sempre: Society for Education, Music and Psychology Research
MUSIC & SCIENCE

Research Article

Music & Science
Volume 4: 1–19

DOI: 10.1177/2059204321030471
journals.sagepub.com/home/mns
SAGE

The Anatomy of Consonance/Dissonance: Evaluating Acoustic and Cultural Predictors Across Multiple Datasets with Chords

Tuomas Eerola and Imre Lahdelma

Abstract

Acoustic and musical components of consonance and dissonance perception have been recently identified. This study expands the range of predictors of consonance and dissonance by three analytical operations. In Experiment 1, we identify the underlying structure of a number of central predictors of consonance and dissonance extracted from an extensive dataset of chords using a hierarchical cluster analysis. Four feature categories are identified largely confirming the existing three categories (roughness, harmonicity, familiarity), including spectral envelope as an additional category separate from these. In Experiment 2, we evaluate the current model of consonance/dissonance by Harrison and Pearce by an analysis of three previously published datasets. We use linear mixed models to optimize the choice of predictors and offer a revised model. We also propose and assess a number of new predictors representing familiarity. In Experiment 3, the model by Harrison and Pearce and our revised model are evaluated with nine datasets that provide empirical mean ratings of consonance and dissonance. The results show good prediction rates for the Harrison and Pearce model (62%) and a still significantly better rate for the revised model (73%). In the revised model, the harmonicity predictor of Harrison and Pearce's model is replaced by Stolzenburg's model, and a familiarity predictor coded through a simplified classification of chords replaces the original corpus-based model. The inclusion of spectral envelope as a new category is a minor addition to account for the consonance/dissonance ratings. With respect to the anatomy of consonance/dissonance, we analyze the collinearity of the predictors, which is addressed by principal component analysis of all predictors in Experiment 3. This captures the harmonicity and roughness predictors into one component; overall, the three components account for 66% of the consonance/dissonance ratings, where the dominant variance explained comes from familiarity (46.2%), followed by roughness/harmonicity (19.3%).

Keywords

Acoustic predictors, consonance, dissonance, familiarity, harmonicity, roughness, spectral envelope

Submission date: 5 October 2020; Acceptance date: 11 June 2021

Introduction

The investigation of musical consonance and dissonance—that is, the relative agreeableness/stability versus disagreeableness/instability of simultaneous and successive pitch combinations—has a long and checkered history (see e.g., Tenney, 1988). The Pythagorean school in ancient Greece held that consonance/dissonance (hereafter referred to as C/D and implying exclusively simultaneous pitch combinations) can be explained through the simplicity of number ratios, and this view was upheld well into the 16th century (e.g., in the work of music theorist Gioseffo Zarlino). In the 17th and 18th centuries, the origins of C/D were elaborated by scholars such as Marin Mersenne, Joseph Sauveur, and Jean-Philippe Rameau, who investigated the role of overtones and their relation to musical

Durham University, Durham, UK

Corresponding author:
Tuomas Eerola, Department of Music, Durham University, Palace Green, Durham DH1 3RL, UK.
Email: tuomas.eerola@durham.ac.uk

harmony. In the 19th century, scholars such as Hermann von Helmholtz (1875) and Carl Stumpf (1898) brought the knowledge of physics, anatomy, perception, and empirical testing to characterize C/D as something that depends on frequencies of the fundamental and the partials of the sound and how these are interpreted within the musical tradition that the listener is familiar with. Twentieth-century psychoacoustics made large strides in charting the sensory aspects of phenomena such as dependence on the frequency (Terhardt, 1984) and critical bands (Plomp & Levelt, 1965). Today, the research field is starting to reach a consensus that the overall perception of C/D in simultaneous sonorities in the Western musical culture is arguably based on a combination of *roughness*, *harmonicity*, and *familiarity* (see e.g., Harrison & Pearce, 2020; McLachlan et al., 2013; Parncutt & Hair, 2011).

Roughness denotes the sound quality that arises from the beating of frequency components (see e.g., Hutchinson & Knopoff, 1978; Kameoka & Kuriyagawa, 1969), and harmonicity indicates how closely a sonority's spectrum corresponds to a harmonic series (see e.g., Parncutt, 1989). Familiarity, which has received the least amount of attention out of these three proposed features, denotes the prevalence of sonorities in a given musical culture which affects how familiar the listeners become with these sonorities (see e.g., Johnson-Laird et al., 2012). The order of importance between these features on the perception of C/D has remained contentious, and it has recently been the focus of a large-scale analysis (see Harrison & Pearce, 2020) that brought a number of roughness, harmonicity, and cultural models under a systematic review and analysis. This impressive modeling identified the strongest acoustic models for roughness and harmonicity, and it also demonstrated that across musical genres, roughness has a more strong and reliable negative effect on chord prevalence than harmonicity. However, it is too early to draw strong conclusions based on these results; these analyses were typically based on other proxy concepts than actual consonance, such as pleasantness in Bowling et al. (2018), and this method of assessing C/D has recently been demonstrated to result in possible confounds (Lahdelma & Eerola, 2020). Moreover, some of the analyzed datasets contained only a limited selection of chords/intervals (e.g., Schwartz et al., 2003) or have a large majority of culturally unfamiliar chords. Smit et al. (2019) found roughness, harmonicity, spectral entropy, derived familiarity, and mean pitch to contribute to C/D ratings in the case of unfamiliar (detuned) chords. Also, the range of register and timbre used in previous C/D research has been limited. Even a cursory analysis of the state-of-the-art literature suggests that a more thorough assessment of the possible contributions of the different main theoretical features is needed.

In the current study we aim to estimate which acoustic and cultural features account for perceptual evaluations of C/D. We assume that the three categories of features—*roughness*, *harmonicity*, and *familiarity*—as identified by Harrison and Pearce (2020) is a solid starting point to refine the model contributions. We will also add a new feature category labeled *spectral envelope*, since descriptors such as sharpness have been previously implicated in C/D studies (Zwicker & Fastl, 1990). Since there is a large number of possible models to include as the predicting features, we will explore and streamline the models in preliminary steps to provide robust, independent predictors for the actual model construction. In Experiment 1, we take a reasonable number (4–7) of predictors for each feature category and carry out cluster analysis of a new chord dataset to verify and possibly redefine the feature categories empirically. In Experiment 2, we use the confirmed feature categories to identify the most effective predictor for each category by comparing the alternative predictors within each feature category to the predictors in the current state-of-the-art model (Harrison & Pearce, 2020) using the raw consonance ratings in three recent studies. To assess the overall contribution of the models, in Experiment 3 we compare the C/D model that has been optimized via Experiment 2 and the model by Harrison and Pearce through building linear models via regression with the mean consonance ratings across nine datasets. We also evaluate the degree to which the feature categories contribute independently to the C/D ratings.

Experiment 1: Analysis of Consonance and Dissonance Predictors

We want to utilize a solid set of predictors for the stimuli when exploring the features of C/D. We obtain our predictors mostly from the compilation of models available in the `incon` library, an open-source R package by Harrison and Pearce (2020). Based on their extensive analysis of the existing datasets, Harrison and Pearce (2020) derive *roughness* predictions from Hutchinson and Knopoff's model (1978), *harmonicity* from a model created by Harrison and Pearce (2020), and *familiarity* from an analysis of Billboard chart hits (Burgoyne, 2012) encoded by Harrison and Pearce (2020). The rationale for starting with these three variables is that they have been shown to be a solid combination for predicting pleasantness ratings (used as a proxy for consonance) in past research (Bowling et al., 2018; Harrison & Pearce, 2020). However, we will also explore the potential of spectral elements that have been put forward as possible contributors to C/D in past research, such as sharpness (Zwicker & Fastl, 1990). In our analyses, we will explore 22 variant predictors (four roughness variants, seven harmonicity variants, six familiarity variants, and five spectral envelope predictors) including the three predictors utilized by Harrison and Pearce (2020). Our selection of predictors is not exhaustive and many possible candidates, such as those offered by Sethares (2005), Krimphoff et al. (1994), or Cook (2017), have been left out due to practical constraints related to reliable implementation of these models.

Definition of the Predictors

For *roughness*, we have four variant models. The model by Hutchinson and Knopoff (1978), hereafter `Hutc78`, sums the dissonance of all harmonics that are based on distances in critical bandwidths. A roughness model by Sethares (1993), abbreviated as `Seth93`, is based on the beating phenomena, which estimates the interference between the amplitudes of the partials on the dissonance curve established by Plomp and Levelt (1965). The roughness model by Vassilakis (2001), hereafter `Vass01` is a variant of Sethares's model, although it assesses the minimum amplitude of each pair of peaks instead of summing up all amplitudes as in the model by Sethares. The model by Wang et al. (2013), hereafter `Wang13`, incorporates an auditory periphery model (Aures, 1985), derives the critical bandwidths from the Bark scale, and utilizes nonlinear filtering to represent excitation levels in critical bands.

For *harmonicity*, there are seven alternative models: `Parn88`, `Parn94`, `Gill09`, `Miln13`, `Har18`, `Stol15` and `Bowl18`. Parncutt (1988) proposed a model (`Parn88`) that builds on Terhardt's (1984) chord-root model, which draws from pattern recognition consisting of harmonic series. The model utilizes pitch classes and assumes octave equivalence, and considers 10 subharmonics, both of which simplify the computation of the template matching. In Terhardt's model, higher subharmonics carry considerably less weight, but to rectify the problem of minor chord and the recognition of the right fundamental for it, Parncutt modifies the weight of the harmonics to give more prominence to the higher harmonics. The second element of the model deals with pitch classes and assigns the root as the pitch class that receives the greatest weight based on the harmonics. Harmonicity is taken as the ambiguity of the root of the chord, which is calculated by dividing the relative weight of the largest root with the number of possible roots for the chord. The implementation utilizes updated weights (Parncutt, 2006). Parncutt and Strasburger (1994) (`Parn94`) is an updated model by Parncutt (1989) and utilizes an idea proposed by Terhardt (1982) that matches different harmonic templates to the input by expanding the pitches into the implied partials. The improvements concern masking and other limitations of the auditory system before carrying out the matching. `Miln13` is a model by Milne (2013) that utilizes pitch classes and supplements the tones to represent a predefined rich harmonic spectrum. These spectral templates are added together and further modified by convolving them with a Gaussian distribution. The distance between this enriched template and a harmonic template is computed using a cosine similarity to determine the best fit and the cosine similarity itself is the estimated harmonicity for the chord.

