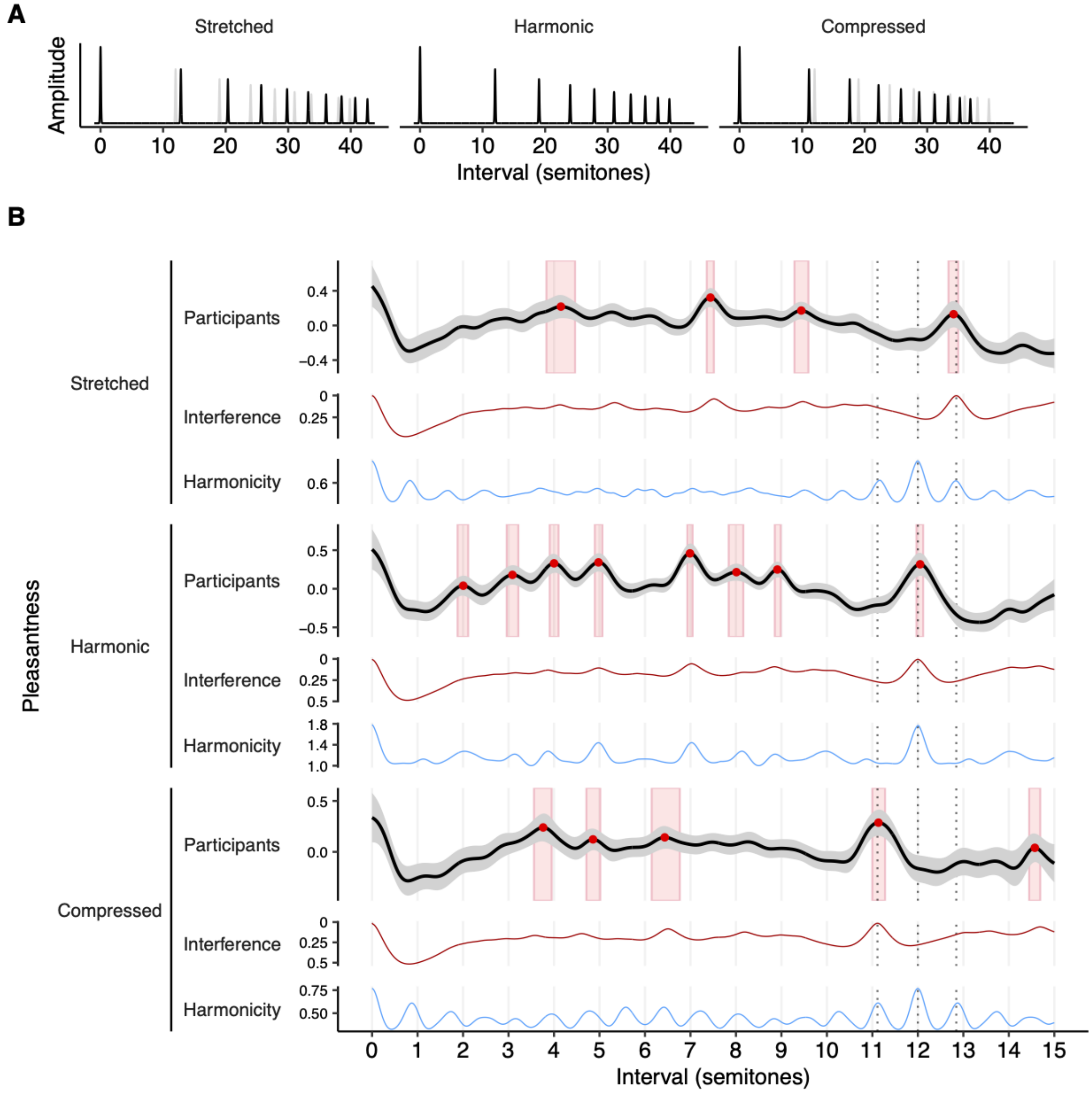


Figure 4. Spectral stretching/compression and consonance (Studies 1A, 2A). (A) Stretched and compressed tone spectra, with a baseline harmonic spectrum (gray) for comparison. **(B)** Dyadic pleasantness judgments for stretched, harmonic, and compressed tones. Behavioral results are summarized using a kernel smoother with a bandwidth of 0.2 semitones, with 95% confidence intervals (bootstrapped) shaded in gray, peak locations plotted as red circles, and 95% confidence intervals (bootstrapped) for peak locations shaded in red. Dotted lines indicate the location of the compressed, harmonic, and stretched octaves.

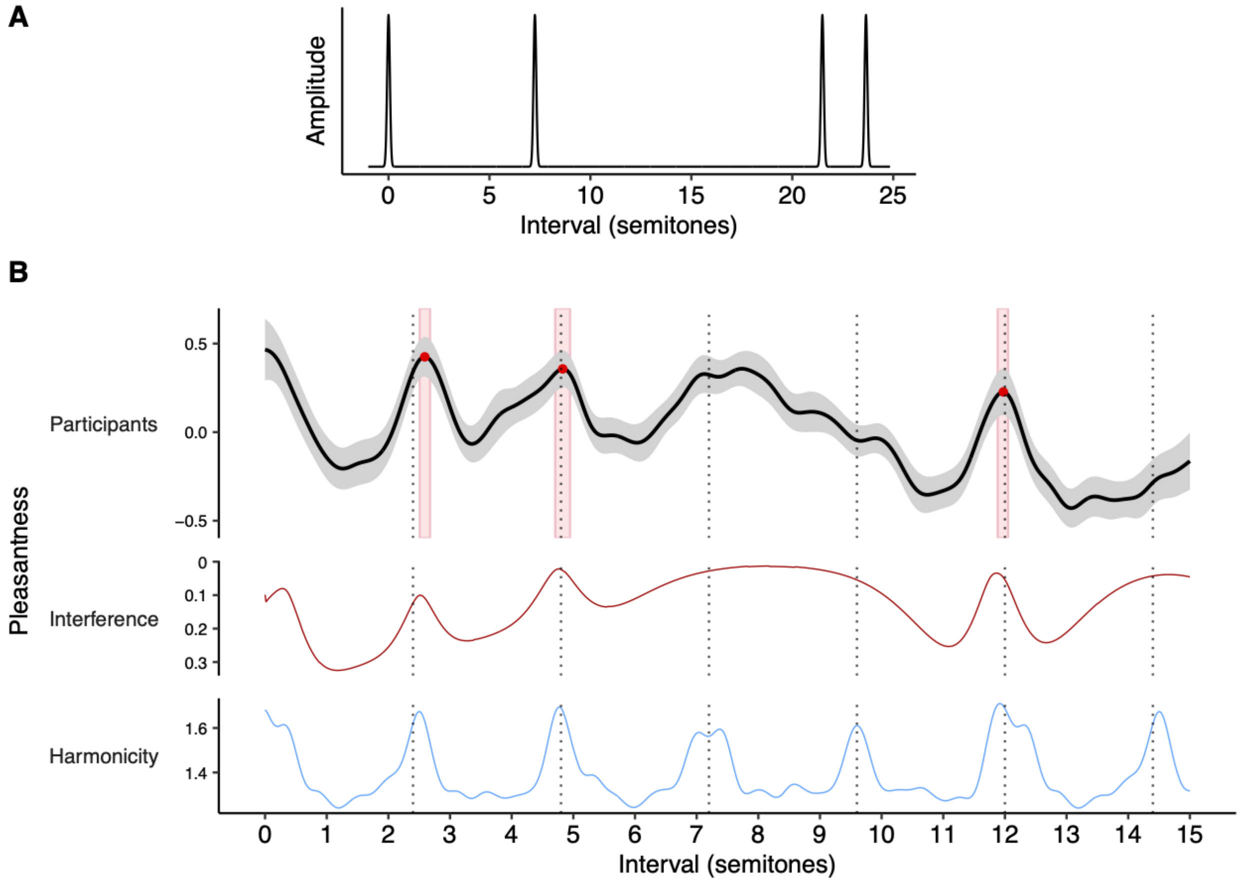


Figure 5. Dyadic pleasantness judgments for the bonang (Study 2B, 170 participants). (A) Idealized spectrum for the bonang (Sethares, 2005). **(B)** Pleasantness judgments (95% confidence intervals) for dyads comprising a harmonic complex tone (lower) combined with an idealized bonang tone (upper) (Study 2B, 170 participants). Peak locations are plotted as red circles with 95% confidence intervals in red rectangles. Interference (Hutchinson & Knopoff, 1978) and harmonicity (Harrison & Pearce, 2018) model predictions are plotted for reference. The slendro scale, approximated as 5-tone equal temperament (Sethares, 2005), is plotted with dashed lines; note how this scale barely overlaps with the 12-tone scale.

Changing harmonic amplitudes (Study 3)

We now consider how consonance profiles may be affected by changing the *amplitudes* of their tones' harmonics. In particular, we focus on the so-called *spectral roll-off* parameter, which determines the rate at which harmonic amplitude rolls off (decreases) as harmonic number increases. For example, a tone with *high roll-off* might have amplitude decrease at a rate of 12 dB per octave, whereas a tone with *low roll-off* might have amplitude decrease at only 3 dB per octave (Figure 6A).

Interference theories would predict that roll-off has a major influence on consonance judgments: increased roll-off should reduce the magnitude of beating effects induced by the upper harmonics, hence producing flatter consonance profiles (Figure 6B, red lines). In contrast, harmonicity theories predict that pleasantness profiles should remain highly differentiated. Harmonicity comes primarily from integer relationships between fundamental frequencies, a phenomenon which is relatively robust in the face of roll-off manipulations (Figure 6B, blue lines).

Interference and harmonicity theories also predict potential main effects of spectral roll-off on pleasantness (with higher overall pleasantness rating for larger roll-off parameters). According to the interference models, almost every interval elicits some interactions between upper harmonics, and hence becomes more pleasant with increased roll off (Figure 6B, red lines; standardized regression coefficients (β) for the main effects = 0.39, 0.83, 0.62 for the three models). Harmonicity models make less consistent predictions: some predict an overall main effect on pleasantness (e.g., Figure 6B, blue lines; β = 0.91), whereas others do not predict a strong effect (β = 0.02, 0.23; *Supplementary Materials*).

We tested these predictions in Study 3 with a dense dyad rating experiment (322 participants) manipulating both pitch interval (0-15 semitones) and roll-off (0-15 dB roll-off/octave, Figure 6A; see *Supplementary Materials* for a video version). We see a clear main effect of roll off (as predicted by both theories), with participants finding greater roll-offs more pleasant (Figure 6B). However, we see no clear effect on pleasantness variability; the profiles remain highly differentiated for all roll-off levels. Indeed, we find that a generalized additive model using just main effects of roll-off and of pitch interval can explain 98% of the variance of smoothed consonance ratings, indicating that, despite its strong main effect (β = 0.89), roll off has a minimal effect on the *shape* of the consonance profile.

These results are clearly inconsistent with the interference model, which predicted that the consonance profile should lose its differentiation at higher levels of spectral roll-off; in contrast, they are highly consistent with the harmonicity model, which predicted that the profile's differentiation should remain preserved. Unlike Study 2A, which yielded support for the interference account of consonance, Study 3 therefore yields support for the harmonicity account. We will return to this apparent paradox in the *Discussion*.

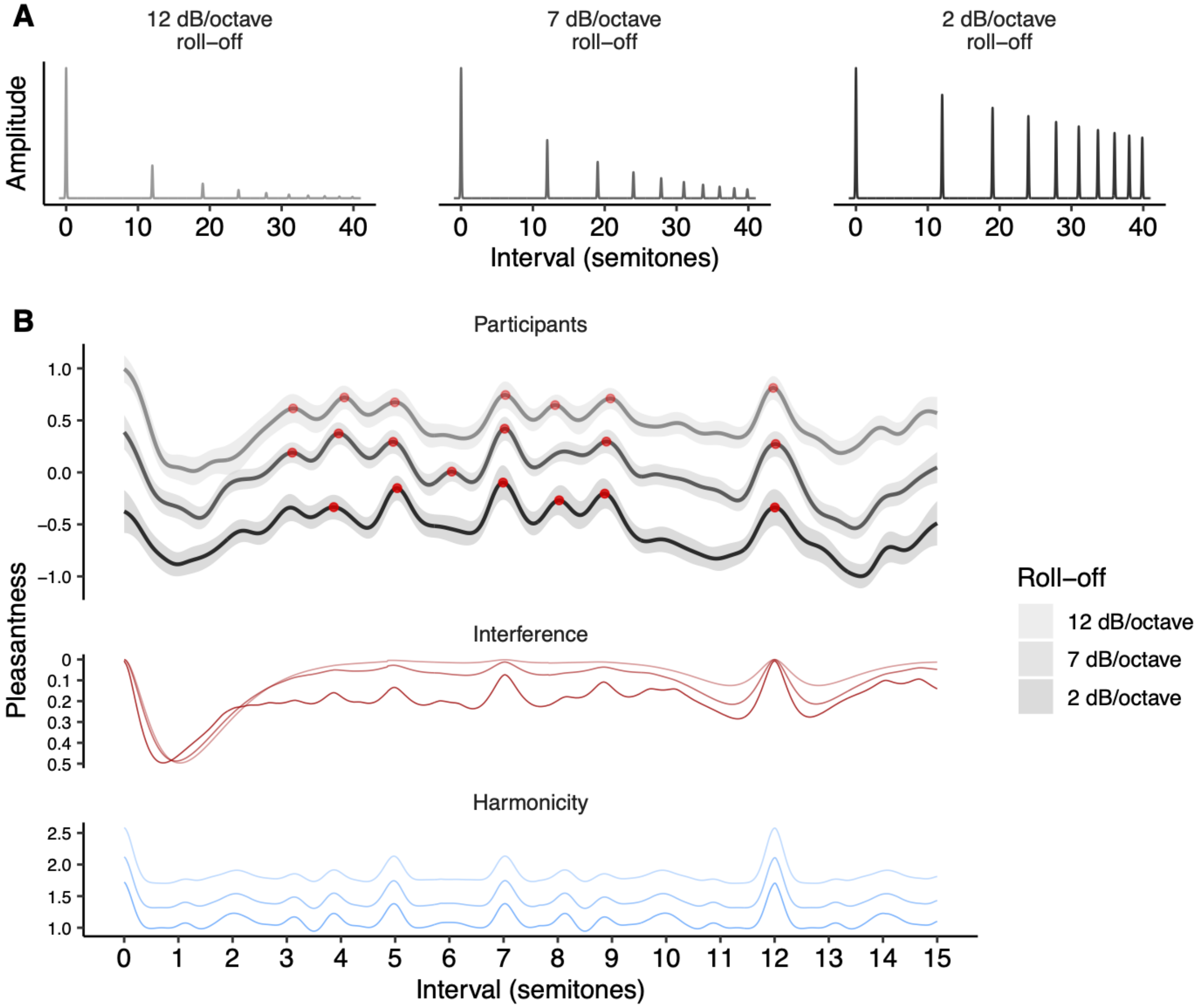


Figure 6. Dyadic consonance as a function of roll-off (Study 3, 322 participants). (**A**) Tone spectra for three representative levels of roll-off (12, 7, and 2 dB/octave). (**B**) Pleasantness judgments (95% confidence intervals) for these three roll-off levels computed using kernel smoothing (bandwidth = 0.2 semitones), plotted alongside interference (Hutchinson & Knopoff, 1978) and harmonicity (Harrison & Pearce, 2018) model predictions. Peak locations are plotted as red circles.

Deleting entire harmonics (Study 4)

Study 3 found that reducing harmonic amplitudes had little effect on the shape of consonance profiles, in contrast to the predictions of the interference model. In Studies 4A and 4B, we now ask whether consonance profiles are affected by a more radical manipulation: completely *deleting* particular harmonics.