The model by Gill and Purves (2009) is also based on a template-matching idea, and has been developed for intervals initially. The algorithm, hereafter `Gill09`, works out the common divisor of each note's fundamental frequency and builds a template that assumes a harmonic complex tone, starting from the inferred root tone. The harmonics created by the actual notes and the template are calculated as the proportion of the match. Although the initial work was done with intervals, the model has been shown to generalize to trichords and tetrachords as well (Bowling et al., 2018). An additional model that works as a tie-breaker for the model by Gill and Purves and accounts for small intervals has been offered by Bowling et al. (2018). This model (`Bowl18`) calculates the minimum distance between the fundamental frequencies of a chord and has been used to distinguish those chords where the fundamentals are within 50 Hz, in which case the chords with the highest overall minimum distance between the fundamentals are assumed to be more consonant. Here we do not couple these two models (`Gill09` and `Bowl18`) together but use them separately, and do not limit the frequency difference to 50 Hz.

Another variant harmonicity model by Stolzenburg (2015) is based on ratio simplicity that takes into account the sensitivity to small tuning deviations in chords that are not just-tuned. In this model (`Stol15`), each chord frequency is expressed as a fractional multiple of the bass frequency and ratio simplicity is then computed as the lowest common multiple of the fractions' denominators. As periodicity and harmonicity are essentially equivalent phenomena (Harrison & Pearce, 2020), this model has a clear motivation to be implemented as a harmonicity model. To be consistent with the other models of harmonicity, we invert the model output as it originally outputs high values for pitch combinations with high period length, which implies lower periodicity and consonance. Finally, Harrison and Pearce (2018) (`Harr18`) have proposed a variant of Milne's harmonicity model (Milne, 2013) where the template matching is not done with cosine distance but rather treating the profile as a probability distribution, which allows to measure the degree of the violation of the profile from the uniform distribution using Kullback-Leibler divergence.

For *familiarity*, the current model relies on corpus-based counts as an index of familiarity of the chords and intervals. This predictor (`Har19`), implemented by Harrison and Pearce (2018), derives pitch-class frequencies from the Billboard corpus (Burgoyne, 2012) consisting of 739 pieces, which is probably currently the best source of information to represent common chords in Western popular music. This model is available in the `incon` library. We created five variant models (`CorpPop`, `CorpClas`, `CorpJazz`, `KeyClar`, and `TonDiss`). The first three corpus-based variants are attempts to mitigate some issues with the `Har19` model. These relate to the encoding of the chords using pitch-class representation and a root. The full range of chord inversions cannot be adequately captured with such a representation: the octave interval will be missing, and there are plenty of possibilities for misattributing

chords due to inversions, especially for chords with four tones or more. For instance, [0, 5, 9] is an inversion of the major chord, and in our view, a more robust encoding of the pitch classes would rely on the fundamental structure of intervals such as the system presented by Forte (1973). In previous empirical studies the consonance ratings of the inversions have only shown marginal differences (Lahdelma & Eerola, 2016). This approach is also motivated by music theory, which suggests that inversions of a chord represent the same chord type as its root position (Rameau [1722], 1971). We re-encoded the Billboard corpus available in `hcorp` (Harrison & Pearce, 2018) as Forte classes using the conversion routine available in `music21` (Cuthbert & Ariza, 2010). Forte classes are defined by two numbers: the number of pitch classes, and the sequential number within that number of pitch-class sets. To derive the pitch-class sets, a chord is expressed as a pitch class (integers from 0 to 11) and further transformed to the so-called prime form, which is the transposed and sorted version of the pitch classes—for example, a 2nd inversion major triad contains pitch classes [7, 0, 4] and its prime form is [0, 4, 7]. The chord's Forte class is 3-11, which refers to three pitches and being eleventh in a set of chords with three pitches. The conversion into Forte classes in the Billboard dataset results in 72 unique chords being used instead of 157 unique chords in Harrison and Pearce's model. Another problematic issue with the corpus analysis is the rarity of intervals, which seldom occur by themselves in this type of music (e.g., in the Billboard corpus, intervals form 2.1% of chord occurrences, mainly in the form of m2/M7, which constitutes the majority—1.9%—of these). To rectify this issue, we estimate the prevalence of the seven intervals in Forte representation by collapsing the occurrence of each pitch class in the Billboard collection and convert these into their normal forms (seven intervals) as an estimation of the interval prevalence in the corpus. These probabilities are not dissimilar to the profiles obtainable from Krumhansl-Kessler key profiles (1982) or Huron's aggregate consonance values (1994), but the advantage here is that the values reflect the idiosyncrasies in the corpus. To combine the empirical probabilities of the chords ($chord_p$) and the inferred probabilities of the intervals (iv_p), we first balance the interval probabilities to be similar in terms of the negative log values to the distribution of the chord probabilities by $iv_p^{1.333}/\sum iv_p^{1.333}$. After this rescaling and normalization operation, the two sources are combined by a simple weighting scheme, where $chord_p \times 0.99$ and $iv_p \times 0.01$ to reflect the rarity of intervals in the corpus and to preserve the sum of the probabilities to 1 before recalculating the negative log values that are used as the model output. This predictor will be called `CorpPop` as it is based on a corpus of popular music. We also carry out the same operation for the two other corpora available in `hcorp` (Harrison & Pearce, 2018), namely classical and jazz. The classical corpus (n = 1,022) contains an assortment of Mozart, Chopin, Haydn, Bach, and Beethoven sonatas and string quartets (`CorpClas`). The jazz corpus (n = 1,186) consists of jazz standards taken from fake books (`CorpJazz`).

We also created two additional predictors of familiarity reflecting classic work on tonality and C/D. We calculate the tonal stability of each pitch class as established by Krumhansl and Kessler (1982), which is known to be a good estimator of pitch-class prevalence in Western classical music (Krumhansl, 1990), popular music (Temperley & Clercq, 2013), and even bebop jazz (Järvinen, 1995). The correlation between the best key profile and the pitch-class profile of the input has been used as a measure of key clarity (`KeyClar`) (see Lartillot & Toiviainen, 2007) and as this measure indexes the cultural conventions and shows higher values for pitch-class distribution with tonic, dominant, and third degree, it is a reasonable link with an aspect of consonance—perhaps *tonal consonance* (Huron, 1991). We also encode the tacit knowledge of tonal principles that Western listeners share in the form of *tonal dissonance*. This idea has been put forward by Johnson-Laird et al. (2012) and is available in the `incon` library (Harrison & Pearce, 2020). According to Johnson-Laird et al. (2012), the relevant principles of tonality are tacitly represented in the minds of listeners as a result of their experiences in listening to tonal music (i.e., enculturation). Their theory relies on three principles that appear to be embodied in tonal music: 1) the increasing trend in dissonance of chords in major scales, in minor scales only, and in neither sort of scale; 2) the privileged status of the major triad as the most consonant chord of all; and 3) the construction of tonal chords out of thirds. As explained by Johnson-Laird et al., within each of these levels, dissonance depends on the psychoacoustic factor of roughness. We feel that their model is aggregating numerous components with separate weights into one model, even though these separate principles and their weights have not been assessed in a wider context beyond their own research. To prune this model, we analyzed the three principles with respect to the consonance ratings in a representative dataset (Bowling et al., 2018). Correlations—and semi-partial correlations (accounting for other principles present in the `Eero21` model, introduced later)—suggested that the first principle of the model correlates better with the empirical data ($r(296) = 0.574$, $sr = 0.193$) than the other two principles (principle 2, r = -.303, $sr = 0.167$ and principle 3, r = -0.424 and $sr = 0.040$) or the aggregated model (r = -0.572, $sr = 0.152$). To take the model parsimony a step further still, we simplified the model by collapsing minor and other scales together to create a simple implementation that we call the *tonal dissonance* model (`TonDiss`), which assesses whether the chord can be constructed from a major scale (1) or not (0). This was motivated by analyzing the contribution of the three principles of the original model by testing each principle as a binary coded variable in regression to predict consonance ratings together with roughness, familiarity, and spectral envelope predictors (all from the

Eero21 model without any harmonicity predictor). This regression showed that only principle 1 contributed to the consonance ratings (unstandardized $\beta = 0.75$, $p < .001$) and the other two principles were not significant predictors of C/D in this dataset (principle 2, $\beta = -0.26$, $p = 0.25$, and principle 3, $\beta = 0.26$, $p = 0.25$). This simplified version of tonal dissonance is better than the original formulation when the two are compared within the context of existing models (Eero21, $\chi^2 = 15.265$, $p < .001$ or Harr20R, $\chi^2 = 16.144$, $p < .001$). This is in line with both musicological and psychological observations of the special role of major tonality in Western music as the norm; minor tonality and atonality are far less frequent compared to major tonality (Parncutt, 2014), and familiarity has been shown to affect stimulus perception as per the mere exposure effect (Zajonc, 2001) which postulates that exposure yields positive valence (in this case, consonance).

We introduce an additional category of predictors that did not feature in Harrison and Pearce's (2020) review, namely *spectral envelope*. This category is related to the shape of the energy distribution along the spectrum. For instance, in sharpness, the energy at high frequencies creates sharp sounds that are found to be less pleasant, and therefore sharpness has been implicated as a predictor of consonance in psychoacoustics studies (Terhardt, 1974; Zwicker & Fastl, 1990). We utilize Zwicker's model to calculate sharpness (hereafter SpecSharp). This model first calculates the loudness of the signal relying on the Bark scale using Zwicker's algorithm (Zwicker & Fastl, 1990) and computes sharpness across these frequency bands using the formula established by Zwicker, which emphasizes sounds with high-frequency content as having sharper timbre. Our calculation of sharpness is based on Matlab functions replicating Zwicker's work, created by Claire Churchill (2004). To widen the field in terms of the spectro-temporal characterization of the signal qualities that are not covered by roughness or harmonicity, we outline several additional predictors that capture additional properties of the spectral envelope. Spectral envelope relates closely to the register (as defined by mean pitch height) of the chord, which has been implicated as a predictor of C/D by at least two studies (Lahdelma & Eerola, 2016; Smit et al., 2019), but it could also be indexed with acoustic measures of spectral center of gravity (spectral centroid) provided that the timbre is the same. We characterize perceptual brightness with spectral centroid and spectral roll-off (SpecCentr, SpecRolloff), which both index the balance of the energy distribution of the frequency spectrum. Additionally, we calculate the irregularity (SpecIrreg) of the adjoining partials (Jensen, 1999), which may further capture aspects of the spectrum that could potentially contribute to C/D. We also calculate the standard deviation of the spectral flux (SpecFlux) of the Euclidean distance between two successive frames (20 ms) of the spectrum, which has also been suggested to contribute to C/D in past research (Herbst, 2018; Terhardt, 1984). All spectral envelope predictors were calculated in Matlab using MIR toolbox (Lartillot et al., 2008). The purpose of these five additional predictors is to explore the role of spectral envelope-related qualities of the sounds that are not captured by roughness or harmonicity in terms of their contribution to C/D.

Cluster Analysis of Predictors

To assess the numerous alternative models for all four categories of C/D features, we first wanted to establish whether the predictors represented the assumed categories. We assume that the predictors from the same feature category would be largely collinear and therefore easily clustered into the same cluster. We examine the degree of collinearity between all predictors by carrying out an analysis of a separate dataset created for this purpose called the Durham Chord Dataset (DCD). This dataset contains all pitch pairings no more than 12 semitones for 2-pitch (12 in total), 3-pitch (66 in total), 4-pitch (220), 5-pitch (495), and 6-pitch (792) combinations across three registers (starting from E_3, E_4, and E_5), resulting in 4,755 unique pitch combinations. To obtain predictions for the additional acoustic predictors, we generated all these pitch combinations using the piano timbre. The sounds were generated with Ableton Live 9 (a music sequencer software), using the Synthogy Ivory Grand Pianos II plug-in. The applied sound font was Steinway D Concert Grand. No reverb was used, and the intervals and chords had a fixed velocity (65). The DCD with audio and all predictors is available at https://github.com/tuomaseerola/DCD.

We first looked at the correlations between the predictors in the DCD dataset. The correlation matrix, shown in Figure 1, suggests that feature categories operate largely as surmised; the predictors within the same feature categories correlate highly positively with each other. To assess the membership of the two potentially mixed categories of roughness and harmonicity empirically, we first estimated the optimal number of clusters that would characterize the similarity of all predictors. We applied gap statistic (Tibshirani et al., 2001) with bootstrapping (1,000 replications) using the $1 - x_{ij}$ where x is the correlation matrix as the input to the hierarchical clustering algorithm to establish the optimal number of clusters. This analysis suggested four clusters as the plausible number of groupings in the data, which we have shown together with the correlations and the hierarchical cluster solution in Figure 1.

In this solution, shown in Figure 1, the harmonicity cluster contains harmonicity-related variables (Miln13, Har18, Parn88, Gill09, Bowl18, Parn94, Stol15) but it also has one familiarity predictor (TonDiss). Predictors representing roughness (Hutc78, Seth93, Vass01, Wang13) are grouped into the same cluster, although SpecFlux also aligns within this cluster. The rest of the spectral descriptors of energy balance (SpecSharp, SpecCentr, SpecRolloff,

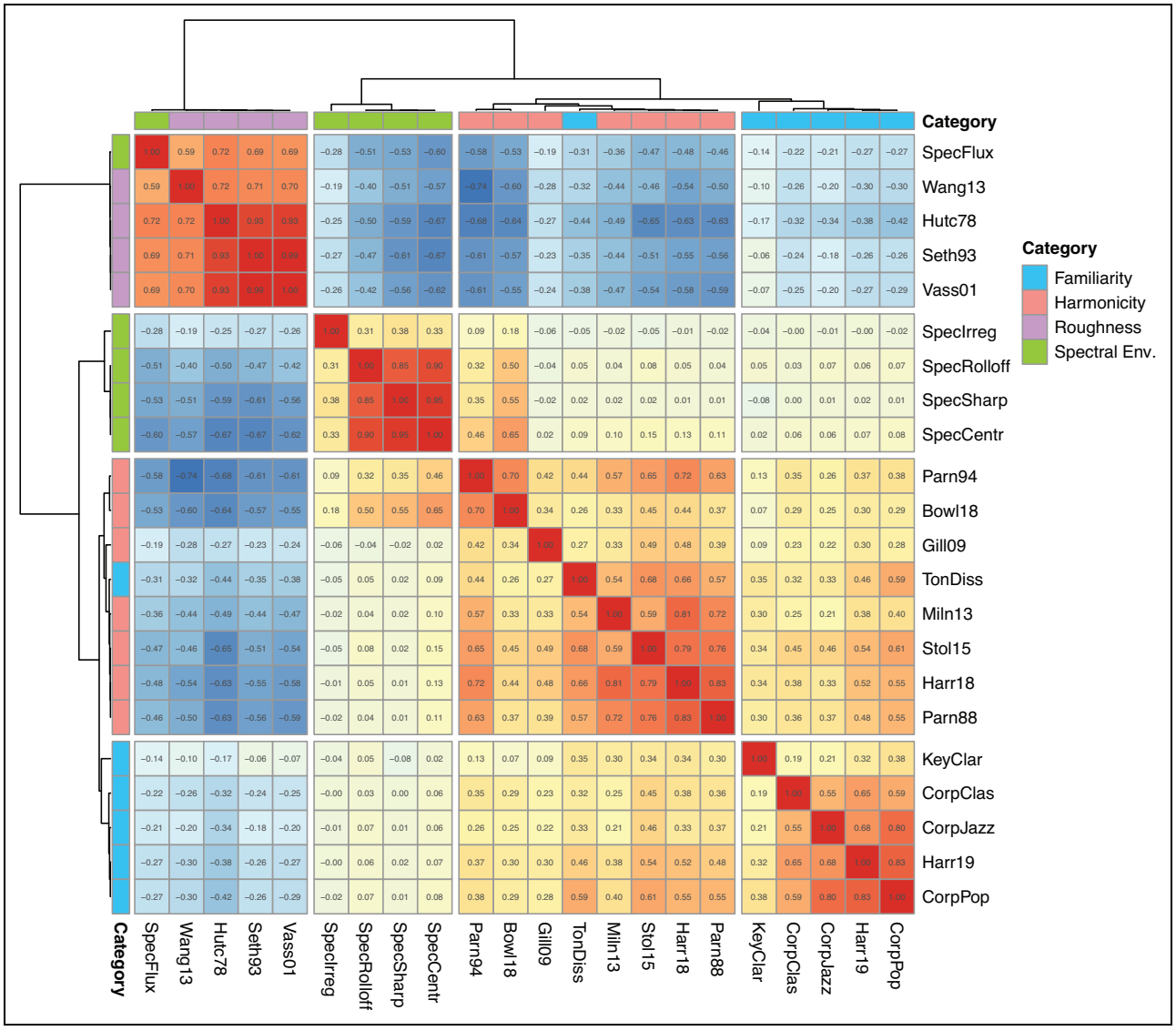

Figure 1. Correlations and a hierarchical cluster solution for predictors of roughness, harmonicity, familiarity, and spectral envelope categories (Durham Chord Dataset, $n = 4755$).

`SpecIrreg`) form a distinct spectral envelope cluster. The predictors representing familiarity of the underlying pitch combinations form their own cluster (`KeyClar`, `CorpPop`, `CorpJazz`, `CorpClas`, `Harl9`). Although the tonal dissonance model (`TonDiss`) that we created out of the model proposed by Johnson-Laird et al. (2012) was assumed to represent the acquired conventions of the tonal system, it is notably absent from this familiarity cluster.

In the subsequent analyses, we will use this empirically established classification of predictors as the basis of selecting predictors for an optimized model. Moreover, it is also worth pointing out that most acoustic predictors are strongly impacted by numerosity and register. Analytically, this can be demonstrated by separate ANOVAs for each predictor with two factors (numerosity, referring to the number of pitches in the chord, and register, referring to the three different octaves in which the chords were created). This analysis shows the main effect of numerosity for all predictors (df(4,4740), all $F > 4.03$) except `SpecRolloff` ($F = 0.35$, $p = 0.84$) and `SpecIrreg` ($F = 1.19$, $p = 0.31$), which are unaffected by numerosity. The majority of predictors display significant main effects of register (df(2,4740), all $F > 690$)) except `Harr18`, `Gill09`, `Miln13`, and `Parn88`. All of this is clearly visible in the predictor distributions across numerosity and register, shown in Figure 2.