Some previous studies have investigated the effects of harmonic deletion, with mostly negative results. Vos (1986) assessed subjective purity judgments in the neighborhood of the major third and the perfect fifth, and found that purity ratings increased when removing the even harmonics from the upper tone, but the overall shape of the consonance profiles remained broadly similar. Nordmark and Fahlén (1988) took the minor ninth dyad, and investigated the effect of deleting the partials in each tone theoretically responsible for the most interference; however, they found no effect on consonance judgments. McLachlan et al. (2013) tried removing (a) all even harmonics and (b) all harmonics above the fundamental frequency for a collection of dyads, but likewise found no clear effect on consonance judgments.

On the face of it, these historic results seem to be conclusive evidence against interference theories of consonance, which clearly predict that the consonance of non-unison intervals should depend on the existence of upper harmonics (McLachlan et al., 2013; Nordmark & Fahlén, 1988). However, there are a couple of reasons to be skeptical here. One is that these historic studies do not test intervals ‘in between’ the scale degrees of the 12-tone chromatic scale, and therefore potentially miss certain changes to the consonance profiles. The second is that historic studies using pure tones (i.e. tones with no upper harmonics) have often identified different (typically flatter) profiles to those that used complex tones (e.g., Plomp & Levelt, 1965).

We therefore reexamined this question in Study 4A (485 participants). We focused on three tone types (Figure 7A): a tone with five equally weighted harmonics (bottom row), a tone with the third harmonic deleted (middle row), and a pure tone with all upper harmonics deleted (top row). We chose this combination of tones because (a) it is possible to delete the third harmonic without changing the tone’s spectral centroid (i.e., mean spectral frequency), which is useful because the spectral centroid is known to contribute to pleasantness (Eerola & Lahdelma, 2021), and because (b) pure tones are good for making comparisons with prior studies (e.g. McDermott et al., 2010; McLachlan et al., 2013).

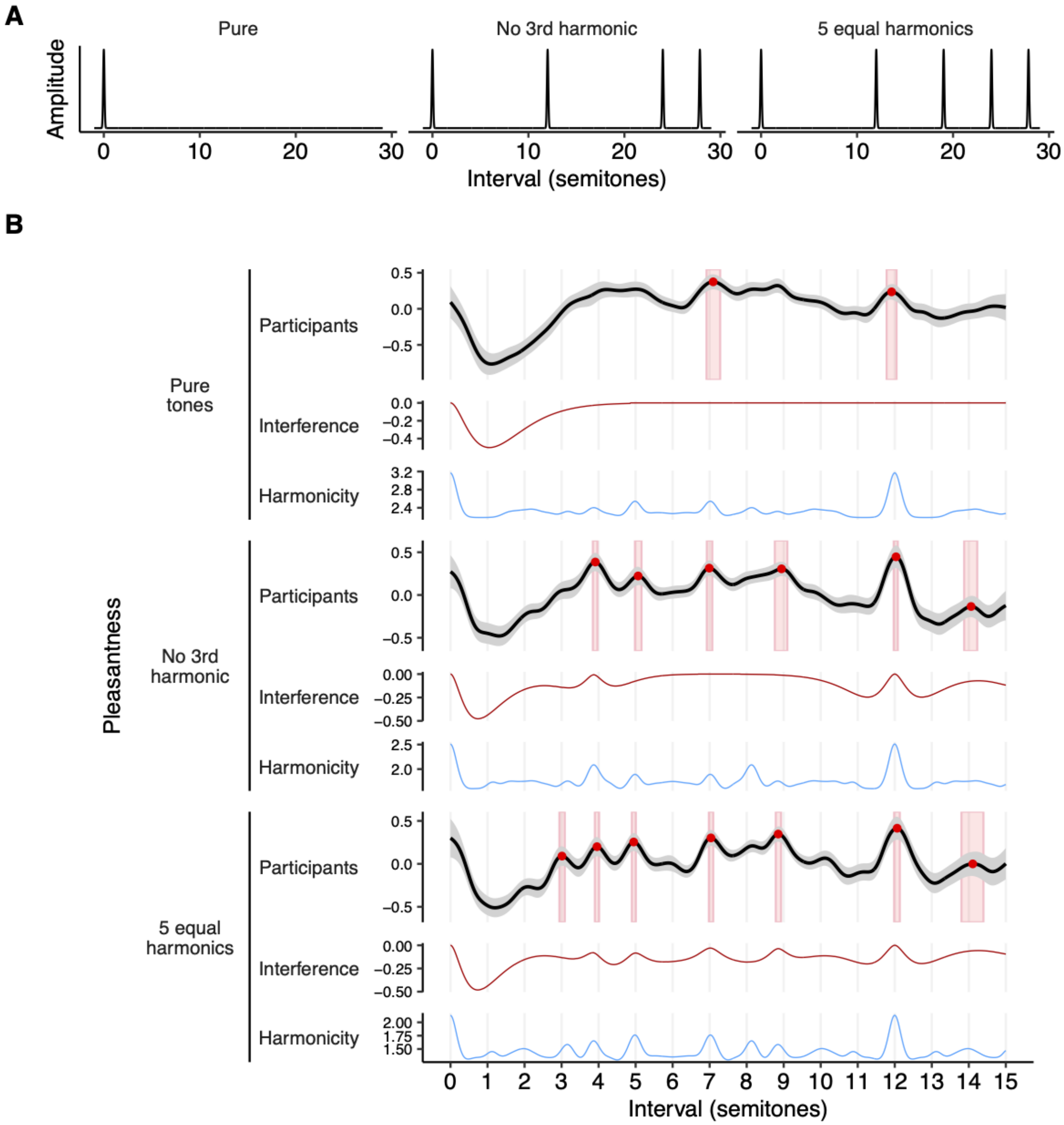


Figure 7. Harmonic deletion and consonance (Study 4A, 485 participants).
(**A**) Tone spectra used in Study 4A. (**B**) Pleasantness judgments (95% confidence intervals) for dyads using these tone spectra, plotted alongside interference (Hutchinson & Knopoff, 1978) and harmonicity (Harrison & Pearce, 2018) model predictions (Study 4A). Peak locations are plotted as red circles with 95% confidence intervals in red rectangles.

We see a clear effect of the manipulation: deleting harmonics reduces the number of peaks in the pleasantness profiles, producing smoother and less differentiated curves (Figure 7B; see *Supplementary Materials* for video versions). In particular, the peak-picking algorithm identifies seven statistically reliable peaks for the full spectrum (minor third, major third, perfect fourth, perfect fifth, major sixth, octave, and major tenth); deleting the third harmonic causes the minor third peak to be lost, and deleting the remaining upper harmonics causes all but the perfect fifth and octave peaks to be lost.

The interference models correctly predict that peaks will be lost by deleting harmonics, but their predictions about specific peaks are often inaccurate. For example, they fail to predict any minor-third peak for any tone type, yet the behavioral data clearly exhibits such a peak for the five-harmonic tones. Interestingly, the Harrison-Pearce harmonicity model successfully predicts this peak, and it successfully predicts that this peak will disappear once the third harmonic is removed; however, the other two harmonicity models fail to make the same prediction (see *Supplementary Materials*). It is also worth noting that the interference model predicts a complete absence of peaks for pure tones, whereas the behavioral data exhibits clear peaks at both the perfect fifth and the octave; similar erroneous predictions are made by the other two interference models (see *Supplementary Materials*).

As a follow-up question, we wondered what effect this timbral manipulation would have on preferred tunings for particular musical intervals. If listeners do indeed prefer different tunings for different musical timbres, this would imply that there is no such thing as an 'ideal' tuning system for maximizing consonance, and instead the ideal tuning system should depend on the timbres being used.

What tunings should Western listeners ordinarily consider most consonant? On the one hand, interference and harmonicity models predict that consonance should be maximized by *just-intoned* intervals, which correspond to exact simple integer ratios (e.g., 2:1 for the octave, 3:2 for the fifth, and so on). On the other hand, Western listeners typically have substantial experience with 12-tone equal temperament, the most common tuning system in Western music. Therefore, we might expect them to consider *equal-tempered* intervals most consonant.

How should harmonic deletion affect these tuning preferences? Interference theories clearly predict that listeners' preferences for just-intoned intervals should be eliminated by this manipulation, on account of eliminating the beating between upper harmonics (Figure 8, red lines). Harmonicity theories meanwhile predict that peaked consonance preferences should still be possible in the absence of upper harmonics; however, in practice certain harmonicity models do predict that consonance preferences will be somewhat flattened by deleting upper harmonics (Figure 8, blue lines).

We examined these possibilities in Study 4B (1,341 participants). In order to maximize statistical power, we focus on just two tone types: (a) harmonic complex tones with low spectral roll-off (3 dB/octave), and (b) pure tones (corresponding to infinite spectral roll-off). We collect ratings in the neighborhood of several prototypical consonances from Western music theory: the major third (5:4), the major sixth (5:3), and the octave (2:1). We chose the major third and the major sixth because they have significantly different tunings in just intonation versus equal temperament, which is helpful for distinguishing the candidate theories. We additionally chose the octave because of previous work indicating listener preferences for stretched octaves (e.g., Dobbins & Cuddy, 1982; McKinney & Delgutte, 1999; Ohgushi, 1983).

The results for harmonic tones were considerably more interesting than we anticipated (Figure 8, solid lines; see *Supplementary Materials* for video versions). For both the major sixth and the octave, we see two clear peaks that sit either side of the just-intoned interval (8.78 [8.77, 8.80] and 8.93 [8.92, 8.94] for the major sixth, and 11.94 [11.92, 11.96] and 12.08 [12.07, 12.1] for the octave). In neither case do these peaks overlap with the equal-tempered interval. For the major third, we see a less clear version: the peak on the sharp side of just intonation is clear (3.95 [3.93, 3.96]), but the equivalent peak on the flat side is more subtle, only being detected in 66% of the bootstrap iterations. Nonetheless, the pattern of results is clearly inconsistent with simple preferences for just intonation or equal temperament.

Why might listeners prefer slight deviations from just intonation? One explanation would be that listeners positively enjoy the slow ‘pulsating’ beats that these slight deviations induce for the feeling of ‘richness’ that they convey. This possibility was hypothesized by Hall (1973) and received some empirical support from Roberts and Mathews (1984). However, this concept is missed in the modeling literature; current interference models assume that beating is universally disliked (Hutchinson & Knopoff, 1978; Sethares, 1993; Vassilakis, 2001).

If these preference patterns are indeed explained by enjoyment of slow beats, then they should be eliminated by deleting the upper harmonics from the complex tones to produce pure tones. Indeed, this is what we see (Figure 8, dashed lines): the preference for slight deviations from just intonation is eliminated, resulting in relatively flat consonance curves.

To summarize: Studies 4A and 4B demonstrate that deleting upper harmonics substantively influences consonance judgments. The pattern of effects in Study 4A (intervals ranging from

0 to 15 semitones) is predicted fairly well by the harmonicity model but not by the interference model. The pattern of effects in Study 4B (tuning preferences) is meanwhile not directly explained either by existing interference or harmonicity models; however, they do seem to be explainable by resorting to a 'slow beats' theory that could theoretically be incorporated into interference models.

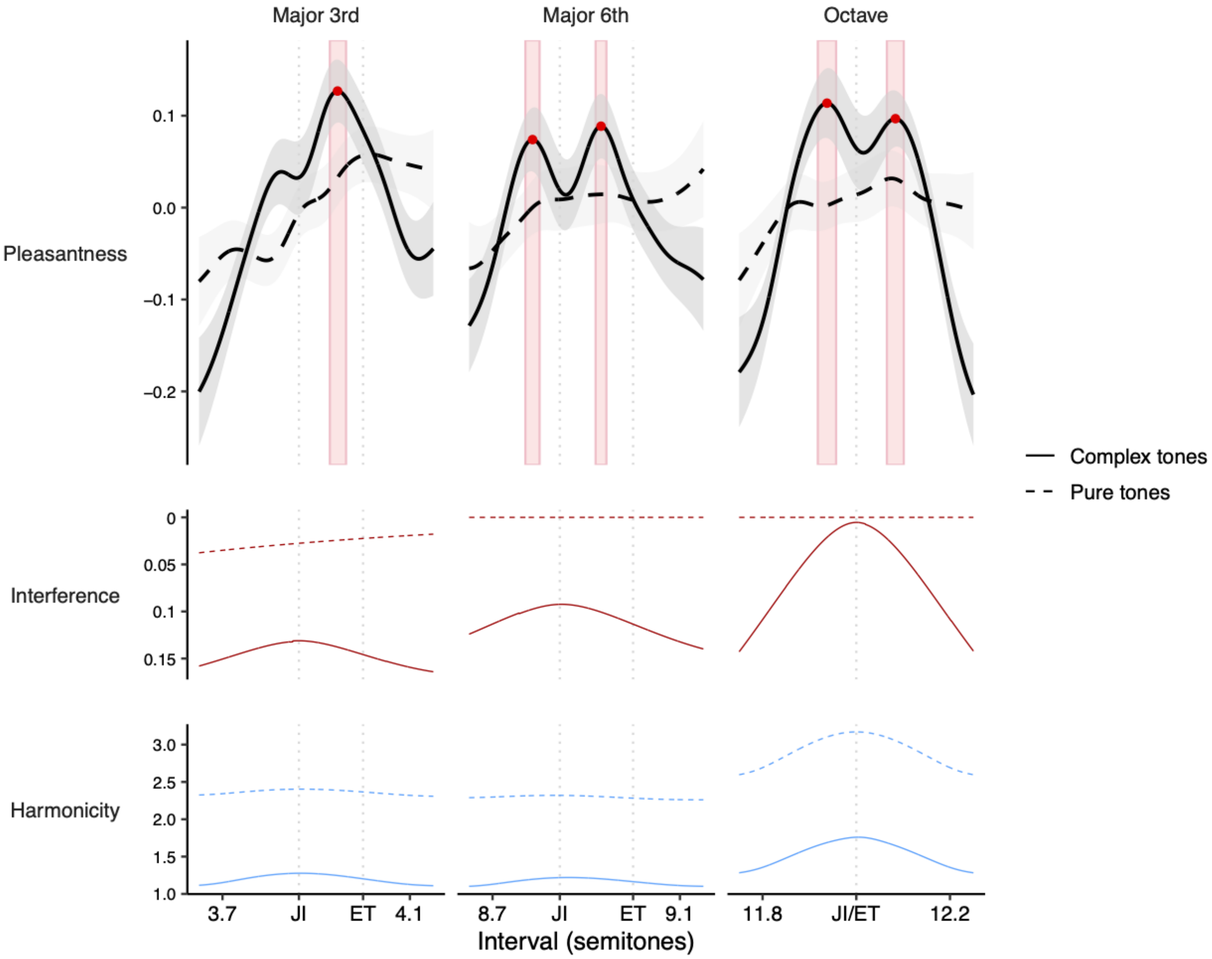


Figure 8. Dyadic preferences for (mis)tunings of the major 3rd, major 6th, and the octave (Study 4B, 1341 participants). Pleasantness judgments (*z*-scored within participants, 95% confidence intervals) are plotted alongside interference (Hutchinson & Knopoff, 1978) and harmonicity (Harrison & Pearce, 2018) model predictions. Note that the between-groups design means that judgments for complex tones and pure tones are *z*-scored independently (but always within participants). Peak locations are plotted as red circles, with 95% confidence intervals for these peak locations (bootstrapped) plotted as red rectangles. Just-intoned and equal-tempered versions of each interval are marked with 'JI' and 'ET' respectively.

Generalizing to triads (Study 5)

The previous studies definitively show that spectral manipulations can affect consonance perception for two-note chords (dyads). However, much of Western music is built from chords comprising at least three notes (triads, tetrads, etc.). It is worth asking whether these spectral effects can generalize to these larger chords.

The dense rating techniques used in the previous studies work well for the one-dimensional domain of dyadic intervals, but they scale less well to higher-dimensional chords such as triads. As the number of dimensions increases, the behavioral ratings are spread out over increasingly wider spaces, making the local averages at any one point less and less reliable.

Here we therefore use an alternative method called GSP (Harrison et al., 2020). This method coordinates participants into collaboratively exploring the stimulus space to find regions of (in this case) high consonance. In our application the stimulus space is two-dimensional, and corresponds to a space of possible triads. Each trial begins at a point on this plane, with the participant being presented with a slider corresponding to either horizontal or vertical motion in the plane. Moving and releasing the slider triggers a new chord to be played corresponding to the updated position in the plane. The participant is told to move the slider to maximize the chord's pleasantness (Figure 9A); when they are finished, the chord is then passed to next participant, who then manipulates the other interval and passes the chord to the next participant, and so on for a chosen number of iterations (typically 40) (Figure 9B). Trials from many participant chains starting at many different points are averaged using a *kernel density estimator*.

Figure 9C shows baseline GSP results for harmonic dyads with 3 dB/octave spectral roll-off (Study 5A, 228 US participants). Analogous to the prior results for dyads (Figure 1), the present results now reflect the Western consonance hierarchy for triads. Some of the structure is inherited directly from the dyadic profile; for example, we see a clear diagonal line corresponding to chords whose intervals sum to 12 semitones (e.g., [5, 7], [7, 5], where the two numbers correspond to the intervals between the lower two notes and the upper two notes respectively). We also see hotspots corresponding to prototypical triadic sonorities, especially the three inversions of the major triad ([4, 3], [3, 5], [5, 4]). Looking specifically at locations corresponding to the Western 12-tone scale, we further find that our results correlate well with the most comprehensive available reference dataset (Bowling et al., 2018) ($r = .73$, 95% CI: [.57, .83]). Additionally, a Monte Carlo split-half correlation analysis showed an excellent internal reliability ($r = .93$, 95% CI: [.91, .96], 1,000 permutations).

Figure 10 shows GSP results for stretched and compressed tones (Study 5B, 462 participants). Similar to the dyad results (Study 2A), we see that these tones elicit stretched and compressed consonance profiles respectively. This effect is particularly visually prominent in the case of the *octave diagonal*, a line running from the middle top to the middle right of the consonance plot corresponding to chords whose lower and upper tones are separated by an octave. For harmonic tones, this diagonal is located at 12.04 [11.87, 12.21] semitones; for stretched tones, the diagonal shifts to 12.80 [12.76, 12.83] semitones, whereas for compressed tones it shifts to 11.20 [11.15, 11.25] semitones. As before, these results are clearly predicted by the interference model but not by the harmonicity model. In summary, Study 5 clearly replicates the results of Study 2A, but generalizes them from dyads to triads.

Discussion

In this paper we sought to understand how consonance perception depends on the spectral properties of the underlying chord tones. We used a pair of novel psychological paradigms (dense rating, GSP) to measure consonance judgments for continuous intervallic spaces (Study 1), systematically varying the spectra of the underlying chord tones (Studies 2A-5), and interpreting the results using computational models of interference and harmonicity (Harrison & Pearce, 2018; Hutchinson & Knopoff, 1978). Our results show that these spectral manipulations do indeed influence consonance perception in a deep way, with important implications for our understanding of its underlying mechanisms.

Studies 2A and 2B investigated the effect of changing harmonic *frequencies*. We found that such manipulations can induce inharmonic consonance profiles, contrasting with the vast majority of previous consonance research, which consistently demonstrates either preferences for harmonic intervals (Bowling et al., 2018; Johnson-Laird et al., 2012; e.g. McDermott et al., 2010; Schwartz et al., 2003) or (for some populations) an absence of any preferences (McDermott et al., 2016). In particular, Study 2A showed that stretching or compressing tone spectra induced stretched and compressed consonance profiles respectively; this finding is particularly interesting because it is predicted by current interference but not harmonicity models of consonance. Study 2B subsequently showed that tone spectra modeled after the Javanese bonang also yield an inharmonic consonance profile; interestingly, as hypothesized by Sethares (2005), this consonance profile aligns with an idealized slendro scale as used in the Javanese gamelan. These results provide an empirical foundation for the idea that cultural variation in scale systems might in part be driven by the spectral properties of the musical instruments used by these different cultures (Li, 2006;

Schneider, 2001; Sethares, 2005). They also provide an empirical justification for certain practices in the experimental music tradition of 'Dynamic Tonality' (Plamondon et al., 2009), where tone spectra and scale tunings are manipulated in tandem (see also Milne et al., 2016).

Study 3 investigated the effect of changing harmonic *amplitudes*. In particular, we studied a *spectral roll-off* manipulation, changing the rate at which amplitude decays as a function of harmonic number. Interference models predict that increasing roll-off should gradually 'flatten' the consonance profile, such that all intervals between four and ten semitones become similarly consonant. In contrast, harmonicity models predict that the consonance profile should be relatively robust to this manipulation. In actuality, the empirical data displayed no such flattening, only a general preference for higher roll-off, consistent with the harmonicity model.

Studies 4A and 4B investigated the effect of deleting harmonics *entirely*. First, Study 4A studied effects of harmonic deletion in the context of intervals spanning 0-15 semitones. We found clear effects of these manipulation, corresponding to the disappearance of certain peaks from the consonance profile. The precise pattern of peak elimination seemed best predicted by the harmonicity rather than the interference models. Second, Study 4B studied harmonic deletion in the context of fine-grained preferences for different tunings of particular consonant intervals. We identified a subtle but interesting phenomenon for tones with strong upper harmonics, whereby listeners prefer slight deviations from exact just intonation, as hypothesized by Hall (1973). This effect seems explainable by listeners enjoying slow beats, a phenomenon that could in theory be straightforwardly incorporated into interference models of consonance perception. Consistent with this hypothesis, these preferences for slight deviations disappeared upon elimination of the upper harmonics, presumably because this eliminates the slow beating effect.

Study 5 then investigated whether spectral manipulations could also affect consonance profiles for chords comprising more notes. In particular, we tested the stretching/compressing manipulation from Study 2A, and applied it to triads using the adaptive technique GSP (Harrison et al., 2020). We successfully replicated the stretching/compressing effect from Study 2A, showing that stretched/compressed tones yielded preferences for stretched/compressed chords respectively.

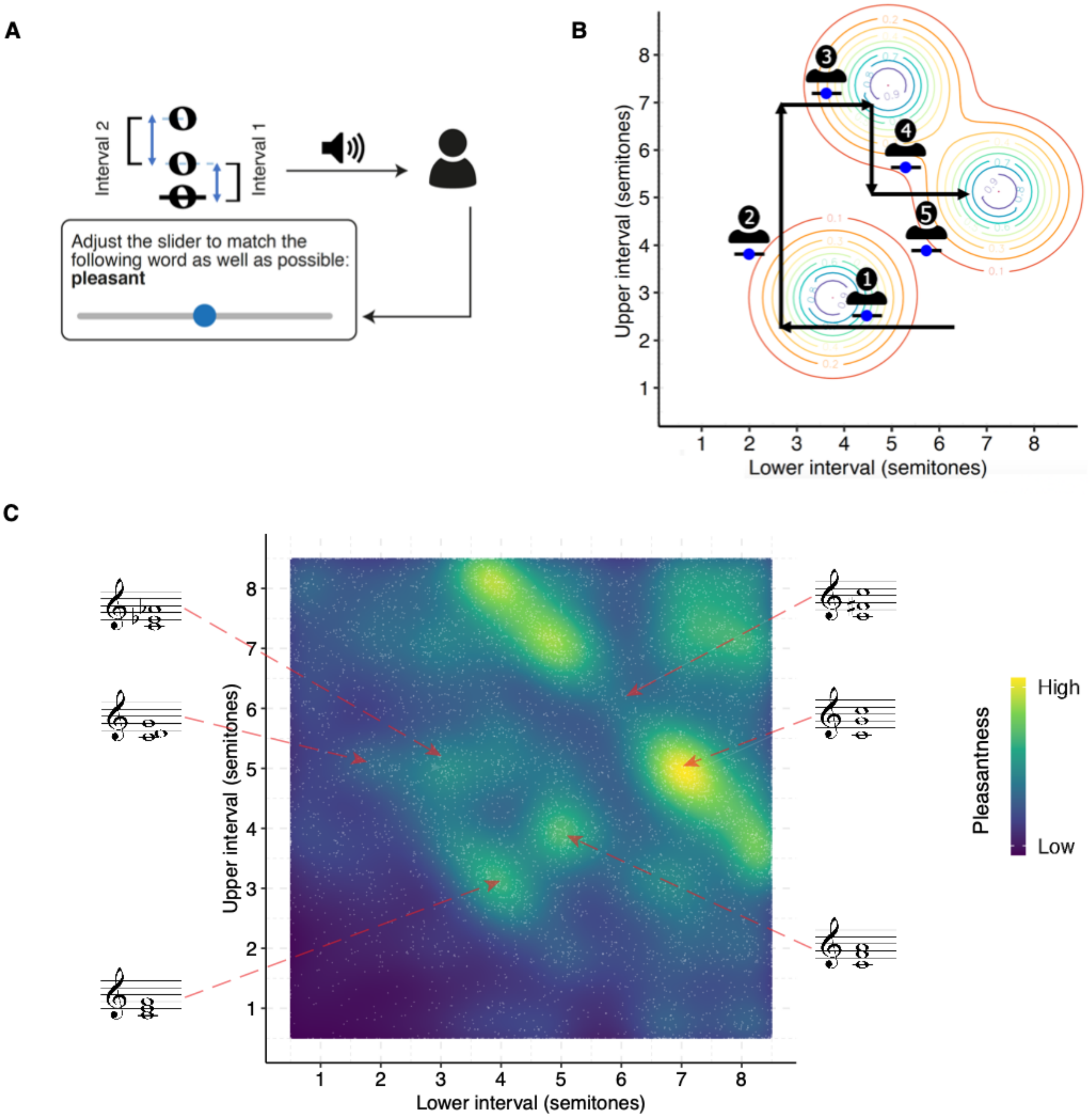


Figure 9. Gibbs Sampling with People (GSP). (**A**) Schematic illustration of the GSP task. (**B**) Example trajectory from a GSP chain overlaid on a kernel density estimate. (**C**) GSP data for harmonic triads (Study 5A, 228 participants). Individual iterations are plotted as white dots, whereas the KDE (bandwidth = 0.375) is plotted as a heat map, with light areas corresponding to high density/consonance.

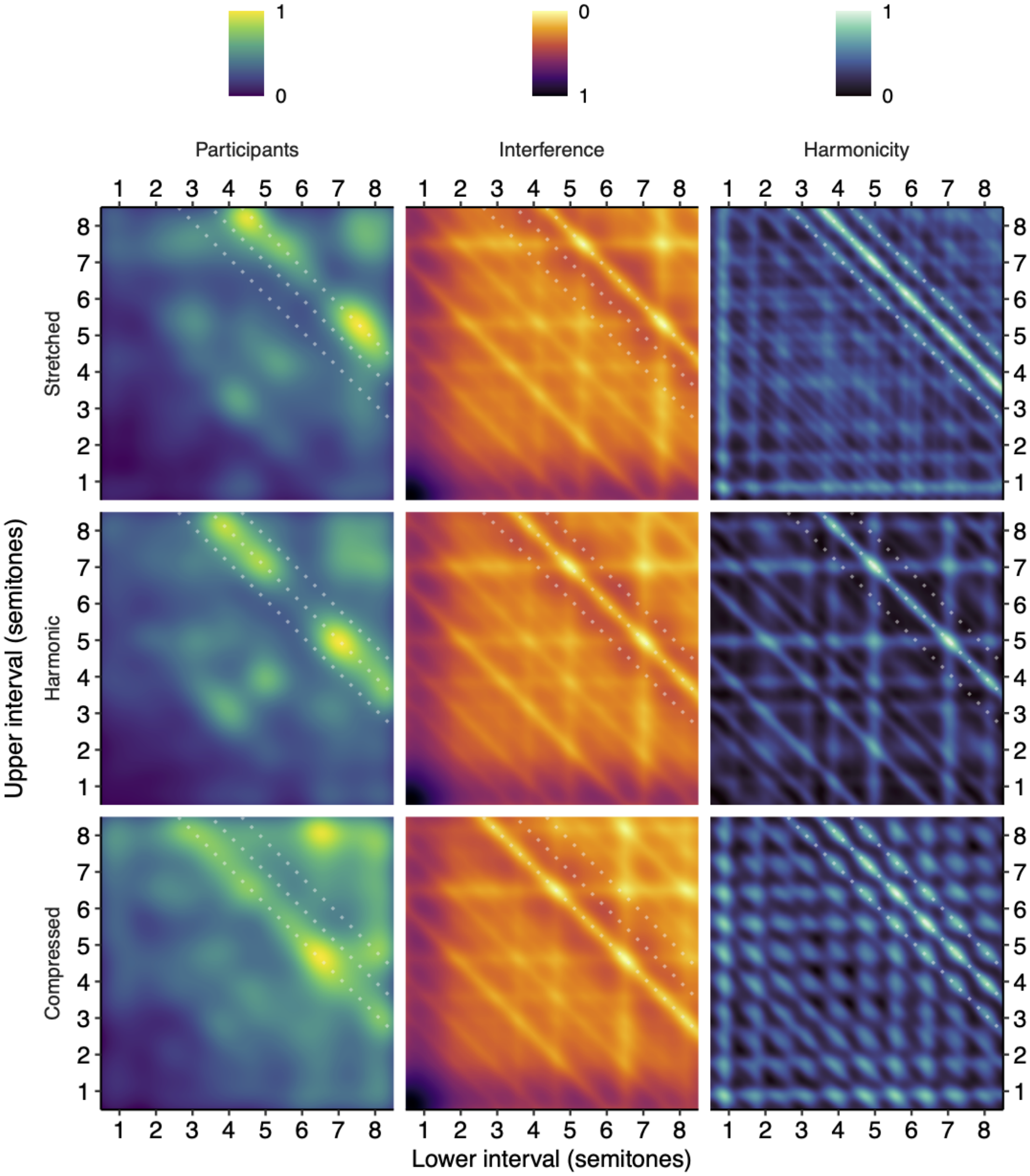


Figure 10. Triadic pleasantness judgments for stretched and compressed tones (Study 5B, 690 participants). GSP results are summarized using a KDE with a bandwidth of 0.375 semitones, and plotted alongside interference (Hutchinson & Knopoff, 1978) and harmonicity (Harrison & Pearce, 2018) model predictions. Analogous pleasantness judgments for harmonic tones (Study 5A) are also included as a reference. The theoretical locations of the compressed, harmonic, and stretched octave diagonals are plotted as dotted white lines. Values within each panel are scaled to the unit interval [0, 1].

Together, these results clearly demonstrate that the consonance and dissonance phenomena documented by Western music theory are *not* universally determined by frequency ratios between chord tones. Rather, these phenomena depend integrally on the spectral properties of the chord tones, as hypothesized by Helmholtz (1875) as well as Plomp and Levelt (1965) but seemingly disproved by later empirical studies (McLachlan et al., 2013; Nordmark & Fahlén, 1988). This is particularly interesting in the context of explaining the inharmonic scale systems used in the Javanese gamelan; in particular, our results support the hypothesis of Sethares (2005) that the slendro scale may have evolved to reflect the consonance profile induced by the inharmonic bonang.