Discussion

Using the extensive Durham Chord Dataset we established how the central models of *roughness*, *harmonicity*, *familiarity*, and *spectral envelope* correlate highly with each other within the four categories that emerge through hierarchical cluster analysis. Some predictors also exhibit high

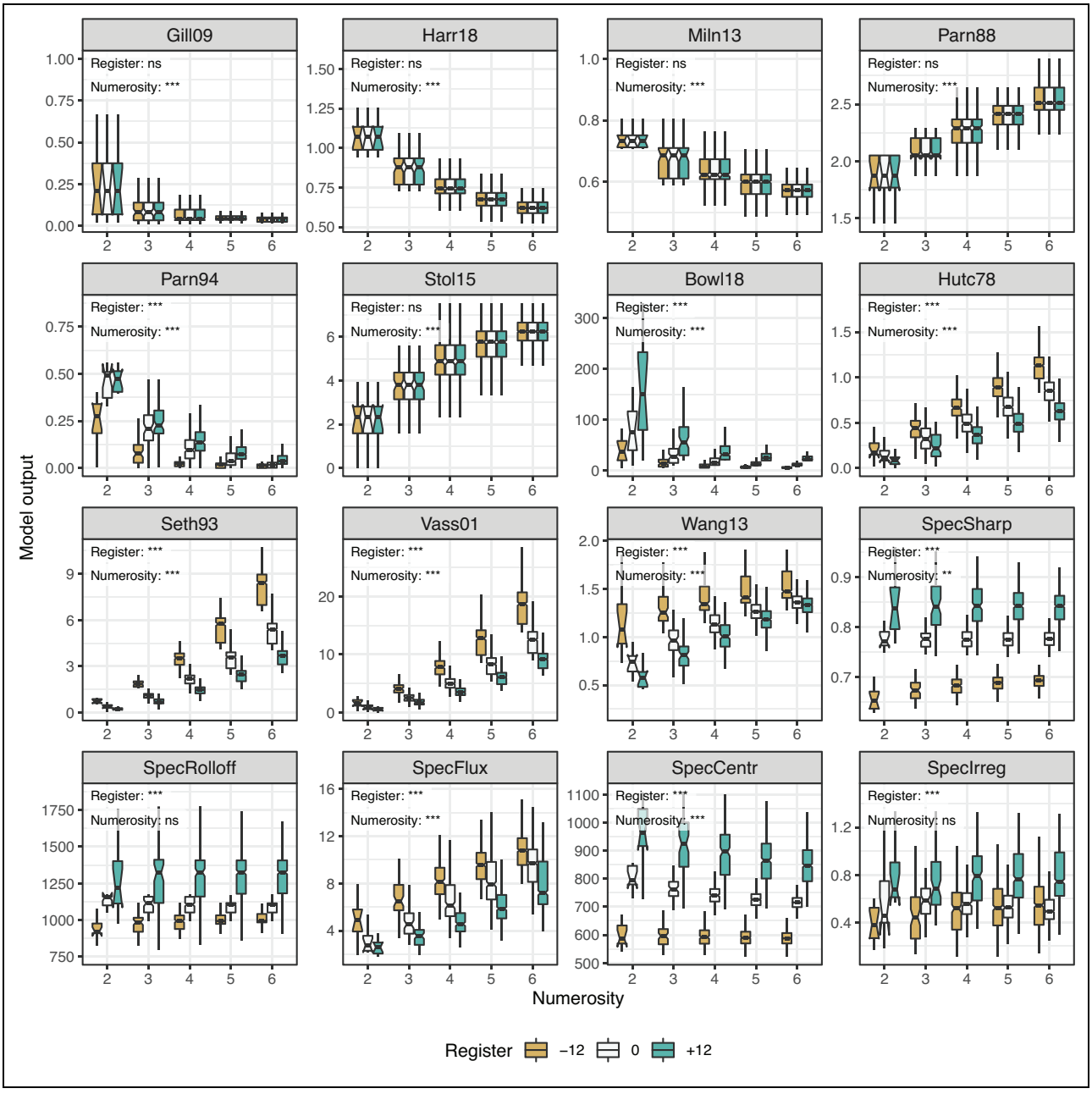

Figure 2. Numerosity and register across roughness and harmonicity models (Durham Chord Dataset, $n = 4755$). Asterisks refer to p-values in the ANOVA analysis across register or numerosity for each variable, * $p < .05$, ** $p < .01$, *** $p < .001$.

correlations between these theoretically derived categories (e.g., Stolzenburg's harmonicity model and Hutchinson's roughness model demonstrate a correlation coefficient of $r = 0.65$). This is an important caveat for understanding the independent contributions of the predictors of C/D in subsequent analyses, as multicollinear predictors hinder the interpretation of the predictor contributions in linear regression. The analysis of the dataset also demonstrated that a host of new predictors representing familiarity and spectral envelope operate more or less independently of roughness and harmonicity, although exceptions were observed as well. Spectral flux and tonal dissonance behave with real chords more akin to the other predictors of roughness and harmonicity, respectively. The predictors representing spectral envelope, which is a notion that has been previously proposed to account for dissonance across a wide frequency range, may not have featured strongly in recent studies such as Harrison and Pearce (2020), since the musical materials have not spanned a large range in register or other spectral differences (instrument timbres).

The present analysis also demonstrated how sensitive the models are as to the number of simultaneous pitches (i.e., pitch numerosity). This is worth paying attention to, since several studies have reported that perceptual

consonance ratings are dependent on pitch numerosity (Bowling & Purves, 2015; Lahdelma & Eerola, 2020; Lahdelma et al., 2020). Judging from the results of the present dataset, the statistical modeling needs to be able to handle numerosity differences in the models. Also, register could be an important determinant of C/D, although its role has remained largely unexplored; we will pay close attention to the register of the empirical datasets in our modeling and model comparisons. Although the current research has not yet extensively manipulated register, this is an important motivation to incorporate predictors to the consonance explanation that will be able to account for a wider variety of register and timbre in future research.

The present analysis, the selection of predictors, and the dataset were not exhaustive, of course; we did not address timbre in our dataset or the analysis which in real sounds has an undeniable contribution to C/D. However, the decision of leaving out timbre for now is related to the materials (models and data) available: the majority of the models do not easily incorporate timbre and the majority of the empirical data on consonance has been collected with relatively homogenous sounds using piano or sine wave timbres. Theoretically we could have created the Durham Chord Dataset with a wider range of intervals (2 octaves, resulting in 166,362 unique pitch combination if the same generation principles were used) and expanded the register (2, 3, or 4 octaves above and below) or the number of simultaneous pitches (to 7 or 8) for a more extensive coverage of these factors, but we felt that the point is already made and the variant formulations of the dimensions would yield little additional insight. There might be a small danger that the current dataset puts too much emphasis on high numerosity pitch combinations (5- and 6-pitch combinations form 81.1% of the dataset), which might undervalue the importance of familiarity predictors, since most of the 4-, 5-, and 6-pitch combinations rarely occur outside rare subgenres of contemporary music.

Next we will evaluate the goodness of individual predictors within the four categories of predictors using empirical data on C/D.

Experiment 2: Features of C/D

In the following analyses we will dissect C/D using four sets of features, namely *roughness*, *harmonicity*, *familiarity*, and *spectral envelope*. As a starting point, we will rely on the best three predictors (`Hutc78` for roughness, `Harr18` for harmonicity, and `Harr19` for familiarity) from the model by Harrison and Pearce (2020) to probe the contribution of these three features. We will call this the `Harr20` model. We will apply this model to three high-quality datasets (Experiments 1 and 2 in Lahdelma & Eerola, 2020; Popescu et al., 2019) where individual ratings across a range of C/D and numerosity are available. The advantage of starting with these datasets is that they contain ratings of actual C/D, whereas several other studies, including Bowling et al. (2018), have collected ratings for pleasantness (as a proxy for consonance), which is in fact not directly equal to consonance (see Lahdelma & Eerola, 2020). In our analysis, we substitute alternative predictors from each of the three feature categories in the original model to see whether the predictive rates can be improved with alternative predictors. We will add the spectral envelope category to the equation and estimate whether any predictor of the four proposed spectral envelope predictors is able to further improve the account of C/D beyond the three categories. We will also examine to what degree a composite model created by Harrison and Pearce (2020) (available in the `incon` package and labeled as `Harr20` hereafter) is able to account for consonance; this model combines roughness, harmonicity, and familiarity based on the analyses of data from Bowling et al. (2018).

Methods

We construct models with the data from Experiments 1 and 2 by Lahdelma and Eerola (2020) and by Popescu et al. (2019). The materials in Experiment 1 by Lahdelma and Eerola (2020) consist of 25 intervals, trichords, and tetrachords initially selected from Bowling et al.'s study (2018) and represent low cultural familiarity. The chords and intervals were presented to participants with piano and sine wave timbre. The materials of Experiment 2 by Lahdelma and Eerola (2020) are a balanced selection of 72 chords again selected from Bowling et al. (2018) representing high, medium, and low familiarity chords, all of which were presented with piano sounds. Popescu et al. (2019) provides rating data for 80 chords taken from real musical examples representing four distinct styles of music (jazz, classical, avant-garde, random) presented with piano sound. For all of these datasets (hereafter `lah20a`, `lah20b`, `pop19`), we model the individual mean ratings of each chord by the participants rather than the grand averages across the chords. This allows the responses to vary by participants and should offer better statistical inferences using mixed effects models. In these studies, participants gave C/D ratings using a 5-point interval scale (from 1 = dissonant to 5 = consonant in `lah20a` and `lah20b`) or a 7-point scale (from -3 = strong roughness to +3 = weak roughness in `pop19`). We rescaled the ratings in the Popescu et al. study (2019) to 1–5 before the analyses.