Our results further reveal limitations in our traditional understanding of the consonance of standard harmonic tones. In particular, Study 4B showed that listeners do not prefer intervals to be *exactly* just-intoned, but rather prefer slight deviations from just intonation. This phenomenon has interesting consequences for evaluating musical tuning systems (e.g., quarter-comma meantone; 12-tone equal temperament); the perfectly pure intervals of just intonation cease being a gold standard, and a certain amount of interval impurity becomes positively desirable (see also Hall, 1973).

The combined results also have important implications for competing theories of consonance perception (Table 1). Unitary explanations -- for example, that consonance is completely due to interference between partials (Plomp & Levelt, 1965), or completely due to harmonicity perception (McDermott et al., 2010) -- seem untenable in the context of these results. The stretching/compressing results from Studies 2A and 5B cannot be explained by current harmonicity models, only by interference models; the preferences for slight deviations from just intonation in Study 4B seem best explained by interference effects; conversely, the effects of deleting individual harmonics in Study 3 are best explained by harmonicity modeling, and the robustness to spectral roll-off demonstrated in Study 4 is consistent with an underlying harmonicity mechanism. Overall, it would seem plausible that consonance perception in Western listeners derives from a combination of (negatively valenced) interference and (positively valenced) harmonicity, which is likely complemented and reinforced by listeners' recognition of culturally familiar chord types (Harrison & Pearce, 2020).

However, there is an obstacle to this 'composite' interpretation in Study 3: here, increasing spectral roll-off had no impact on the peakiness of the consonance profile, in contrast with the predictions of the interference model. If interference truly contributes to consonance perception, then why doesn't the consonance profile become flatter with increased roll-off?

An answer is suggested by Study 4A, where we see that eliminating upper harmonics entirely *does* flatten the consonance profile to a significant extent. This is interesting, because according to the interference model there should be essentially no difference between tones with 12 dB/octave roll-off and pure tones (see *Supplementary Materials*). This suggests a potential explanation for the results of Study 3: that existing interference models underestimate the contribution of low-amplitude partials to dissonance perception. If the interference models were updated to upweight the contribution of such partials, then the discrepancy between Studies 3 and 4A would trivially disappear, and all the results would be consistent with a composite theory of consonance. This possibility needs to be investigated carefully alongside an analysis of potential mechanisms for such an effect (see Vassilakis, 2001 for a discussion of related issues).

Table 1

Summary of empirical results and their theoretical implications

Study	Description	Theory	
		Interference	Harmonicity
1, 5A	Harmonic tones induce harmonic consonance profiles.	✓	✓
2A, 5B	Stretched/compressed tones induce stretched/compressed consonance profiles.	✓	✗
2B	The synthetic bonang elicits an idiosyncratic inharmonic consonance profile.	✓	(✓)
3	Consonance profiles are robust to spectral roll-off.	✗	✓
4A	Deleting harmonics causes particular peaks to disappear from the consonance profile.	(✓)	✓
4B	Deleting harmonics eliminates listener preferences for slight deviations from just intonation.	(✓)	✗

Note. Theories are marked as successful (✓) or unsuccessful (✗) to the extent that their models successfully predict the empirical data, or could clearly do so after well-motivated extensions.

Our results suggest some other model improvements too. The preferences for slight deviations for just intonation observed in Study 4B could be captured by modifying interference models to incorporate a liking of slow beats. Moreover, the stretching/compressing effects observed in Studies 2A and 5B might be captured by updating the template-matching algorithms to incorporate the possibility of template stretching/compressing. Our datasets, included as supplementary material to this paper, should provide a useful crucible for developing and evaluating such improvements.

One interesting manipulation that we did not study here is dichotic versus diotic presentation. Several papers have argued that presenting dyads dichotically (i.e., one tone to the left ear, and the other tone to the right ear) eliminates interference effects (e.g., roughness), because the two tones no longer interact in the auditory periphery (Bidelman & Krishnan, 2009; McDermott et al., 2010); comparing dichotic and diotic consonance profiles should therefore isolate the contribution of interference to consonance. Our methods should generalize well to this dichotic/diotic manipulation.

Consonance perception is known to vary between individuals, even when the individuals are drawn from a relatively homogeneous cultural group (McDermott et al., 2010; Popescu et al., 2019). Here we focused on consonance perception at the population level, but in the *Supplementary Materials* we break down the results by musicianship. We found that the results were generally similar across the participant groups, but a systematic investigation might well yield more interesting conclusions. An interesting future path is to step outside the traditional musician versus nonmusician dichotomy, and instead use unsupervised clustering methods to identify population subgroups with different response strategies (e.g., Pearce et al., 2010).

Cross-cultural studies are important for identifying the universality of these phenomena and the potential moderating role of cultural exposure. On the one hand, the prominence of harmonic scales across different musical cultures suggests that humans are somewhat predisposed to prefer harmonic intervals (Gill & Purves, 2009); however, case studies of particular musical traditions indicate that these preferences are by no means inevitable (Ambrazevičius, 2017; Florian, 1981; McDermott et al., 2016; Vassilakis, 2005; Vyčinienė, 2002). We see a need for additional careful case studies of consonance perception in individual musical cultures (Butler & Daston, 1968; Lahdelma et al., 2021; Maher, 1976; McDermott et al., 2016) complemented by more global studies assessing the statistical prevalence of these phenomena across the world's cultures. Previous success in cross-cultural studies using rating tasks (McDermott et al., 2016; McPherson et al., 2020) and slider tasks

(Sievers et al., 2013) suggests that our dense rating and GSP paradigms should also be feasible cross-culturally.

As an initial proof of concept, we replicated the stretching/compressing dyadic consonance study (Study 2A) using Korean participants ($N = 68$) recruited with the aid of a research assistant in the local area (Study 6). Participants were required to be native Korean speakers and resident in South Korea; all experiment instructions were translated to Korean by a native speaker. The general effect clearly replicates with the new participant group: stretched spectra produce a stretched consonance profile, whereas compressed spectra produce a compressed consonance profile (Figure 11; see *Supplementary Materials* for video versions). However, there are certain interesting divergences between the results that already hint at cultural variation in pleasantness judgments. For example, the Korean participants display a general dislike of large intervals for the stretched and compressed tones, manifested as descending slopes (stretched tones: -0.84 [-1.10, -0.58] SD/octave; compressed tones: -0.75 [-0.97, -0.54] SD/octave; bootstrapped over participants), which is not matched in the US participants (stretched tones: -0.17 [-0.35, 0.01] SD/octave; compressed tones: -0.04 [-0.22, 0.13] SD/octave). These kinds of cultural discrepancies deserve to be explored systematically in future work.

An appealing property of the dense rating and GSP paradigms is that they free the experimenter from having to specify particular chords to study *a priori*. Here we then chose particular tone spectra that we expected to elicit interesting differences in consonance profiles. However, GSP can also support fully exploratory studies, where participants adjust both intervals and tone spectra in order to match certain adjectives. Figure 12 plots results from a proof-of-concept experiment applying this method to the adjectives ‘pleasant’, ‘bright’, and ‘dark’ (Study 7, 394 participants). Since individual harmonics have relatively subtle perceptual effects, we used *aggregated* GSP, where the slider responses of multiple participants are combined before progressing to the next GSP iteration (Figure 12A; see *Methods* for details). The derived prototypes are plotted in Figures 12B (spectral weights) and 12C (pitch intervals). Compared to the reference conditions ‘bright’ and ‘dark’, the ‘pleasant’ condition displays a moderate amount of energy in the upper harmonics, and a relatively complex intervallic space corresponding to a ‘cleaned’ version of the non-aggregated GSP experiment conducted earlier (Study 5A). Meanwhile, brightness is associated with large intervals and high energy in the upper harmonics; conversely, darkness is associated with small intervals and low energy in upper harmonics.

Pleasantness, darkness, and brightness are just a few examples of many subjective concepts that listeners can associate with auditory stimuli. Here we focused on pleasantness as a near-synonym of consonance, but there are many other near-synonyms of consonance that might be studied (Lahdelma & Eerola, 2020; van de Geer et al., 1962), and a myriad of other connotations ranging from the emotional to the physical that might also be studied (e.g., Knöferle & Spence, 2012; Palmer et al., 2013; Zacharakis et al., 2014). Our dense rating and GSP methods should generalize well to these different domains, and we look forward to what findings they might uncover.

Taken together, this work demonstrates how large-scale online experiments can contribute to the study of classical questions in psychology, psychophysics, and auditory perception. Such experiments enable us to test significantly more conditions than in most previous studies (24 experiments), while also allowing us to relax restrictive assumptions from previous studies (e.g., discrete scale systems) and providing results that are novel yet consistent with previous research. Large-scale online experiments also facilitate new generations of advanced experimental paradigms such as GSP, which are able to explore high-dimensional stimulus spaces in audition such as chord sequences, melodies, and timbres.

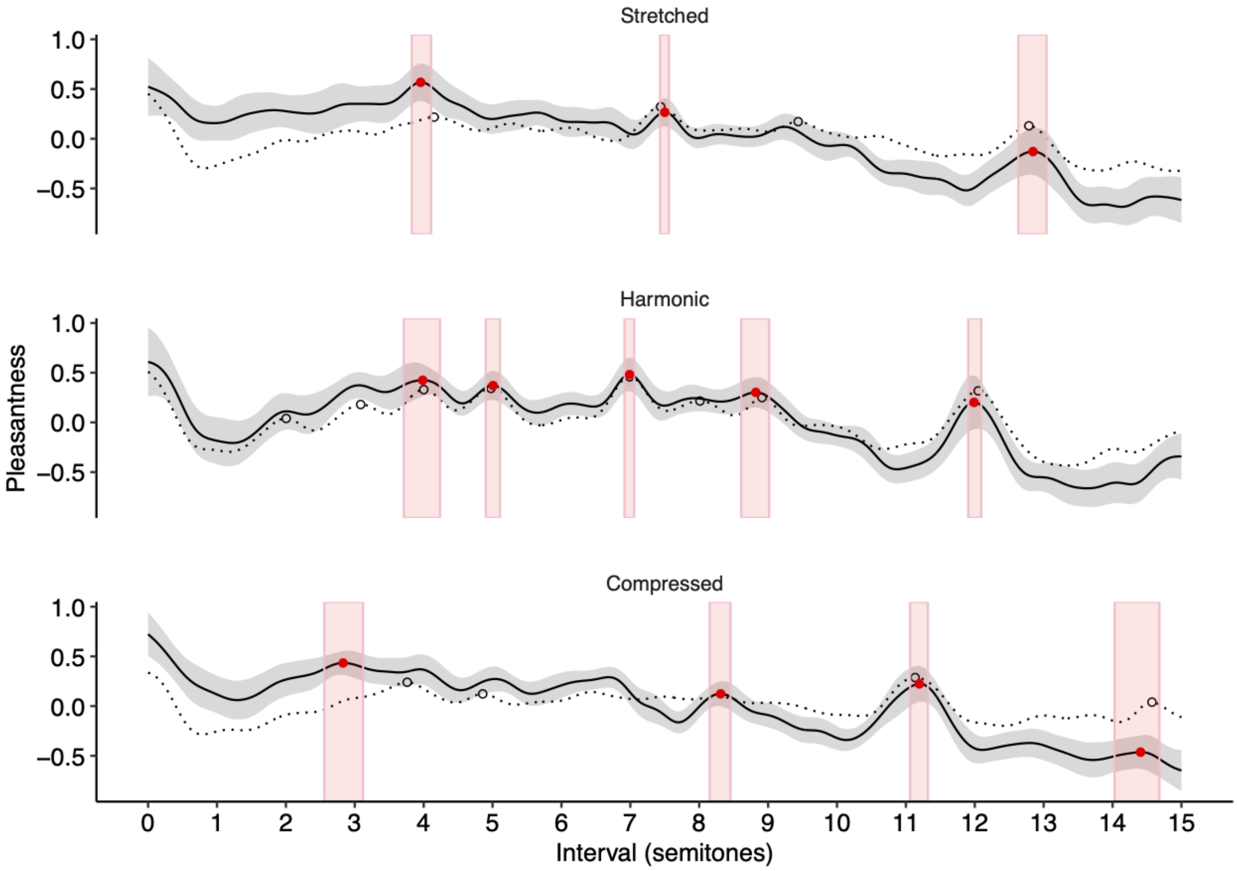


Figure 11. Dyadic consonance as a function of spectral stretching for Korean participants (Study 6). Consonance profiles (*z*-scored) are plotted with 95% bootstrapped confidence intervals (bandwidth of 0.2 semitones). Equivalent consonance profiles for US participants (Study 2A) are plotted as reference dotted lines.

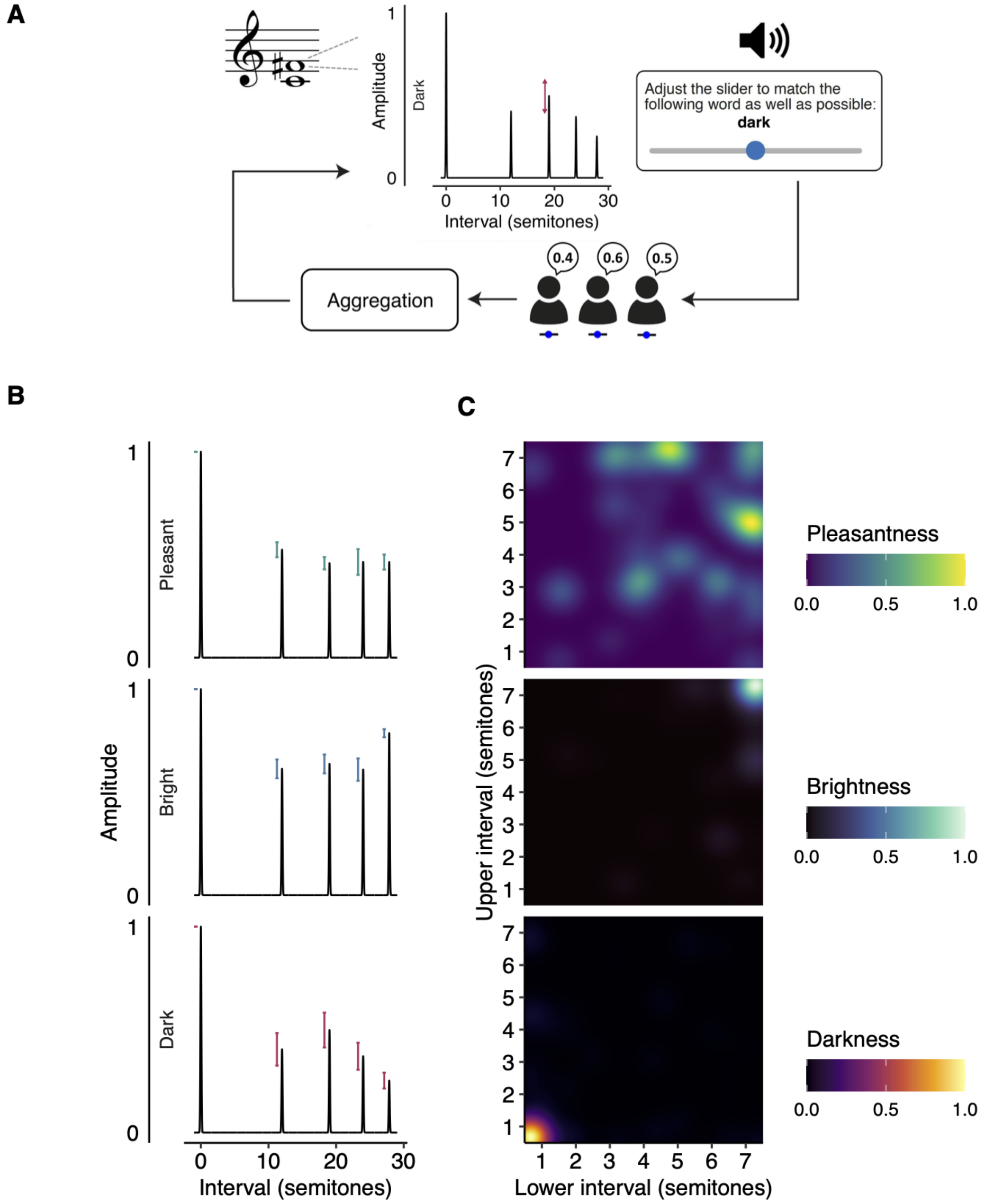


Figure 12. Joint interval and timbre GSP. (**A**) Schematic illustration of the timbre GSP task. (**B**) Average preferred tone spectra and interval KDEs for three adjectives: pleasant, bright, dark (Study 7, 394 participants). Error bars indicate 95% bootstrapped confidence intervals.

Methods

Paradigms

We use two behavioral paradigms in this paper: *dense rating* and *GSP*. The special feature of these paradigms is that they do not make any *a priori* assumptions about culturally specific scale systems, but instead characterize consonance as a smooth function over continuous space.

Dense rating

In the dense rating paradigm, participants are played chords that are randomly and uniformly sampled from continuous intervallic space. For *dyads* (chords comprising two tones), we typically study intervals in the range [0, 15] semitones; in successive trials we might therefore see dyads such as 4.87 semitones, 12.32 semitones, or 1.83 semitones. Each trial receives a pleasantness rating on a scale from 1 (completely disagree) to 7 (completely agree) (Figure 1A). We then summarize the results from all the trials using a *Gaussian kernel smoother*. The degree of smoothing is set by the *bandwidth* parameter, which can be set by the experimenter in order to achieve an arbitrary balance between bias and variance: decreasing the bandwidth allows the smoother to capture more detail (less bias) at the cost of lower reliability (higher variance). To help with the interpretability of the data, we use a single bandwidth parameter for all experiments. In particular, we use a bandwidth of 0.2 semitones for all experiments spanning 0-15 semitones, and a bandwidth of 0.035 semitones for experiments spanning 0.5 semitones. We verify the appropriateness of these bandwidths by computing Monte Carlo split-half correlations (1,000 replicates) for two reference datasets (Study 1: harmonic dyads; Study 4B: major 3rd with 3 dB/octave roll-off). The results indicate excellent reliability in both cases (harmonic dyads: $r = .87$, 95% CI: [.74, .94]; major third: $r = .93$, 95% CI: [.82, .99]). We compute 95% confidence intervals for the smoothed ratings through nonparametric bootstrapping over participants; for computational tractability, we approximate these by computing bootstrapped standard errors ($N = 1{,}000$ replicates) and multiplying by 1.96 (Gaussian approximation). To facilitate interpretation, we also estimate peaks of the consonance curves using a custom peak-picking algorithm, and compute confidence intervals for these peaks using the same bootstrapping algorithm (see *Methods* for details).

GSP

Dense rating paradigms scale poorly to higher dimensions on account of the *curse of dimensionality* (e.g., Hastie et al., 2001), which makes exhaustive sampling of the stimulus space impractical. For studying consonance in *triads* (chords comprising three tones), we therefore use Gibbs Sampling with People (GSP), a recent technique designed to tackle this dimensionality problem (Harrison et al., 2020). We parametrize the space of triads as pairs of intervals, building cumulatively from the bass note, meaning that (for example) a major triad is represented as the tuple (4, 3). The space of possible triads can then be represented as a two-dimensional (2D) plane. Each trial begins at a point on this plane, with the participant being presented with a slider corresponding to either horizontal or vertical motion in the plane. Moving and releasing the slider triggers a new chord to be played corresponding to the updated position in the plane. The participant is told to move the slider to maximize the chord's consonance (Figure 9A); when they are finished, the chord is then passed to next participant, who then manipulates the other interval and passes the chord to the next participant, and so on for a chosen number of iterations (typically 40) (Figure 9B). This process can be modeled as a *Gibbs Sampling* process, well-known in computational statistics (Harrison et al., 2020). Following this model, we can estimate consonance for the full 2D space by applying a *kernel density estimator* (KDE) over the locations of the different trials generated by the process with a fixed bandwidth of 0.375 (see methods). As before, the kernel bandwidth parametrizes a trade-off between bias and variance; here we chose a bandwidth of 0.375 semitones, and verified it as before with a Monte Carlo split-half correlation analysis (1,000 replicates), which demonstrated excellent reliability ($r = .93$, 95% CI: [.91, .96]).

Software implementation

All experiments were implemented using PsyNet (https://www.psynet.dev), our under-development framework for complex experiment design (Harrison et al., 2020). This framework builds on Dallinger, a platform for experiment hosting and deployment (Figure 13). Participants engage with the experiment through a front-end interface displayed in the web browser, which communicates with a back-end Python server cluster that organizes the experiment timeline. The cluster is managed by the web service Heroku which orchestrates a collection of virtual instances that share the experiment management workload, as well as an encrypted Postgres database instance for data storage. Code for the implemented experiments can be found in the *Supplementary Materials*.

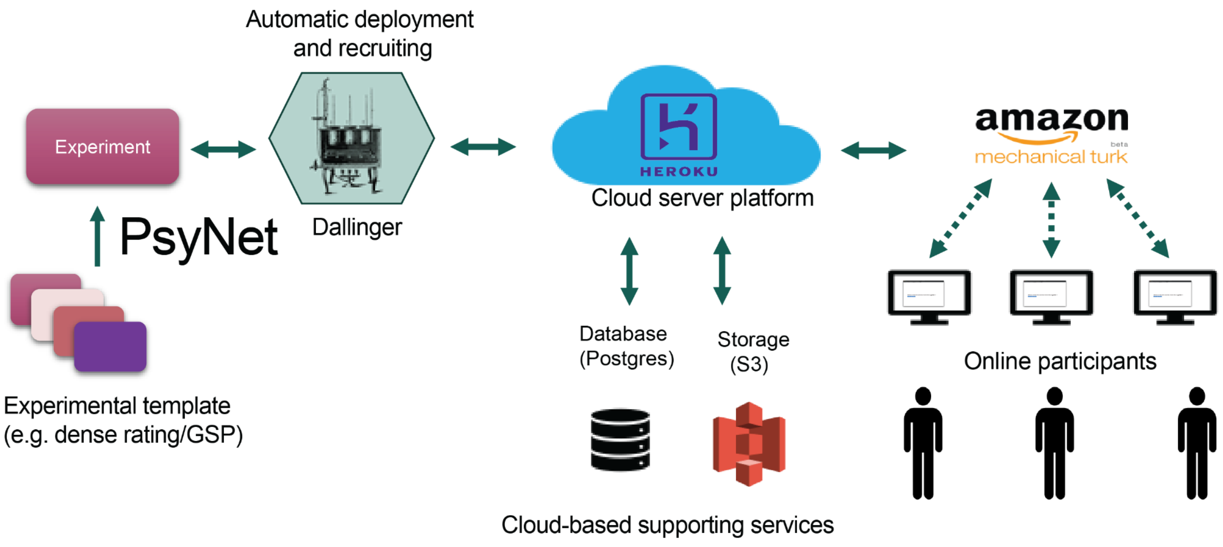


Figure 13. Schematic of the data collection infrastructure. Reproduced from Harrison et al. (2020) with minor modifications.

Stimuli

Each of our chords can be expressed as a collection of absolute pitches. We express absolute pitches using 'MIDI' notation, which maps each note of the 12-tone piano keyboard to a positive integer. Concert A (A4, 440 Hz) is by convention mapped to the value 69. Adjacent integers in this scale then correspond to adjacent semitones. Formally, the mapping is expressed as follows;

$$f = 440 \times 2^{(p-69)/12}$$

where p is the MIDI pitch and f is a frequency measured in Hz. An equal-tempered major triad rooted on middle C can then be expressed as the following tuple: $[60,\ 64,\ 67]$. This notation approach is useful for capturing the logarithmic nature of pitch perception (Jacoby et al., 2019)

In our experiments we also give chords intervallic representations. Intervallic representations are common in consonance experiments because consonance judgments are relatively insensitive to small-to-moderate changes in absolute pitch (though see Eerola & Lahdelma, 2022 for experiments exploring large changes). The intervallic representation expresses each non-bass tone as a pitch interval from the tone immediately below. Since there is no tone below the bass tone, this tone is omitted from the intervallic representation. Our experiments randomize over (dense rating) or manipulate (GSP) the intervallic representation; the bass tone is then treated as a separate parameter that is randomly sampled from a finite range on a trial-by-trial basis. For example, the dense rating procedure might generate the intervallic representation $[4.1, 2.9]$; in a given trial, this will be converted to an absolute representation of the form $[p_0, p_0 + 4.1, p_0 + 4.1 + 2.9]$, where p_0 is the randomly generated MIDI pitch of the bass tone.

In all experiments the bass tone was randomized by sampling from a uniform distribution over the MIDI pitch range 55-65 (G3-F4, 196-349 Hz) . We performed this randomization to discourage participants from perceiving systematic tonal relationships between adjacent stimuli. The chord's pitch intervals were then constrained within a particular range, depending on the experiment: $[0.5, 8.5]$ in Study 5, $[0, 15]$ in Studies 1-3, 4A, and 6, $[0.5, 7.5]$ in Study 7, and $[I_c - 0.25, I_c + 0.25]$ with $I_c = 3.9, 8.9, 12$ in Study 4B.

We synthesized stimuli using Tone.js (https://tonejs.github.io/), a Javascript library for sound synthesis in the web browser. Details of stimulus synthesis are provided below, split into the different tone types used in our experiments (see also Tables 2 and 3). Most of our experiments used tones generated through additive synthesis. Each tone generated through additive synthesis can be expressed in the following form:

$$s(t) = A \sum_{i=0}^{n_H - 1} w_i \sin(2\pi f_i t)$$

where $s(t)$ is the instantaneous amplitude and t is time. Different choices of weights w_i and frequencies f_i correspond to different tone spectra types. The parameter A represents the overall amplitude.

Type I: Harmonic tones comprise $n_H = 10$ harmonic partials with ρ dB/octave roll off. Concretely, $f_i = f_0 \times (i + 1)$ and $w_i = 10^{-\omega_i/20}$ where $\omega_i = \rho \times \log_2(i + 1)$ for $i = 0,..., n_H - 1$.

Type II: Stretched and compressed tones are identical to harmonic tones except that $f_i = f_0 \gamma^{\log_2(i+1)}$ where $\gamma = 2.1, 1.9$ for stretched and compressed tones respectively. When $\gamma = 2$ we recover standard harmonic tones.

Type III: Pure tones comprise a single frequency ($n_H = 1$), f_0 and $w_0 = 1$.

Type IV: Complex tones with/without a 3rd harmonic comprise $n_H = 5$ harmonic partials with zero spectral roll-off. Formally, we write $f_i = f_0 \times (i + 1)$ and $w_i = 1$ (type IV$_+$) or $w_{i\neq 2} = 1$ and $w_2 = 0$ for $i = 0,..., 4$ (type IV$_-$).

Type V: Bonang tones correspond to a synthetic approximation to a bonang tone, after Sethares (2005). Each tone comprises a custom complex tone given by $(f_0, 1.52f_0, 3.46f_0, 3.92f_0)$ with $w_i = 1$.

All additively synthesized timbres were presented with an ADSR envelope, comprising a linear attack segment lasting 200 ms reaching an amplitude of 1.0, a 100 ms exponential decay down to an amplitude of 0.8, a 30 ms sustain portion, and finally an exponential release portion lasting 1 s.

Some experiments additionally used a collection of more complex tones:

Type VI: Naturalistic instrument tones were based on samples from the Midi.js Soundfont database (https://github.com/gleitz/midi-js-soundfonts). The original database only provides samples for integer pitch values (i.e., 12-tone equal temperament); we therefore used the 'Sampler' tool in the Tone.js library to interpolate between these values, synthesizing arbitrary pitches by pitch-shifting the nearest sample to the required frequency. By allowing our bass tones to rove over non-integer pitch values, we ensured that integer pitch intervals were no more or less likely to exhibit pitch-shifting artifacts than non-integer intervals.

Study 7 allowed participants to create their own timbres. We call the resulting tones **Type VII tones.** These were harmonic complex tones with $n_H = 5$; participants controlled timbre by manipulating the individual weights of the upper harmonics w_i ($i = 1, 2, 3, 4$) in the range [0, 1]; w_0 (the weight of the fundamental frequency) was meanwhile fixed to 1. A naive implementation of this approach would result in changes to individual harmonics being confounded with changes of overall amplitude. To control for this we additionally allowed participants to manipulate the amplitude parameter (A), thereby dissociating timbre from volume. Since perceived loudness is approximately a logarithmic function of amplitude (Schnupp et al., 2011), we did not have participants manipulate A directly, but instead had them manipulate a log-amplitude parameter $\alpha = \log(A)$ in the range of $\alpha \in [-0.15, 0.1]$. We then report our results averaging over the amplitude paramenter.

Procedure

Dense rating

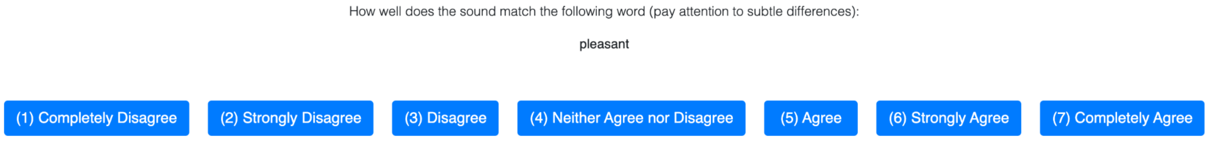
Procedure. After completing a consent form and passing the pre-screening test, participants received the following instructions:

> "In each trial of this experiment you will be presented with a word and a sound. Your task will be to judge how well the sound matches the word.
>
> You will have seven response options, ranging from 'Completely Disagree' to 'Completely Agree'. Choose the one you think is most appropriate."

This was then followed by a prompt informing participants of the response quality bonus (for further details, see *Performance incentives*). The experiment then proceeded as follows. In each trial, participants heard a sound (e.g., a dyad) and were presented with the following prompt:

> "How well does the sound match the following word (pay attention to subtle differences): **pleasant**".

Participants then delivered their judgements on a rating scale (Figure 14) ranging from 1 (Completely Disagree) to 7 (Completely Agree). Participants were assigned randomly to stimuli and the stimuli themselves were sampled uniformly from the stimulus space.



How well does the sound match the following word (pay attention to subtle differences):

pleasant

Figure 14. Experimental interface of a dense rating trial.

Gibbs Sampling with People

Procedure. After completing a consent form and passing the pre-screening test, participants received the following instructions:

> "In this experiment, you will listen to sounds by moving a slider. You will be asked to pick the sound which best represents the property in question."

To familiarize participants with the interface, each participant had to complete three training examples prior to the start of the actual experiment. These were presented with the following instructions:

> "We will now play some training examples to help you understand the format of the experiment. To be able to submit a response you must explore at least three different locations of the slider."

The training examples took an identical format to the main experiment trials, but with their intervals generated randomly rather than through a GSP process. Moreover, prior to the training trials participants were informed of the performance quality bonus (for further details, see Performance incentives).