Data Analysis. We use linear mixed models (LMMs) to identify which acoustic or musical predictors are most consistent with the perceptual ratings of consonance. Analyzing `lah20a`, we collapse the repeated ratings (two for each stimulus) for each participant. Since we are not particularly interested in the number of pitches or timbre per se, we incorporate these as random effects. The number of pitches will have seven levels and timbre two levels. We also consider participants as a random factor. In addition, we define datasets (with three levels) as a random factor to permit

slight variations in the use of scales, terms, and interfaces. In this way, the dataset has 11,260 observations (25 stimuli × 2 timbre × 62 participants plus 72 stimuli × 80 participants plus 80 stimuli × 30 participants in `lah20a`, `lah20a`, and `pop19`, respectively). We also carried out control analyses where we treated numerosity as a fixed factor and eliminated datasets and timbre as random factors due to potential concerns over the small number of levels in the latter factors and possible masking effects in the former. However, these control analyses yield basically the same pattern of results (https://github.com/tuomaseerola/anatomy-of-consonance).

In our analysis, we will start with the `Harr20` model (three predictors from Harrison and Pearce's analyses) and estimate the contribution of each of the three predictors to the consonance ratings. Next we test whether this model can be improved by replacing each component of the model by each variant predictor in turn from the same category. We start this by testing whether any of the four variant roughness predictors will improve the model if these replace the predictor used in the original model (`Hutc78`). If a predictor is able to improve the model as indexed by Wald's χ^2 tests, we replace the original predictor in that category with the new candidate before moving on to the next category of predictors. In the next iteration, we move to harmonicity, test seven variant predictors and again if an alternative predictor is able to improve the model, we replace the original predictor with the new candidate. In case we have several candidates capable of improving the model, we take the strongest one as defined by the unique contribution to the model (*sr*). This analysis is sensitive to the order of the predictor categories. We carried out an auxiliary analysis with all 24 permutations of the predictor category orders and this demonstrated that the order does not affect the choice of the best predictor in any category. This auxiliary analysis is available in the electronic materials. In our view, this analysis strategy is cautious and controlled in comparison to an alternative strategy where one would start with all the predictors and carry out feature selection via regression. The alternative strategy would be prone to problems of multicollinearity, and would not have a sufficient number of observations to test all possible predictor combinations. Also, the regression approach cannot easily incorporate the variation offered by the experimental and participant factors. It is worth mentioning that we also checked whether musical expertise, gender, and age implemented as random effects in the analyses using the two datasets by Lahdelma and Eerola (2020). These extra factors implemented in this way did not impact the models in a significant fashion and therefore we leave them out of the analysis, as they have been reported in detail previously.

Results

First we assessed the previously established features of consonance (roughness, harmonicity, familiarity) with the data. This `Harr20R` model consists of `Hutc78` for roughness, `Harr18` for harmonicity, and `Harr19` for familiarity. The results are summarized in Table 1, which shows that two predictors, roughness and familiarity, contribute beyond the variance explained by the other two predictors in the model. The semi-partial correlations (*sr*) provide a convenient yardstick of the unique contribution of the predictors. The classic roughness model by Hutchinson and Knopoff (1978) has a strong contribution to the ratings of consonance, while the harmonicity model by Harrison and Pearce (2020) does not add anything to the overall model in this data when the two other predictors are already in the model. The strongest predictor, `Harr19`, is familiarity in the form of the Billboard corpus probabilities (Harrison & Pearce, 2018).

Table 1. LMM results for the original model by Harrison and Pearce (`Harr20R`) for consonance ratings across a sample of the three datasets and a comparison of alternative predictors within the predictor categories.

Predictor	χ^2	*p*-value	*sr*
Harr20 model			
`Hutc78` (Roughness)	446.50	< .001	-0.141
`Harr18` (Harmonicity)	2.21	ns	0.047
`Harr19` (Familiarity)	509.39	< .001	-0.216
Roughness variants			
`Wang13`	0	ns	-0.075
`SpecFlux`	0	ns	-0.060
`Seth93`	0	ns	-0.051
`Vass01`	0	ns	-0.047
Harmonicity variants			
`Stol15`†	148.76	< .001	0.108
`TonDiss`	124.33	< .001	0.087
`Parn88`	18.437	< .001	-0.058
`Bowl18`	72.502	< .001	-0.050
`Gill09`	7.72	< .01	0.037
`Miln13`	0	ns	0.034
`Parn94`	0	ns	0.001
Familiarity variants			
`CorpPop`†	305.15	< .001	-0.228
`CorpJazz`	0	ns	-0.147
`KeyClar`	0	ns	0.131
`CorpClas`	0	ns	-0.086
Spectral envelope variants			
`SpecIrreg`†	16.84	< .001	-0.047
`SpecRolloff`	0	ns	-0.021
`SpecSharp`	0.70	ns	-0.011
`SpecCentr`	0	ns	-0.007

†refers to the predictor taken forward to the model dubbed `Eero21`. The alternative predictors have been sorted based on the magnitude of the semi-partial correlations (*sr*).

For the model improvements within the four predictor categories, we replaced the predictor of one category of the original model with each variant predictor of the same category. For instance, for roughness, we tested whether replacing `Hutc78` with `Seth93` would improve the model when the model still has the two other predictors, `Harr18` and `Harr19`, present. The improvement is tested by

Table 2. Summary of the LMM results for predicting consonance ratings with different models across three datasets. The unstandardized beta coefficients are shown and the random effect significance testing is displayed with Likelihood Ratio Test (LRT). The two measures of overall fit refer to variance related to fixed factors (R^2_m) and to both random and fixed factors (R^2_c). Akaike Information Criteria (AIC) is reported to allow comparison of model complexity. The asterisks refer to *p*-values for the significance of the random factors where * is $p < .05$, ** is $p < .01$, and *** is $p < .001$.

	Null model	Harr20R model	Eero21 model
Fixed effects	*b* [95% CI]	*b* [95% CI]	*b* [95% CI]
Intercept	-	4.40 [3.92, 4.88]	4.97 [4.52, 5.42]
Roughness	-	-2.14 [-2.34, -1.94]	-1.32 [-1.55, -1.09]
Harmonicity	-	0.22 [0.03, 0.41]	0.17 [0.14, 0.20]
Familiarity	-	-0.11 [-0.11, -0.10]	-0.10 [-0.11, -0.09]
Spectral Envelope	-	-	-0.20 [-0.29, -0.12]
Random effects	LRT	LRT	LRT
Dataset	10.50**	28.66 ***	28.79 ***
Timbre	0.82	1.32	1.42
Numerosity	201.81 ***	65.97 ***	62.81 ***
Participant	896.47 ***	1210.88 ***	1276.34 ***
$R^2_{LMM(m)}$	0.000	0.210	0.250
$R^2_{LMM(c)}$	0.172	0.397	0.435
AIC	34485	31876	31407

comparing the strongest predictors established (which will be the `Harr20` model at the start) to the revised model with Wald χ^2 test. We will look at the unique contribution (semi-partial correlation, *sr*) of the predictor when comparing several predictor contributions to the model. Table 1 shows the breakdown of this iterative process, starting from the `Harr20` model. For the sake of comparability to the fixed weight composite model (`Harr20`), we display the unstandardized beta coefficients for the two new models and other diagnostic values, including the random effects, in Table 2.

Looking at the breakdown of the variant predictors in Table 1, *roughness* variants indicate that none of the variant formulations of roughness increase the `Harr20R` model significantly. For *harmonicity*, several harmonicity predictors are able to improve the `Harr20` model; the strongest contribution is offered by `Stol15`, which shows a significant improvement over the original model when `Harr18` is replaced by this variant predictor. A variant *familiarity* model (`CorpPop`) is able to improve the model that has the best roughness (`Hutc78`) and harmonicity component (now `Stol15`) with largest unique contribution of the harmonicity variants to the model ($sr = -0.238$). Finally, adding separately each new predictor in the category titled *spectral envelope* suggests that `SpecIrreg` has the highest χ^2 value and the largest unique contribution to the model, albeit a modest one ($sr = 0.047$). From this analysis we can tentatively draw together a new model labeled as *Eero21* that will consist of the best predictor from each category, namely `Hutc78` as the best roughness predictor, `Stol15` as the most robust harmonicity predictor, `CorpPop` as the superior familiarity predictor, and `SpecIrreg` as the best new spectral envelope predictor. The order of the analysis sequence followed the logic of the original model (roughness, harmonicity, familiarity). Spectral envelope was added as the final category when identifying the optimal predictors. This theory-driven sequential order of the analysis categories may have had an impact on the outcome of the analysis. We also conducted auxiliary analysis, available in the digital supplementary materials, including all 24 permutations of the predictor category orders. The results of this alternative analysis do not challenge the conservative analysis procedure reported here.

The overall performance of the `Eero21` model in comparison to the `Harr20R` model is shown in Table 2. This table also shows the random factors, which account for an additional variance, particularly *numerosity*, *dataset*, and *participant factor* in all datasets. Overall, `Harr20` and `Eero21` models achieve healthy marginal R squared value (see Nakagawa et al., 2017): 0.210 for the `Harr20R` model, 0.250 for the `Eero21` model that account for variance explained by the fixed effects. The difference between the `Harr20R` and the `Eero21` models is highly statistically significant (Wald $\chi^2 = 470.75$, $p < .001$) even when the addition of one extra predictor is accounted for using Akaike Information Criterion, which shows the lowest value for the `Eero21` model (Table 2). It is worth noting that in both models, all predictors are statistically significant (β 95% confidence intervals do not cross the zero). The `Harr20R` model, which is now fitted with this dataset, is statistically better than than the implementation of the same model with fixed beta weights (`Harr20`, $\chi^2 = 184.86$, $p < .001$), which is no surprise as the model coefficients have been adjusted for this dataset by the LMM analysis. Nevertheless, the `Harr20` model and its re-weighted version `Harr20R` have similar model coefficients (`Harr20` model has `Hutch78` = -1.62, `Har18` = 1.78, and `Har19` = -0.09; see Table 2 for coefficients for the `Harr20R` model in the present data). The major exception is the lower unstandardized beta coefficient for harmonicity, which may not stem from the inclusion of the

numerosity in the Harr2O model but rather could simply reflect the differences of chord choices in the data. Despite this difference, the broad similarities suggest that the model by Harrison and Pearce (2020) has appreciable capacity to generalize to other materials.