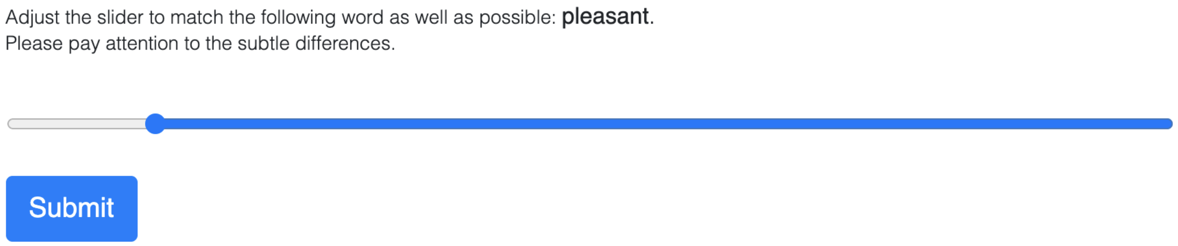
After completing the training phase, the main experiment began. Participants received the following instructions:

> "The actual experiment will begin now. Pay careful attention to the various sounds! Sometimes the differences between the sounds can be subtle, choose what seems most accurate to you. Remember: the best strategy is just to honestly report what you think sounds better! You must explore at least three locations of the slider before submitting a response. Also, sometimes the sounds might take a moment to load, so please be patient."

The experiment then proceeded as follows: in each trial, the participant was assigned to one of the available GSP chains (see below), and was provided a stimulus (e.g., a chord) and a slider (Figure 15). The slider was coupled to a particular dimension of the stimulus space (e.g., an interval) and changed from trial to trial. The participant was presented with the following prompt:

> "Adjust the slider to match the following word as well as possible: **pleasant**.
> Please pay attention to the subtle differences."

In other words, the participant was instructed to explore the slider to find the sound that was most associated with the word in question. The resulting stimulus was then passed along the GSP chain to the next participant, with each successive participant optimizing a different dimension. To circumvent any potential biases toward left or right slider directions, the direction of the slider was randomized in each trial so that in approximately half of the trials the right of the slider corresponded to bigger values, and in the other half it corresponded to smaller values.



Adjust the slider to match the following word as well as possible: pleasant.
Please pay attention to the subtle differences.

Figure 15. Experimental interface of a GSP trial.

Aggregation. Pilot studies indicated that GSP manipulations involving single harmonics were rather subtle, causing participants to produce relatively noisy responses. To compensate for this noise we adopted the *aggregation* strategy from the original GSP paper (Harrison et al., 2020)**.** In each iteration of aggregated GSP, multiple participants (here: 15) contribute one slider choice for the same stimulus, with slider starting locations randomized between participants (Figure 16). An aggregation function is then used to combine the choices into a single value that is passed to the next iteration. Our aggregation function involved constructing a kernel density estimate (KDE) of the distribution of slider values and taking the mode (Harrison et al., 2020). Unlike simpler aggregation functions (e.g., the mean), KDE aggregation is well-suited to capturing multimodal distributions.

Starting values. The starting location of each GSP chain was sampled from uniform distributions over the permissible ranges of the active stimulus dimensions.

Assigning participants to chains. The GSP process involves constructing chains of trials from multiple participants. We achieved this by applying the following process each time a participant was ready to take a new trial:

- Find all chains in the experiment that satisfy the following conditions:
    - The chain is not full (i.e., it has not yet reached the maximum allowed number of iterations);
    - The participant has not already participated in that chain;
    - The chain is not waiting for a response from any other participant.
- Randomly assign the participant to one of these chains.

Participant timeouts. Sometimes a participant will unexpectedly stop participating in the experiment. In order to prevent chains being blocked by perpetually waiting for such participants, we implemented a timeout parameter, set to 60 seconds, after which point the chain would stop waiting for the participant and instead open itself up to new participants. If the blocking participant did eventually submit a trial, they would be allowed to continue the experiment, but their trial would not contribute to the GSP chain.

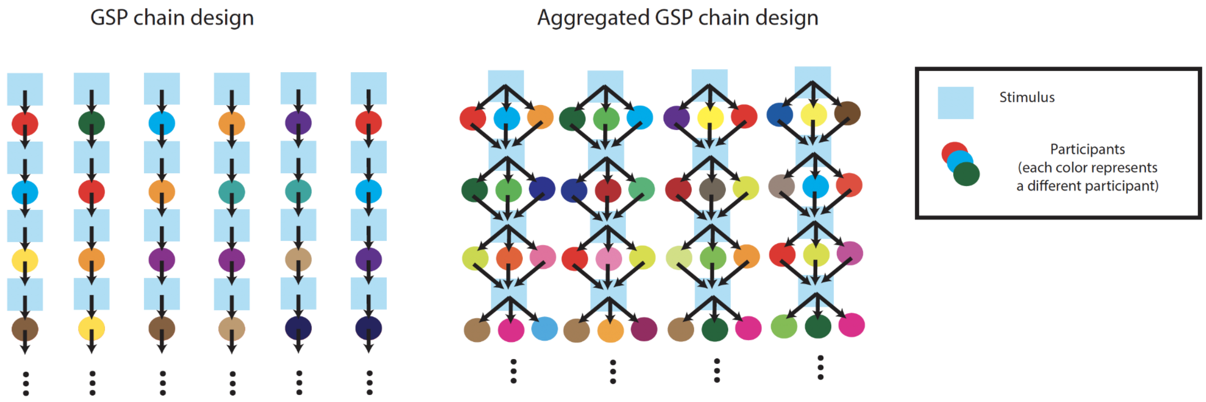


Figure 16. Regular and aggregated GSP chain designs.

Headphone screening test

To ensure high-quality listening conditions we used a previously validated headphone screening test (Woods et al., 2017). Each trial comprises a three-alternative forced-choice task where the participant is played three tones and asked to identify the quietest. The tones are designed to induce a phase cancellation effect, such that when played on loudspeakers the order of their quietness is altered, causing the participant to respond incorrectly. To pass the test the participant had to answer at least four out of six trials correctly. As well as selecting for headphone use, this task also helps to screen out automated scripts ('bots') masquerading as participants (Chmielewski & Kucker, 2020), since successful completion of the task requires following written instructions and responding to auditory stimuli.

Performance incentives

Although our tasks are subjective in nature, meaning that there are no *a priori* right or wrong answers, we nevertheless wanted participants to listen carefully and perform the task honestly. To do that, prior to starting the main experiment, participants received one of the following instructions:

> "The quality of your responses will be automatically monitored, and you will receive a bonus at the end of the experiment in proportion to your quality score. The best way to achieve a high score is to concentrate and give each trial your best attempt." (*Dense rating task*)
>
> "The quality of your responses will be checked automatically, and high quality responses will be bonused accordingly at the end of the experiment. The best strategy is to perform the task honestly and carefully, and then you will tend to give good results." (*GSP task*)

We purposefully did not tell participants exactly how these quality scores were computed, so as to avoid biasing the participants to answer in a particular way (for example answering in a way that matches the responses of other participants). In actuality, we computed these scores by quantifying the participant's *self-consistency*, reasoning that participants who take the task carefully are likely to provide similar responses when presented with the same trial multiple times. This seemed the most reasonable option given the lack of ground truth for subjective tasks such as these.

Self-consistency was estimated as follows. Upon completion of the main experiment trials, participants received a small number of trials (three for GSP, and five for dense rating) which repeated randomly selected trials that were encountered earlier. The data from these trials contributed only to consistency estimation and not to the construction of the main experiment. Consistency was quantified by taking the Spearman correlation between the two sets of numerical answers. Participants were not informed of their consistency score, but at the end of the experiment they received a small monetary bonus in proportion to their score, constrained to range between 0 to 0.5 dollars. The exact mapping between score and bonus was $min(max(0,\ 0.5\ \times\ score),\ 0.5)$ for dense rating and $min(max(0,\ score\ -\ 0.5),\ 0.5)/2$ for GSP.

Participants

The US cohort was recruited from Amazon Mechanical Turk (AMT), a well-established online crowd-sourcing platform. We specified the following recruitment criteria: that participants must be 18 years of age, that they reside in the United States, and that they have a 95% or higher approval rate on previous AMT tasks. All participants provided informed consent in accordance with the Max Planck Society Ethics Council approved protocol 2020_05; all data collection was anonymous in order to protect participants' privacy. While we only allowed each individual to participate once in a given experiment, we did not regulate or track whether individuals participated in multiple experiments. Our reported AMT participant numbers therefore correspond to the total number of times someone participated in our experiments, rather than the total number of unique individuals across experiments.

We ran each AMT experiment for about a day, targeting about 150-200 participants for 1D experiments, and targeting somewhat larger cohorts (~200-400) for the multi-dimensional and tuning experiments; the latter experiments required exploring larger stimulus spaces and/or more subtle perceptual effects. All together the AMT cohort comprised 4,598 verified participants, in addition to 1,533 participants who failed to complete the pre-screening tasks or the main experimental tasks.

Three additional experiments were completed by a cohort of South Korean participants which were recruited through a research assistant residing in the local area (Lee et al., 2021); AMT was not possible in this case as AMT does not currently run in South Korea. Participants were required both to be born in South Korea and to be current residents there. In order to maximize the amount of available data, each participant was allowed to participate in the same experiment up to three times. Each participant provided informed consent following the Max Planck Society Ethics Council approved protocol 2702_12, and took a Korean-language version as translated by a native speaker.

The overall fee for participating in each experiment was computed by estimating the total duration of the experiment and multiplying by a rate of $9/hour. Individual participants were paid in proportion to the amount of the experiment that they completed. Participants were still compensated even if they left the experiment early on account of failing a pre-screening task.

Questionnaire

Upon completion of each experiment we collected demographic information as well as years of musical experience from participants ("Have you ever played a musical instrument? If yes, for how many years did you play?"). In the US cohort, reported ages varied in the range 18-81 ($M = 37.6$, $SD = 11.1$), and 40.2% identified as female (58.6% male and 1.2% other). Participants self-reported 0-55 ($Med = 2$, $M = 4$, $SD = 6.6$) years of musical experience. In the South Korean cohort, the reported age statistics were ($M = 27$, $SD = 10.50$) and those of the years of musical experience were ($Med = 1$, $M = 2.19$, $SD = 2.67$). Details of the demographic information of each experiment are provided in Tables 2-3.

Individual experiments

Dense rating

Studies 1-4 and 6. These experiments elicited pleasantness judgments for dyads of various timbres. Interval sizes were randomly sampled from uniform distributions, with the ranges of these distributions varying between studies: Studies 1-3 and 6 used a range of [0, 15] semitones, whereas Study 4 used ranges of the form $[I_c - 0.25, I_c + 0.25]$ where $I_c = 3.9, 8.9, 12$ respectively. The I_c values correspond to averages of just-intoned and equal-tempered tuning values at the major third, sixth and octave, rounded to one decimal place. In principle, each stimulus was to receive exactly one rating; occasionally for technical reasons the same stimulus was nonetheless administered to more than one participant, but then the latter response was excluded from the data analysis. The exact number of participants, timbres and average number of ratings per participant and per stimulus are summarized in Table 2.

Study 3. This experiment elicited pleasantness judgments from 322 participants for dyads with Type I timbre and varying spectral roll off (ρ). Stimuli were sampled uniformly from $[0, 15] \times [0, 15]$ where the first range corresponds to interval size in semitones and the second range corresponds to spectral roll-off in dB/octave.

Study	Description	Dataset	Ratings per participant	Total number of stimuli	Tone spectra	Participants			
						N	N_f	μ_m	σ_m
1A	Harmonic dyads	*dyh3dd*	37.9	7,500	Type I ($\rho = 3$)	198	73	4.1	6.9
1B(i)	Piano dyads	*harpno*	74.2	15,000	Type VI (piano) and I ($\rho = 3$)	202	78	3.9	6.3
1B(ii)	Guitar dyads	*hargtr*	71.4	15,000	Type VI (guitar) an I ($\rho = 3$)	210	86	4.5	6.9
1B(iii)	Flute dyads	*harflt*	78.9	15,000	Type VI (flute) and I ($\rho = 3$)	190	62	4.2	6.6
2A(i)	Stretched dyads	*dys3dd*	38.7	7,500	Type II ($\rho = 3, \gamma = 2.1$)	194	67	3.6	6.2
2A(ii)	Compressed dyads	*dyc3dd*	37.1	7,500	Type II ($\rho = 3, \gamma = 1.9$)	202	71	4.2	6.7
2B	Bonang dyads	*gamdyrt*	44.1	7,500	Type V	170	97	5.0	7.0
3	Dyads with varying roll-off	*rodyrt*	46.6	15,000	Type I ($0 \leq \rho \leq 15$)	322	145	4.0	6.0
4A(i)	Dyads with 5 equal harmonics	*w3rdd*	78.9	11,754	Type IV$_+$	149	56	3.3	5.2
4A(ii)	Dyads without third harmonic	*wo3rdd*	75	12,000	Type IV$_-$	160	74	3.6	5.5
4A(iii)	Pure dyads	*purdyrt*	42.6	7,500	Type III	176	67	4.2	6.2
4B(i)	Major third (harmonics)	*tun3p9*	49.8	11,796	Type I ($\rho = 3$)	237	99	3.7	5.9
4B(ii)	Major third (no harmonics)	*tunp39*	49.8	13,250	Type III	266	118	3.8	6.9
4B(iii)	Major sixth (harmonics)	*tun8p9*	49.6	11,397	Type I ($\rho = 3$)	230	101	4.4	7.3
4B(iv)	Major sixth	*tunp89*	49.9	11,346	Type III	227	95	3.8	6.8

	(no harmonics)								
4B(v)	Octave (harmonics)	*tunoch*	76.5	15,000	Type I ($\rho = 3$)	196	78	4.1	7.3
4B(vi)	Octave (no harmonics)	*tunocp*	78.2	14,471	Type III	185	68	3.1	5.7
6(i)	Harmonic dyads (Korean)	*korean-dyad-harm*	174.5	4,188	Type I ($\rho = 3$)	24	13	2.1	2.7
6(ii)	Stretched dyads (Korean)	*korean-dyad-str*	198.1	3,961	Type II ($\rho = 3, \gamma = 2.1$)	20	10	2.2	2.9
6(iii)	Compressed dyads (Korean)	*korean-dyad-comp*	199.1	4,777	Type II ($\rho = 3, \gamma = 1.9$)	24	11	2.3	2.8

Table 2. Overview of dense rating experiments. This includes (left to right): Study ID, description, dataset ID (see *Supplementary Materials*), number of ratings per participant, total number of stimuli used, tone spectra of the stimuli, N the total number of participants, N_f the number of female participants, and μ_m/σ_m the mean/standard deviation of reported years of musical experience.

GSP

Study 5. In each experiment participants completed a collection of 200 GSP chains each of length 40 (excluding the random seed). The stimuli consisted of triads composed with different fixed timbres and parametrized by two intervals in the range $[0.5, 8.5]$ semitones. A summary for each individual experiment can be found in Table 3.

Study 7. An overall of 394 participants completed 42 aggregated GSP chains each of length approximately 26. Each of the three adjectives ('pleasant', 'bright', 'dark') was represented by 14 chains. The stimuli consisted of triads comprising three complex harmonic tones ($n_H = 5$) and variable timbre weights (type VII). Overall, the stimuli had 7 active dimensions: two intervals in the range $[0.5, 8.5]$ semitones, four partial weights in the range $[0, 1]$ corresponding to partials 2-5 and a global log-amplitude parameter α taking values in the range $[-0.15, 0.1]$ (see *Stimuli*). The first partial weight was set to 1 to keep the fundamental frequency of the tone unambiguous. We aggregated 15 responses per iteration, and summarized them using a Gaussian kernel of size $\sigma = 0.3$ (see *Paradigms*).

Study	Description	Dataset	Dim.	Iterations	Agg.	Chains	Tone spectra	Participants			
								N	N_f	μ_m	σ_m
5A	Harmonic triads	*trdh3d*	2	40	1	200	Type I $(\rho = 3)$	228	86	3.8	6.2
5B(i)	Stretched triads	*trds3d*	2	40	1	200	Type II $(\rho = 3, \gamma = 2.1)$	229	83	4.6	7.5
5B(ii)	Compressed triads	*trdc3d*	2	40	1	200	Type II $(\rho = 3, \gamma = 1.9)$	233	90	4.4	7.7
7	Intervals and timbre	*mixintm*	7	26.1	15	42	Type VII	394	176	4.3	6.8

Table 3. Overview of GSP experiments. This includes (left to right): Study ID, description, dataset ID (see *Supplementary Materials*), the number of parameters to manipulate (dimension), the number of iterations per chain (excluding the random seed), the amount of aggregation, the number of chains collected, the tone spectra of the stimuli used, N the total number of participants, N_f the number of female participants, and μ_m/σ_m the mean/standard deviation of reported years of musical experience.

Data analysis and visualization

Each of our behavioral experiments involves sampling from some kind of continuous space. In each case we apply some kind of nonparametric smoothing to infer a smooth consonance terrain from these discrete samples. The precise nature of this smoothing varies depending on the paradigm type (rating versus GSP), the dimensions of the stimulus space, and the nature of the dimension (pitch interval versus timbre).

1D rating experiments (Studies 1, 2, 4, 6)

We computed the 1D consonance profiles over a grid of 1,000 points spanning the interval range of interest. In Studies 1, 2, 4A, and 6, the range spanned 15 semitones; in Study 4B, the range spanned 0.5 semitones.

The behavioral profiles were computed by taking the trial-level rating data, *z*-scoring the ratings within participants, and then applying a Gaussian kernel smoother. For experiments spanning 0-15 semitones, we used a kernel with standard deviation of 0.2 semitones; for experiments spanning 0.5 semitones, we used a kernel with standard deviation of 0.035 semitones.

The interference model profiles were computed directly from the corresponding computational models, supposing a bass note corresponding to C4 (~ 262 Hz), and modeling the tone spectra on the corresponding experimental stimuli (see below for more details on the models). For the Milne (2013) and Harrison-Pearce (2018) model profiles, we applied an additional post-processing step of Gaussian kernel smoothing (bandwidth 0.03 semitones); this was introduced to compensate for the fact that the model implementations discretize pitch into bins of 0.01 semitones, which creates artifacts at high resolutions.

We computed confidence intervals for the behavioral profiles by nonparametric bootstrapping over participants. To keep computation time tractable we used 1,000 bootstrap replicates to estimate the standard error and then estimated 95% confidence intervals by making a Gaussian approximation ([mean - 1.96 × *SE*, mean + 1.96 × SE]). We used this same approach for all bootstrapped analyses in the paper.

We computed peaks for the behavioral profiles by applying a custom peak-picking algorithm comprising the following steps:

1. Take the kernel-smoothed behavioral profile as an input; this corresponds to a vector of intervals and a vector of corresponding pleasantness values, both of length 1,000.
2. Approximate this profile using a cubic smoothing spline (implemented as 'smooth.spline' in R). We set the equivalent number of degrees of freedom to 100 to ensure perfect interpolation. Write this spline function as $f(x)$, where x is the interval in semitones.
3. Compute the first and second derivatives of this spline function ($f'(x), f''(x)$).
4. Compute the range of the spline function, writing it as $range(f(x))$.
5. Compile a list of all *peaks* in the function. A peak is defined as a value x_i where the following holds: $f'(x_i) > 0 > f'(x_{i+1})$ and $f''(x_i) < - range(f(x)) / 20$.
6. Compile a list of all *troughs* in the function. A trough is defined as a value x_i where the following holds: $f'(x_i) < 0 < f'(x)_{i+1}$ and $f''(x_i) > range(x) / 20$.
7. Merge peaks that aren't separated by *deep enough* troughs, in each case keeping the tallest peak. Formally: write P_i for the location of the ith peak, P_{i+1} for the next peak, and T_{min} for the lowest trough in between; we consider T_i to be 'deep enough' if $min(f(P_i), f(P_{i+1})) - f(T_{min}) \geq range(f(x)) / 100$.
8. Discard peaks that aren't sufficiently *sharp*. A peak P is considered sharp if it satisfies both of the following conditions:

$$\exists\, a \in [P - 0.5,\, P] : f(P) - f(a) \geq range(f(x)) / 100$$

$$\exists\, b \in [P,\, P + 0.5] : f(P) - f(b) \geq range(f(x)) / 100$$

We estimated the reliability of these peaks via the same nonparametric bootstrapping procedure described above. We took the 1,000 bootstrap replicates of the kernel-smoothed behavioral profiles created previously, and ran the peak-picking algorithm on each of these, producing 1,000 sets of peak estimates. For each of the peaks obtained from the original behavioral profile, we defined a neighborhood as being within +/- 0.5 semitones of that peak, and counted the proportion of bootstrap iterations where a peak was observed within that neighborhood. If this proportion was greater than 95%, we considered that peak to be statistically reliable. We then computed 95% confidence intervals for the location of that peak by taking the standard deviation of those bootstrapped peak locations (i.e., the bootstrapped standard error), and then performing the same Gaussian approximation described above ([mean - 1.96 × *SE*, mean + 1.96 × SE]).

2D rating × roll-off experiment (Study 3)

We computed consonance profiles for Study 3 by factorially combining three spectral roll-off levels (2, 7, and 12 dB/octave) with the same 1,000-point interval grid from the previous analyses. We smoothed the behavioral data using an analogous kernel smoothing process to before, using a Gaussian function with an interval standard deviation of 0.2 semitones (as before) and a roll-off standard deviation of 1.5 dB/octave. All other aspects of the analysis (smoothing the harmonicity models, estimating peak locations, and computing confidence intervals) were identical to before.

2D interval × interval experiment (Study 5)

We computed consonance profiles for Study 5 over a grid of 500 x 500 points spanning 0.5-8.5 semitones on both dimensions. Following the standard GSP approach (Harrison et al., 2020), the behavioral consonance profiles correspond to kernel density estimates over populations of trials. Here we used a kernel density estimator with a bandwidth of 1.5 semitones. We computed consonance model outputs assuming a bass note of C4 as before, and smoothed the harmonicity models using the same approach as before.

6D interval × interval × timbre experiment (Study 7)

We computed consonance profiles using the same approach as in Study 5, but with the grid spanning 0.5-7.5 semitones. We computed spectral profiles by taking mean spectral amplitudes over all trials, bootstrapping 95% confidence intervals using the same approach as before.

Spectral approximations for naturalistic instruments

Study 1B reports spectral approximations for several naturalistic musical instruments (flute, guitar, piano). We calculated these spectra using the method of McDermott, Schemitsch, and Simoncelli (2013) as applied to audio samples for a C4 tone. First the signal was passed through an array of equally spaced cosine filters expressed on the MIDI scale, with 0.5 semitone spacing between filters, 50% overlap between adjacent filters, and filter locations spanning the range 15-119 semitones. Temporal envelopes at each of the filter locations were then extracted by taking the modulus of each filter's analytic signal as computed using the Hilbert transform, then applying a non-linearity ($f(s) = s^{0.3}$), and then resampling to 400 Hz. Temporal envelopes for the first 20 harmonics were then estimated by reversing the non-linearity (i.e., applying $g(s) = s^{1/0.3}$) and evaluating a Gaussian kernel smoother (σ =

1.598 Hz) at each harmonic frequency over the entire temporal trajectory. Each harmonic's temporal envelope was then averaged to produce a single amplitude score.

Consonance models

While many psychoacoustic accounts of consonance perception have been presented over the centuries, recent literature has converged on two main candidate explanations: (i) interference between partials, and (ii) harmonicity (see Harrison & Pearce, 2020 for a review). We address both accounts in this paper using computational modeling, as described below.

Interference models

According to interference accounts, consonance reflects interference between the partials in the chord's frequency spectrum. The nature of this interference depends on the distance between the partials. Distant partials and very close partials elicit minimal interference; however, partials separated by a moderately small distance (of the order of a semitone) elicit a large amount of unpleasant interference. This interference is most commonly thought to derive from fast amplitude fluctuation ("beating"; Helmholtz, 1875) but may also reflect masking (Huron, 2001; Parncutt, 1989).

The literature contains many interference-based consonance models (see Harrison & Pearce, 2020 for a review). These models vary in their mechanistic complexity, but interestingly the older and simpler models seem to perform better on current empirical data (Harrison & Pearce, 2020). Here we use a collection of so-called 'pure-dyad' interference models (Hutchinson & Knopoff, 1978; Sethares, 1993; Vassilakis, 2001) which calculate interference by summing contributions from all pairs of partials in the acoustic spectrum, where each pairwise contribution is calculated as an idealized function of the partial amplitudes and the frequency distance between the partials. We focus particularly on the model of Hutchinson and Knopoff (1978), which performed the best of all 21 models evaluated in Harrison and Pearce (2020), but we also explore the other models in the *Supplementary Materials*. We avoided testing more complex waveform-based models (e.g. Daniel & Weber, 1997; Wang et al., 2013) because of their high computational demands and relatively low predictive performance (Harrison & Pearce, 2020).

At the core of the Hutchinson-Knopoff model is a dissonance curve that specifies the relative interference between two partials as a function of their frequency distance, expressed in units of critical bandwidths (Figure 17). This relative interference is converted into absolute interference by multiplying with the product of the amplitudes of the two partials. The main differences between the Hutchison-Knopoff, Sethares, and Vassilakis models correspond to the precise shapes of the dissonance curves and the precise nature of the amplitude weighting.

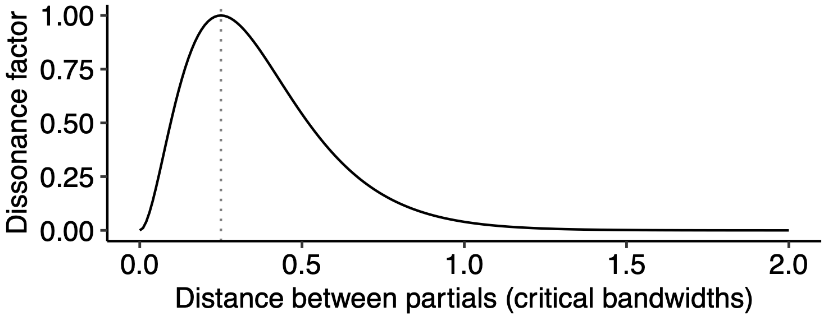


Figure 17. The dissonance curve of the Hutchinson-Knopoff model. The distance between partials that achieves maximal dissonance (0.25) is annotated with a dotted line. Here and elsewhere we use a parametric version of the Hutchinson-Knopoff dissonance curve introduced by Bigand et al. (1996): $D(x) = (4x\exp(1 - 4x))^2$.

Harmonicity models

According to harmonicity accounts, consonance is grounded in the mechanisms of pitch perception. Pitch perception involves combining multiple related spectral components into a unitary perceptual image, a process thought to be accomplished either by template-matching in the spectral domain or autocorrelation in the temporal domain (see de Cheveigné, 2005 for a review). Consonance perception can then be modeled in terms of how well a particular chord supports these pitch perception processes. Here we test three such models: two based on template-matching (Harrison & Pearce, 2018; Milne, 2013) and one based on autocorrelation, after Boersma (1993) (see *Methods* for details). We focus particularly on the model of Harrison and Pearce (2018) because of its high performance in Harrison and Pearce (2020), but we also explore the other models in the *Supplementary Materials*. We excluded several other candidate models because they are insensitive to spectral manipulations, the main focus of this paper (Gill & Purves, 2009; Parncutt, 1988; Stolzenburg, 2015).

The Harrison-Pearce and Milne models

Following Milne (2013), the Harrison-Pearce model uses a harmonic template corresponding to an idealized harmonic complex tone. The template is expressed in the pitch-class domain, a form of pitch notation where pitches separated by integer numbers of octaves are labeled with the same pitch class. It can be transposed to represent different candidate pitches; Figure 18A shows templates for C, D, E, and F.

Each input chord is likewise expressed as an idealized spectrum in the pitch-class domain, after Milne (2013) (Figure 18B). This involves expanding each chord tone into its implied harmonics, making sure to capture any available information about the strength of the harmonics (e.g., spectral roll-off) and their location (e.g. stretched versus non-stretched).

A profile of 'virtual pitch strength' is then created by calculating the cosine similarity between the chord's spectrum and different transpositions of the harmonic template, after Milne (2013) (Figure 18B). For example, the virtual pitch strength at '2' corresponds to the cosine similarity between the chord's spectrum and a harmonic template with a pitch class of '2' (i.e., a D pitch template).

Finally, harmonicity is estimated as a summary statistic of the virtual pitch strength profile. The Harrison-Pearce model treats this profile as a probability distribution, and computes the information-theoretic uncertainty of this distribution, equivalent to the Kullback-Leibler divergence to this distribution from a uniform distribution; high uncertainty means an unclear pitch and hence low harmonicity. Milne's (2013) model takes the same approach, but instead returns the height of the highest peak of this distribution.

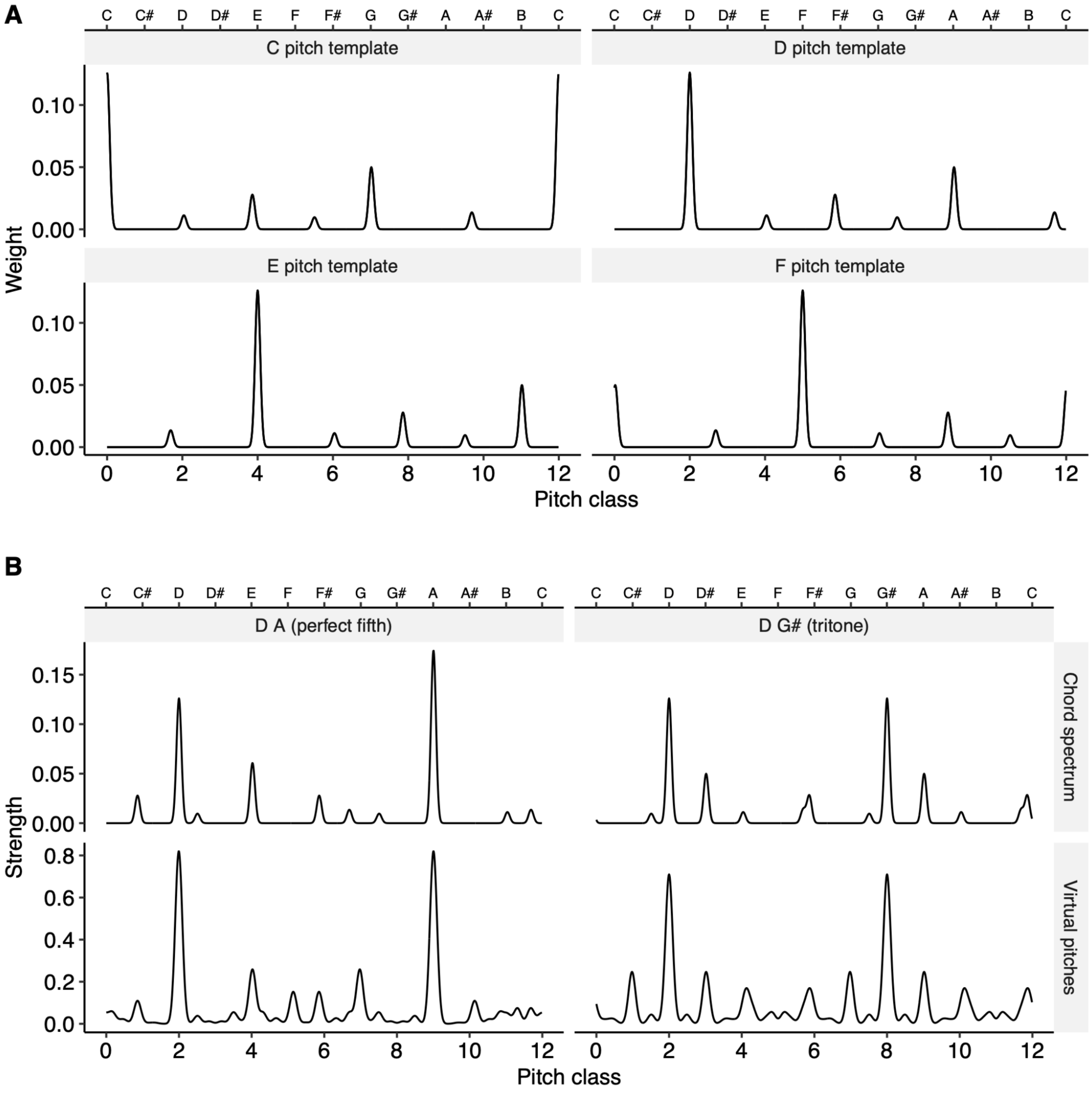


Figure 18. The Harrison-Pearce and Milne harmonicity models. (**A**) Idealized harmonic templates corresponding to harmonic complex tones with different fundamental frequencies. (**B**) Idealized input chord spectra (top row) and corresponding virtual pitch strength (bottom row), with the latter computed as the cosine similarity between the chord spectrum and harmonic templates of different candidate pitches.