Discussion

An analysis of three datasets using LMMs was utilized to probe the merits of several alternative predictors to the composite model offered by Harrison and Pearce (2020). This model (Harr2OR) operates relatively well and in a stable fashion in these datasets. However, the shortcoming of the model seems to be harmonicity, where it fails to contribute significantly to the overall model. Substantial improvements could be identified by substituting another harmonicity predictor (Stol15) to the model. Another significant improvement was seen to come from familiarity and a revised corpus model, which utilizes a simpler account of the chord classifications (CorpPop). It seems that revising the calculation of the chord frequencies in the Billboard data by recoding them with unambiguous classes that do not make a distinction for chord inversions is able to capture more variation in the data than the previous encoding of the chord frequencies.

Adding a predictor of the regularity of the energy in the partials of the sounds (SpecIrreg) also improved the model significantly, albeit this contribution was the weakest overall. This finding is in line with theorizing by Zwicker and Fastl (1990) and previous empirical findings by Lahdelma and Eerola (2016). It is worth noting that in these analyses numerosity was treated as a random effect and we did not explore the impact or interactions with the predictors. We also ran the same analyses where we took numerosity, timbre, and datasets as fixed factors to the models but without observing any material changes to the results. When interactions between numerosity and the other predictors were tested, most of these were significant and suggest that building the models with an explicit numerosity predictor could lead to different results.

Next we will probe the contribution of the predictor categories across a larger selection of datasets to investigate the shortcomings and advantages of the models.

Experiment 3: Assessing C/D Features with Multiple Datasets

To explore the contribution of the acoustic predictors to C/D thoroughly, we compiled nine relatively recent datasets that contain consonance ratings (or one of its variant proxy terms, e.g., pleasantness) of intervals and chords. Our purpose is to apply the two variant models (Harr2OR, Eero21) established in Experiment 2 to these datasets. Three of the datasets (lah20a, lah20b, pop19) are those that were sampled in Experiment 2 to identify the optimal predictors. Some of the datasets are small and may only contain intervals or trichords, but the overall diversity in numerosity, register, rating scales as well as the countries and institutions in which these datasets have been collected should be an asset and guard against over-fitting and offer at least some level of generalizability of the results.

Datasets

The oldest dataset is from Schwartz et al. (2003), who compiled historic rankings of consonance of all intervals within an octave. Johnson-Laird et al. (2012) organized two experiments where they collected pleasantness (consonance) ratings for trichords and tetrachords organized according to their theory of dual-process theory of dissonance. Lahdelma and Eerola (2016) collected ratings of consonance for a small set (15) of trichords, tetrachords, pentachords, and hexachords. Likewise, Arthurs et al. (2018) carried out an experiment with a collection of trichords and tetrachords ($n = 12$) which were presented in two timbres and rated in terms of consonance, pleasantness, stability, and relaxation. Bowling et al. (2018) established the consonance (pleasantness) ratings for all 2-, 3-, and 4-pitch combinations within an octave ($n = 298$). Popescu et al. (2019) expanded the choice of chords by deriving them from real music spanning four styles (jazz, classical, avant-garde, random); the chords spanned a wide range in pitches and each style had 20 exemplars ($n = 80$ in total). Finally, Lahdelma and Eerola (2020) collected two substantial datasets of ratings of variant concepts of consonance for chords that were selected from the extensive collection established by Bowling et al. In most experiments, participants were Western, young, and educated, also the subject pools in many of these studies are relatively small; we have summarized the studies in Table 3. To make the datasets comparable, we have made sure that the ratings are in the same direction (high ratings indicate high consonance), reversing some of the scales (Johnson-Laird et al., 2012; Schwartz et al., 2003). We also have rescaled the ratings within the datasets to a range between 1 and 10 for consistency and comparability.

Data Analysis

We will carry out two linear regression analyses—unpooled and pooled—to probe the model performance within (unpooled) and across (pooled) the datasets. In the unpooled analysis, we explore the generalizability of the models by training each model within a dataset utilizing a cross-validation and applying the constructed model to a testing portion of the dataset. The performance of the model is indexed with prediction rate (R^2) in the unseen data (training portion). This diagnostic operation aims to outline the differences in the datasets and the ways in which the models pick these up. In the pooled analysis, however, we aggregate all observations across the datasets and assess the model fit using regression across the data. In the

Table 3. Description of the datasets including concepts, number of unique chords/intervals (*N*), numerosity, and pitch range in the stimuli.

Study	Abbrev.	*N*	Concept	Numerosity	Range
Schwartz (2003)	sch03	12	Consonance	2	C_4–C_5
Johnson-Laird (2012) (Exp. 1)	jl12a	48	Pleasantness	3	A_2–G_5
Johnson-Laird (2012) (Exp. 2)	jl12b	55	Pleasantness	4	G_2–D_5
Lahdelma (2016)	lah16	15	Smoothness	3, 4, 5, 6	G_4–C_6
Arthurs (2018)	art18	12	Consonance	3, 4	C_4–B_4
Bowling (2018)	bow18	298	Pleasantness	2, 3, 4	D_3–G_4
Popescu (2019)	pop19	80	Roughness	3, 4, 5, 6, 7, 8	E_1–C_7
Lahdelma (2020) (Exp. 1)	lah20a	25	Consonance	2, 3, 4	F_3–F_4
Lahdelma (2020) (Exp. 2)	lah20b	72	Consonance	2, 3, 4	E_3–G_4

Table 4. Results from two models showing prediction rates (R^2), standardized betas and semi-partial correlations (*sr*) for each predictor category for each dataset. $\bar{x}$ stands for weighted mean of the column.

Dataset	R^2	β_c	β_R	β_H	β_F	β_S	sr_R	sr_H	sr_F	sr_S
Harr20R										
sch03	1.00	-1.88	-1.02	0.44	-0.06	–	0.52	0.17	0.00	–
jl12a	0.62	-0.15	-0.52	0.43	0.14	–	0.29	0.19	0.21	–
jl12b	0.32	0.75	-0.40	0.56	0.30	–	0.17	0.38	0.21	–
lah16	0.95	-0.47	-0.33	0.54	0.53	–	0.10	0.22	0.23	–
art18	1.00	-0.81	0.53	0.21	0.93	–	0.04	0.03	0.66	–
bow18	0.60	-0.04	-0.30	0.16	0.49	–	0.22	0.10	0.38	–
pop19	0.73	0.24	-0.18	0.39	0.62	–	0.10	0.16	0.45	–
lah20a	0.85	0.24	-0.61	-0.22	0.85	–	0.35	0.10	0.33	–
lah20b	0.67	-0.24	-0.82	-0.06	0.41	–	0.38	0.04	0.49	–
$\bar{x}$	0.64	-0.03	-0.38	0.22	0.47	–	0.23	0.13	0.37	–
Eero21										
sch03	1.00	-2.11	0.00	1.41	-0.25	-0.21	0.10	0.25	0.18	0.03
jl12a	0.47	-0.01	-0.02	0.71	-0.18	-0.03	0.02	0.46	0.11	0.02
jl12b	0.79	0.67	-0.10	0.96	-0.15	-0.14	0.01	0.52	0.14	0.07
lah16	0.99	-0.46	-0.31	0.58	-0.47	0.03	0.11	0.15	0.20	0.07
art18	1.00	-1.19	-0.94	0.32	-0.59	0.58	0.13	0.12	0.20	0.18
bow18	0.78	-0.06	-0.06	0.35	-0.48	0.11	0.04	0.21	0.37	0.11
pop19	0.82	0.03	-0.19	0.21	-0.63	-0.07	0.10	0.08	0.43	0.07
lah20a	0.96	0.25	-0.64	0.00	-0.49	0.52	0.23	0.02	0.19	0.19
lah20b	0.88	-0.26	-0.33	0.37	-0.52	-0.04	0.17	0.17	0.54	0.11
$\bar{x}$	0.79	-0.10	-0.15	0.42	-0.45	0.06	0.07	0.22	0.34	0.10

construction of the models, we utilize a cross-validation scheme and predict the responses in the unseen part of the data. For the unpooled analyses, we utilize a 80/20% random split between training and prediction subsets and construct the model using a 10-fold cross-validation with 10 repeats. For pooled analysis, we have a similar split between training and testing data, but we increase the random repeats of the 10-fold cross-validation to 50. Again, the overall success of the model is captured with R^2 in the unseen portion of the data. To index the unique contribution of each predictor category within the models, we report semi-partial correlations (*sr*) between the predictor and C/D ratings when the contribution of all other predictors in the model have been partialled out. In contrast to the analyses in the previous section, we operate with mean data and relatively low number of observations for each dataset (see Table 4).

In addition, we carried out an auxiliary analysis where we identify the principal components of the predictor matrix and use either the component scores or the predictors that best represent the components as predictors in the regression. The purpose of this analysis is to offer a reliable assessment of the predictor contributions to C/D, as several of the feature categories are known to be highly collinear and hinder the interpretation of the model components.

Results

Unpooled Analysis Results. The results of the linear regression where both models have been trained and assessed on each dataset separately are shown in Table 4 with model fits, standardized beta coefficients, semi-partial correlations, and weighted means.

Table 5. Model summaries across the pooled data showing unstandardized beta coefficients, semi-partial correlations, and Akaike Information Criterion (AIC) for model parsimony.

Model	R^2	β_R	β_H	β_F	β_S	sr_R	sr_H	sr_F	sr_S	AIC
Harr20	0.57	-	-	-	-	-	-	-	-	1757
Harr20R	0.62	-4.89	1.86	0.41	-	-0.24	0.10	0.38		1725
Eero21	0.73	-2.41	0.66	-0.28	-3.22	-0.05	0.24	-0.35	-0.05	1643

Table 4 indicates that both models are able to produce adequate ($R^2 = 0.32$ for `jl12b`) to near-perfect fit ($R^2 = 1.00$ for `scho3` and `art18`) to different datasets, the weighted average being $\bar{R}^2 = 0.64$ in the `Harr20R` model. Also, the model coefficient directions and amplitudes are consistent, although some notable exceptions can be observed. For instance, in the `Harr20R` model, the standardized beta coefficient for roughness in `art18` is not in an inverse relationship to consonance ratings ($\beta = 0.53$), whereas the roughness coefficient is negative in all other models for different datasets. Such minor discrepancies may be traced to the unique combinations of the pitch collections the experiment contains, the case in point being the study by Arthurs et al. (2018), which contained a small set of familiar chords. The semi-partial correlations in the `Harr20R` model suggest that *familiarity* is carrying the bulk of the predictions (the weighted mean sr_F of 0.37) and *harmonicity* the least ($sr_F = 0.13$), consistent with the analysis of the three datasets in the previous section. In the `Eero21` model, the model prediction rates are generally higher ($\bar{R}^2 = 0.79$) than the rates obtained by the `Harr20R` model, although one exception is evident as well (`jl12a`). Interestingly, there are some inconsistencies in the way the predictor weights operate between the datasets in the `Harr20R` model. For example, *harmonicity* does have a negative sign for two datasets (`lah20a, lah20b`) in the `Harr20R` model, perhaps linked to the overall familiarity of the chords used those studies.

The semi-partial correlations spell different stories between the two models. In the `Eero21` model, *familiarity* is still one of the strongest predictor categories ($sr_F = 0.34$) and *harmonicity* comes close second ($sr_H = 0.22$), whereas *roughness* has lesser contribution to the ratings ($sr_R = 0.07$). The new predictor category of *spectral envelope* has about the same unique contribution as roughness ($sr_S = 0.10$) to the regression models. One way of looking at the differences between the two models is to examine the variation with semi-partial correlations between models in specific datasets; for instance, in the small dataset `art18` that can be perfectly predicted by both models, familiarity seems to deliver all variance ($sr_F = 0.66$) with the `Harr20R` model, whereas in `Eero21`, the harmonicity predictor also contributes individually ($sr_H = 0.15$) to the C/D ratings. One might assume that the contribution of familiarity is related to the question of whether a dataset has used familiar chords (e.g., `art18, lah20a`) in comparison to datasets comprised largely of unfamiliar stimuli (e.g., `bowl18, pop19`). However, the picture emerging from the *sr* values does not suggest that the differences in contributions of familiarity are related to the use of familiar chords in the datasets. The other noteworthy difference between the predictors is the shifting of the balance from roughness in `Harr20` to harmonicity in `Eero21`. This probably reflects the change of the harmonicity predictor in the model, but we will return to this question when we address the multicollinearity of the predictors that may impact the interpretation of these individual contributions.

Pooled Analysis Results. Turning our attention to how well the models operate across the datasets, we ran separate linear regressions for the `Harr20R` and `Eero21` models. For these analyses, the datasets were pooled together ($n = 617$) and a 10-fold cross-validation with 50 random repeats was carried out with a random 80% of the data to estimate the model coefficients. Again, the model prediction rates were estimated by applying the model relying on the coefficients from the training data to predict the unseen data. Table 5 summarizes these sets of analyses with the pooled data and also offers the baseline performance with the composite model (`Harr20`).

The baseline comparison is to the composite model (`Harr20`) by Harrison and Pearce (2020), which is the model without adjustable components (except constant), since the three predictors it contains (`Hutch78, Har18, Har19`) have predetermined coefficients based on the analysis of the data of Bowling et al. (2018). It can explain 57% ($R^2 = 0.572$) of the variance in the data, which is a solid and respectable quantity considering the overall challenge of being able to capture the C/D ratings in nine separate studies done with slightly different stimuli, raters, concepts, and participant backgrounds. The `Harr20R` model, which has the same three components but with optimized beta coefficients, is able to improve the model significantly ($\chi^2 = 68.34$, $p < .001$) although the increment is modest in variance explained ($R^2_\Delta = 0.043$). The coefficients of the `Harr20R` model resemble the ones in `Harr20`, which is the one with fixed beta coefficients (`Hutch78` = -1.62, `Har18` = 1.80, `Har19` = -0.089). The latter model puts more emphasis on roughness and less emphasis on the familiarity predictor. It is also worth pointing out that the familiarity predictor carries the dominant unique contribution in the model ($sr = 0.38$) whereas harmonicity plays a relatively small role ($sr = 0.10$), consistent with the analyses presented with the three datasets earlier. Looking at

the Eero21 model, the prediction rate has increased to 73% ($R^2 = 0.729$) and this improvement stems from a substantially higher unique contribution of the new harmonicity predictor (Stol15, $sr = 0.24$) and the new predictor category (spectral envelope, $sr = 0.05$) represented by spectral irregularity. Puzzlingly, the contribution of roughness has dwarfed (sr = -0.05) in this model, which could imply that the model components are volatile due to high multicollinearity. We will address this in the next analysis. Even though the model now has one more component, the improvement to the Harr20R model is statistically highly significant ($\chi^2 = 144.25$, $p < .001$) and exhibits lower AIC (1643) than the Harr20R model (1725).

Pooled Analysis with Principal Components. A close look at the models, and particularly the unique contributions of the feature categories, suggests that roughness and harmonicity may be highly collinear and this would hamper the interpretation of the component contributions. In the Harr20R model, the correlation between the roughness (Hutc78) and harmonicity (Harr18) predictors is -0.67 ($p < .001$). In a regression equation, this transforms into a variance inflation factor (VIF) of 2.08, which indicates that the variance of the predictor coefficient is over two times greater than it would be otherwise. In the Eero21 model, the correlation between the Hutch78 and the Stol15 predictors is -0.81 and the VIF in the regression is 3.16. There are different rules of thumb for the threshold of VIF values such as 2.5, 3, 5, and 10 (see Graham, 2003; Johnston et al., 2018; O'Brien, 2007) or correlations above 0.80 (e.g., Abu-Bader, 2016) that spell problems for separating out the independent contributions of the predictors. Here we take heed of the lower end of the recommendations (VIF > 3 and $r > |0.80|$) and also consider the conspicuous changes in the signs of the betas and the large variations in the unique contribution of the model categories related to harmonicity and roughness observed above (Tables 4 and 5). These variations are surprising, lead to drastically different interpretations of the model, and may signal that the predictor contributions are not well defined (O'Brien, 2007). To assess the contributions of the predictors representing these feature categories more robustly, we carried out a principal component analysis of all 22 features with the 617 chords. We started with the correlation matrix and estimated the number of components with parallel analysis that compares the eigenvalues from the principal component analysis to the similarly sized matrix of random data to estimate the chance level of eigenvalues (Zwick & Velicer, 1986). This analysis yielded three components as sufficient that accounted for 69% of the variance (34%, 21%, and 14% by each component) of the original correlation matrix. The first component captures both the roughness and harmonicity predictors, where harmonicity has high negative loadings while roughness has high positive loadings with the first component. The second component is related to the familiarity predictors and the third encapsulates the spectral envelope predictors (see Table 6).

Table 6. Loadings of the predictors from the principal component analysis ($n = 617$). Loadings under 0.55 are not shown.

Predictor	PC1	PC2	PC3
Vass01	0.90	-	-
Seth93	0.90	-	-
Wang13	0.87	-	-
Parn94	-0.86	-	-
Harr18	-0.80	-	-
Hutc78	0.77	-	-
Stol15	-0.74	-	-
Parn88	0.74	-	-
Miln13	-0.74	-	-
Bowl18	-0.63	-	-
Gill09	-0.56	-	-
SpecFlux	0.56	-	-
TonDiss	-	-	-
CorpClas	-	0.75	-
CorpJazz	-	0.80	-
Harr19	-	-0.89	-
CorpPop	-	0.93	-
SpecIrreg	-	-	-
SpecCentr	-	-	0.93
KeyClar	-	-0.63	-
SpecRolloff	-	-	0.89
SpecSharp	-	-	0.89
Var. expl.	34%	21%	14%

We used the scores of the three principal components as predictors in regression to predict C/D ratings in the pooled dataset using the same evaluation routine as with the other models (cross-validation and assessment of model prediction rate with the unseen data). This yields a model, labeled as *PCA components,* reported in Table 7, which puts the prediction rate at $R^2 = 0.67$. It is better than the Harr20R model ($\chi^2 = 49.22$, $p < .001$), but poorer than the Eero21 model. However, the most interesting part of this analysis is the unique contribution of the model components (sr) as there is zero correlation between the predictors. The semi-partial correlations suggest that familiarity accounts for the largest part of the variance in the model (sr = -0.68, 46.2% of variance), whereas the combined roughness and harmonicity component is the second major element (sr = -0.44, or 19.3% of variance). The spectral envelope component is left with negligible contribution ($sr = 0.04$, less than 0.2% of variance). The purpose of this analysis was to rethink the predictor category contributions and find a solution to handle the collinear feature categories of roughness and harmonicity. However, as a model of C/D, the model with PCA components is not a convenient one despite the lucrative independence of the components, as it needs a linear combination of all 22 predictors to create the model. An often used strategy (Jolliffe, 2002) is to represent the principal components with predictors that receive the highest loadings with the components. In this case, this would

Table 7. Results from two models related to principal component analysis showing prediction rates (R^2), standardized β, and semi-partial correlations (*sr*) for each predictor category.