The autocorrelation model

The autocorrelation model uses the fundamental-frequency estimator of Boersma (1993), as implemented in the Praat software and accessed via the Parselmouth package (Jadoul et al., 2018)). The algorithm works by looking for the maximum of the sound's autocorrelation function (i.e., the temporal interval at which the sound correlates maximally with itself).

The following steps were used to estimate the harmonicity of a given chord:

1. Synthesize the chord to an audio file using additive synthesis;
2. Estimate fundamental frequency from the audio file using Boersma's (1993) algorithm, using a time-step of 0.1 s, and bounding candidate fundamental frequencies to lie above 10 Hz but more than four semitones below the lowest chord tone.
3. Take the median of these fundamental frequency estimates. After Stolzenburg (2015), high fundamental frequencies are taken as implying high periodicity and hence high harmonicity.

Code and data availability

We provide a full set of data and code to enable readers to explore our data, replicate our analyses, and reproduce our experiments. These are all available in the *Supplementary Materials*, but can also be accessed individually at the following links:

- Raw data, individual experimental implementations, and experiment templates: https://gitlab.com/raja.marjieh/consonance-and-timbre-data

- Analysis code: https://gitlab.com/pmcharrison/timbre-and-consonance-paper

- Interactive web app: https://pmcharrison.gitlab.io/timbre-and-consonance-paper/supplementary.html

Acknowledgments

Author contribution

Authors RM, PH and NJ designed the experiments. Author RM conducted the experiments with US participants (Studies 1-5, 7); authors HL and FD conducted the experiments with

Korean participants (Study 6). Author PH performed the psychological modeling and prepared the supplementary materials. Authors RM, PH and author NJ wrote the paper. All authors edited and commented on the manuscript.

Funding statement

Fotini Deligiannaki was supported by an Erasmus+ scholarship.

Conflict of interest

The authors declare that they have no competing interests.

References

Ambrazevičius, R. (2017). Dissonance/roughness and tonality perception in Lithuanian traditional Schwebungsdiaphonie. *Journal of Interdisciplinary Music Studies*, *8*(1&2), 39–53. https://doi.org/10.4407/jims.2016.12.002

Bidelman, G. M., & Krishnan, A. (2009). Neural correlates of consonance, dissonance, and the hierarchy of musical pitch in the human brainstem. *The Journal of Neuroscience*, *29*(42), 13165–13171. https://doi.org/10.1523/JNEUROSCI.3900-09.2009

Bigand, E., & Poulin-Charronnat, B. (2006). Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100–130. https://doi.org/10.1016/j.cognition.2005.11.007

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, *17*(1193), 97–110.

Bowling, D. L., & Purves, D. (2015). A biological rationale for musical consonance. *Proceedings of the National Academy of Sciences*, *112*(36), 11155–11160. https://doi.org/10.1073/pnas.1505768112

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

Brown, S., & Jordania, J. (2013). Universals in the world's musics. *Psychology of Music*, *41*(2), 229–248. https://doi.org/10.1177/0305735611425896

Butler, J. W., & Daston, P. G. (1968). Musical consonance as musical preference: A cross-cultural study. *The Journal of General Psychology*, *79*, 129–142.

Chiba, G., Ho, M.-J., Sato, S., Kuroyanagi, J., Six, J., Pfordresher, P., Tierney, A., Fujii, S., & Savage, P. E. (2019). *Small-integer ratios predominate throughout the world's musical scales*. https://doi.org/10.31234/osf.io/5bghm

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Cousineau, M., McDermott, J. H., & Peretz, I. (2012). The basis of musical consonance as revealed by congenital amusia. *Proceedings of the National Academy of Sciences*, *109*(48), 19858–19863. https://doi.org/10.1073/pnas.1207989109

Daniel, P., & Weber, R. (1997). Psychoacoustical roughness: Implementation of an optimized model. *Acta Acustica United with Acustica*, *83*(1), 113–123.

de Cheveigné, A. (2005). Pitch perception models. In C. J. Plack & A. J. Oxenham (Eds.), *Pitch: Neural coding and perception* (pp. 169–233). Springer. https://doi.org/10.1007/0-387-28958-5_6

Dobbins, P. A., & Cuddy, L. L. (1982). Octave discrimination: an experimental confirmation of the "stretched" subjective octave. *The Journal of the Acoustical Society of America*, *72*(2), 411–415. https://doi.org/10.1121/1.388093

Eerola, T., & Lahdelma, I. (2021). The anatomy of consonance/dissonance: Evaluating acoustic and cultural predictors across multiple datasets with chords. *Music & Science*, *4*.

https://doi.org/10.1177/20592043211030471

Eerola, T., & Lahdelma, I. (2022). Register impacts perceptual consonance through roughness and sharpness. *Psychonomic Bulletin & Review*, *29*(3), 800–808. https://doi.org/10.3758/s13423-021-02033-5

Euler, L. (1739). *Tentamen novae theoria musicae*. Academiae Scientiarum.

Florian, G. (1981). The two-part vocal style on Baluan Island Manus Province, Papua New Guinea. *Ethnomusicology*, *25*(3), 433–446.

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

Gill, K. Z., & Purves, D. (2009). A biological rationale for musical scales. *PloS One*, *4*(12). https://doi.org/10.1371/journal.pone.0008144

Hall, D. (1973). The objective measurement of goodness-of-fit for tunings and temperaments. *Journal of Mathematics & Music*, *17*(2), 274–290. https://doi.org/10.2307/843344

Harrison, P. M. C., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs Sampling with People. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2008.02595

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.

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

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer. https://doi.org/10.1007/978-0-387-21606-5

Helmholtz, H. L. F. (1875). *On the sensations of tone as a physiological basis for the theory of music* (A. J. Ellis, trans.). Longmans, Green and Co.

Huron, D. (1994). Interval-class content in equally tempered pitch-class sets: Common scales exhibit optimum tonal consonance. *Music Perception*, *11*(3), 289–305. https://doi.org/10.2307/40285624

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, *19*(1), 1–64. https://doi.org/10.1525/mp.2001.19.1.1

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

Jacoby, N., Undurraga, E. A., McPherson, M. J., Valdés, J., Ossandón, T., & McDermott, J. H. (2019). Universal and non-universal features of musical pitch perception revealed by singing. *Current Biology*, *29*(19), 3229–3243.e12. https://doi.org/10.1016/j.cub.2019.08.020

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

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

Kameoka, A., & Kuriyagawa, M. (1969). Consonance theory Part I: Consonance of dyads. *The Journal of the Acoustical Society of America*, *45*(6), 1451–1459. https://doi.org/10.1121/1.1911623

Knöferle, K., & Spence, C. (2012). Crossmodal correspondences between sounds and tastes. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-012-0321-z

Krueger, F. (1904). Differenztöne und Konsonanz. *Archiv Fur Die Gesamte Psychologie*, *2*, 1–80.

Krueger, F. (1910). Die Theorie der Konsonanz. *Psychologische Studien*, *5*, 294–411.

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*(1), 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*(1), 8693. https://doi.org/10.1038/s41598-020-65615-8

Lee, H., Hoeger, F., Schoenwiesner, M., Park, M., & Jacoby, N. (2021). Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms. *ArXiv*. http://arxiv.org/abs/2108.00768

Li, G. (2006). The effect of inharmonic and harmonic spectra in Javanese Gamelan tuning (1): A theory of the Sléndro. *Proceedings of the 7th WSEAS International Conference on Acoustics & Music: Theory & Applications*, 65–71.

Maher, T. F. (1976). "Need for resolution" ratings for harmonic musical intervals: A comparison between Indians and Canadians. *Journal of Cross-Cultural Psychology*, *7*(3), 259–276.

McBride, J., & Tlusty, T. (2021). Convergent evolution in a large cross-cultural database of musical scales. In *PsyArXiv*. https://doi.org/10.31234/osf.io/eh5b3

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., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*(4), 493–498. https://doi.org/10.1038/nn.3347

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

McKinney, M. F., & Delgutte, B. (1999). A possible neurophysiological basis of the octave enlargement effect. *The Journal of the Acoustical Society of America*, *106*(5), 2679–2692. https://doi.org/10.1121/1.428098

McLachlan, N., Marco, D., Light, M., & Wilson, S. (2013). Consonance and pitch. *Journal of*

Experimental Psychology: General, *142*(4), 1142–1158. https://doi.org/10.1037/a0030830

McPherson, M. J., Dolan, S. E., Durango, A., Ossandon, T., Valdés, J., Undurraga, E. A., Jacoby, N., Godoy, R. A., & McDermott, J. H. (2020). Perceptual fusion of musical notes by native Amazonians suggests universal representations of musical intervals. *Nature Communications*, *11*(1), 2786. https://doi.org/10.1038/s41467-020-16448-6

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468). https://doi.org/10.1126/science.aax0868

Milne, A. J. (2013). *A computational model of the cognition of tonality*. Unpublished doctoral dissertation, The Open University.

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

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). https://doi.org/10.1371/journal.pone.0089642

Nordmark, J., & Fahlén, L. E. (1988). Beat theories of musical consonance. *STL-QPSR*, *29*(1), 111–122.

Ohgushi, K. (1983). The origin of tonality and a possible explanation of the octave enlargement phenomenon. *The Journal of the Acoustical Society of America*, *73*(5), 1694–1700. https://doi.org/10.1121/1.389392

Palmer, S. E., Schloss, K. B., Xu, Z., & Prado-León, L. R. (2013). Music-color associations are mediated by emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(22), 8836–8841. https://doi.org/10.1073/pnas.1212562110

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. (1989). *Harmony: A psychoacoustical approach*. Springer-Verlag.

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, *39*(10), 1365–1389. https://doi.org/10.1068/p6507

Plamondon, J., Milne, A., & Sethares, W. (2009). Dynamic tonality: Extending the framework of tonality into the 21st century. *Proceedings of the CMS South Central Chapter Conference*.

Plomp, R., & Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, *38*(4), 548–560. https://doi.org/10.1121/1.1909741

Popescu, T., Neuser, M. P., Neuwirth, M., 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*. https://doi.org/10.1038/s41598-018-35873-8

Preyer, W. (1879). *Akustiche Untersuchungen* (pp. 44–61). Verlag G. Fischer.

Rahn, J. (1996). Perceptual aspects of tuning in a Balinese gamelan angklung for North American students. *Canadian University Music Review/Revue de Musique Des Universités Canadiennes*, *16*(2), 1–43.

Rameau, J.-P. (1722). *Treatise on harmony*. Jean-Baptiste-Christophe Ballard.

Roberts, L. A., & Mathews, M. V. (1984). Intonation sensitivity for traditional and nontraditional chords. *The Journal of the Acoustical Society of America*, *75*(3), 952–959. https://doi.org/10.1121/1.390560

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992. https://doi.org/10.1073/pnas.1414495112

Schneider, A. (2001). Sound, pitch, and scale: From “tone measurements” to sonological analysis in ethnomusicology. *Ethnomusicology*, *45*(3), 489–519. https://doi.org/10.2307/852868

Schnupp, J., Nelken, I., & King, A. (2011). *Audio neuroscience: Making sense of sound*. MIT Press.

Schwartz, D. A., Howe, C. Q., & Purves, D. (2003). The statistical structure of human speech sounds predicts musical universals. *The Journal of Neuroscience*, *23*(18), 7160–7168.

Sethares, W. A. (1993). Local consonance and the relationship between timbre and scale. *The Journal of the Acoustical Society of America*, *94*(3), 1218–1228. https://doi.org/10.1121/1.408175

Sethares, W. A. (2005). *Tuning, timbre, spectrum, scale*. Springer.

Sievers, B., Polansky, L., Casey, M., & Wheatley, T. (2013). Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 70–75. https://doi.org/10.1073/pnas.1209023110

Smit, E. A., & Milne, A. J. (2021). The need for composite models of music perception: Consonance in tuning systems (familiar or unfamiliar) cannot be explained by a single predictor. *Music Perception*, *38*(3), 335–336. https://doi.org/10.1525/mp.2021.38.3.335

Stolzenburg, F. (2015). Harmony perception by periodicity detection. *Journal of Mathematics & Music*, *9*(3), 215–238. https://doi.org/10.1080/17459737.2015.1033024

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

Stumpf, C. (1898). Konsonanz und dissonanz. *Beiträge Zur Akustik Und Musikwissenschaft*, *1*, 1–108.

Tenney, J. (1988). *A history of “consonance” and “dissonance.”* Excelsior Music Publishing Company.

Terhardt, E. (1974). Pitch, consonance, and harmony. *The Journal of the Acoustical Society of America*, *55*(5), 1061–1069. https://doi.org/10.1121/1.1914648

van de Geer, W. J., Levelt, W. J. M., & Plomp, R. (1962). The connotation of musical consonance. *Acta*

Psychologica, *20*(4), 308–319. https://doi.org/10.1016/0001-6918(62)90028-8

Vassilakis, P. N. (2001). *Perceptual and physical properties of amplitude fluctuation and their musical significance*. Unpublished doctoral dissertation, University of California.

Vassilakis, P. N. (2005). Auditory roughness as a means of musical expression. In R. A. Kendall & R. H. Savage (Eds.), *Selected Reports in Ethnomusicology: Perspectives in Systematic Musicology* (Vol. 12, pp. 119–144). Department of Ethnomusicology, University of California.

Vos, J. (1986). Purity ratings of tempered fifths and major thirds. *Music Perception*, *3*(3), 221–257.

Vyčinienė, D. (2002). Lithuanian Schwebungsdiaphonie and its south and east European parallels. *The World of Music*, *44*(3), 55–57.

Wang, Y. S., Shen, G. Q., Guo, H., Tang, X. L., & Hamade, T. (2013). Roughness modelling based on human auditory perception for sound quality evaluation of vehicle interior noise. *Journal of Sound and Vibration*, *332*(16), 3893–3904. https://doi.org/10.1016/j.jsv.2013.02.030

Woods, K. J. P., Siegel, M. H., Traer, J., & Mcdermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Young, R. W. (1952). Inharmonicity of plain wire piano strings. *The Journal of the Acoustical Society of America*, *24*(3), 267–273. https://doi.org/10.1121/1.1906888

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception*, *31*(4), 339–358. https://doi.org/10.1525/mp.2014.31.4.339