	R^2	β_c	$\beta_{R/H}$	β_F	β_S	$sr_{R/H}$	sr_F	sr_S
PCA components	0.67	5.06	-1.02	-1.48	-0.10	-0.44	-0.68	-0.04
PCA predictors	0.65	5.05	-0.69	-0.24	-1.48	-0.27	-0.68	-0.10

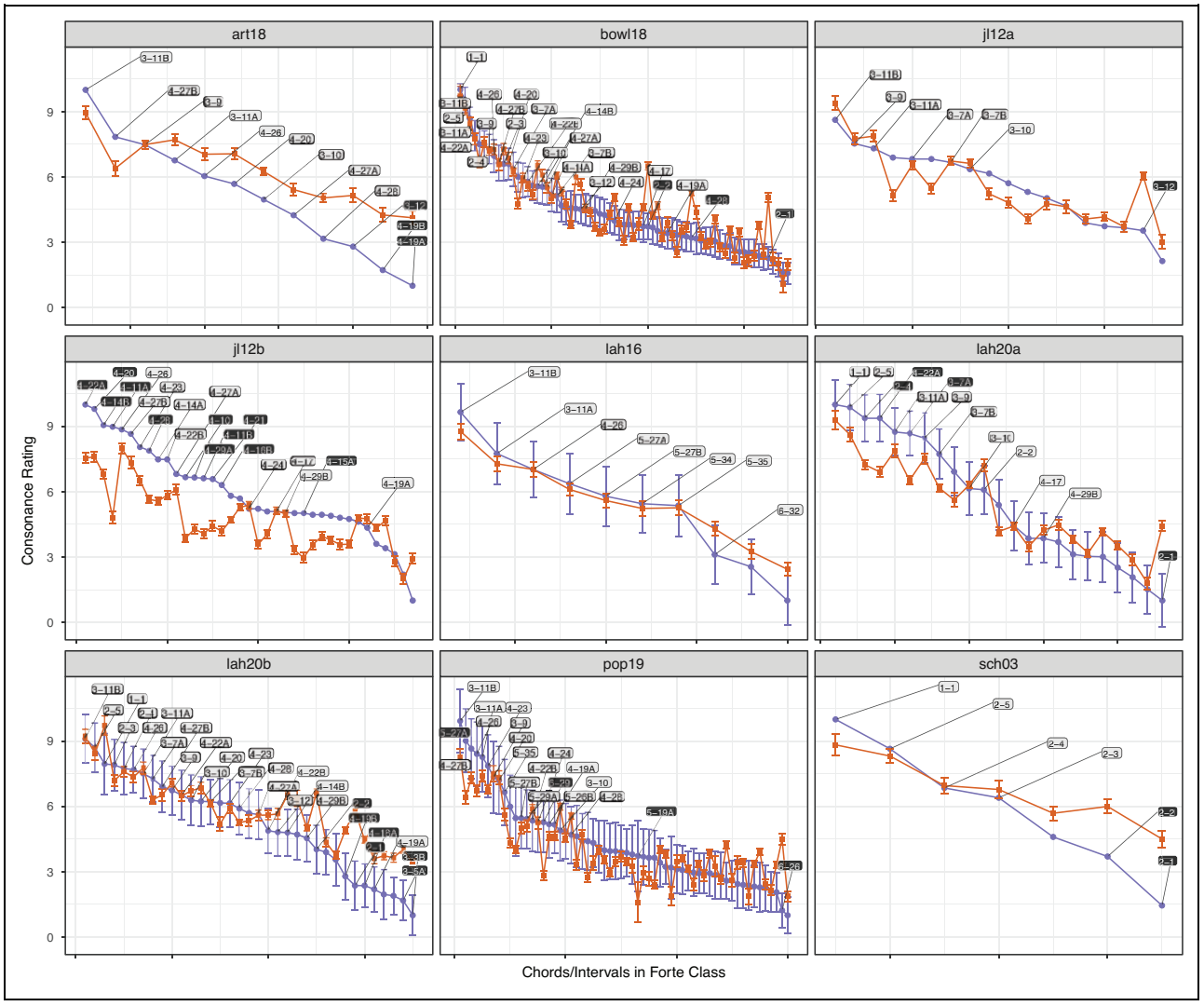

Figure 3. Model predictions and C/D ratings for unique Forte chords within each dataset. The red markers and lines indicate the model (`Eero21`) from the pooled analysis, showing also the 95% confidence intervals, and the blue markers and lines show ratings and standard deviations. The labels in black denote large ($> |2|$) prediction errors.

suggest a model where `Vass01`, `CorpPop`, and `SpCentr` are used as the proxies for the three components. The model with these predictors, called `PCApredictors` in Table 7, is able to achieve R^2 of 0.65 with minimal correlations between the predictors (all under $r = |0.19|$), although the model offers no improvement in terms of performance compared with the `Harr20R` model ($\chi^2 = 0$).

Going back to the most successful model, a visualization of the model prediction of the C/D ratings is shown in Figure 3, where the `Eero21` model is taken from the pooled analysis. The general pattern reflects the success of the model in picking up the variation in C/D ratings, although there are some curious and potentially systematic errors that none of the predictors are able to pick up. The visualization displays only the chords with unique Forte classes within each dataset. The chords predicted with an absolute error over 2 in the C/D ratings are shown with black labels for easier interpretation. Overall, the figure underscores that generally the model predicts the ratings reasonably well but there are specific failures. The first

observation is that these failures are not concentrated on one or two datasets. An analysis of the errors suggests that some of these may relate to familiarity, since several errors relate to intervals (m2/M7, M2/m7) that could probably be assessed more robustly than was done in the `CorpPop` model. Beside these failures, at this point we can only conclude that no other systematic errors are clearly apparent, but a more systematic analysis of the errors is probable the best way to explore any future gains when modeling C/D.

Discussion

The two regression analyses demonstrated a consistent pattern where the model proposed by Harrison and Pearce (2020) explains about 64% and 62% of the variance in C/D ratings in unpooled and pooled data, respectively, whereas the revised model with new predictors is able to account for about 10% more variance. Although the `Harr20R` model is satisfactory in many ways and combines the acoustic and cultural elements of C/D, the revision of the model is a clear improvement that arrives through better formulation of predictors related to familiarity as well as adding one missing element—spectral envelope—to the model.

A recent experiment by Lahdelma and Eerola (2020) also found roughness to be a more important predictor of C/D than harmonicity, corroborating the results of Harrison and Pearce (2020), but also found that cultural familiarity has a strong contribution to C/D ratings. Also, somewhat surprisingly, replacing the harmonicity predictor by Harrison and Pearce with another harmonicity model (i.e., Stolzenburg's model) had a large impact on the model and decreased the unique contribution of roughness in particular. This has an important implication here; a fairly small change in the actual predictors can have a knock-on effect on how the overall model operates, which in this case seemed to relate to an increased collinearity between roughness and harmonicity.

The additional analysis of the predictor matrix with the principal components addressed the problematically high collinearity between predictors representing roughness and harmonicity. Although the ensuing models with principal components or predictors best capturing the components scores did not improve the model beyond the level already offered with the `Eero21` model, the model with PCA components or predictors representing the components removed the collinearity of the predictors and allowed to estimate the independent contributions of the revised feature categories (roughness/harmonicity, familiarity, spectral envelope) to the variance in the C/D ratings. It remains to be seen whether the problematic multicollinear predictors can be set apart in other ways in future analyses or datasets. Manipulating the intonation of the intervals such as done with Bohlen-Pierce tuning could be one solution (Smit et al., 2019), and recent empirical data suggests that harmonicity and roughness exhibit lower correlations ($r(89) = .36$) in a sample of BP chords (Friedman et al., 2021). It might also be possible to postulate models that selectively apply to extremes of dissonance (roughness) or consonance (harmonicity) to try to avoid this conundrum. The alternative approach is to embrace the partial collinearity of these main elements of C/D and concede that roughness, harmonicity, and familiarity all undoubtedly reflect similar source constraints, namely the physics of natural, harmonically complex sounds and the properties of the auditory system, and how both have shaped musical conventions (Parncutt et al., 2019). The assumption that fully independent predictors can be developed is a challenge, but if the alternative is to keep working with two influential feature categories (roughness and harmonicity) that have avid proponents, conflicting accounts are bound to rise.

We revised and improved the familiarity measure by simplifying the classification of the chords. We believe that the elimination of chord inversions made the frequency distribution of the chords in the corpus more consistent and closer to the perceptual assessment of the chords. It remains to be seen whether a better corpus could be established, and even better if the unit of analysis in the tabulation of chords corresponds to the way listeners recognize the familiarity of the chords and intervals. It would be beneficial for future endeavors in modeling C/D if familiarity could also rely on acoustic properties, as it would eliminate any need for symbolic representation in the models and would allow for a greater range of tunings, timbres, and musical conventions to be readily applied to the model.

We brought the spectral envelope as a new element to C/D and it turned out to be a significant, albeit small, addition to the model. However, we think that spectral envelope might play a more pronounced role once the stimulus materials in future empirical work span a wider range of registers and different timbres. This assumption is based on previous literature where, for instance, sharpness has been implicated as an important factor contributing to consonance/dissonance perception in addition to roughness and tonalness, that is, periodicity/harmonicity (see Zwicker & Fastl, 1990, p. 313). Moreover, Lahdelma and Eerola (2016) have empirically demonstrated that chords played in a higher register tend to be perceived as more dissonant than chords in a lower register. In their study, this observation was explained indeed with the effect of sharpness: the higher-register chords were lower in roughness but higher in sharpness compared to the lower-register chords, where the ratio between these two acoustic factors was the opposite.

Conclusions

The musical and acoustical aspects of C/D have not yet been fully accounted for. Despite the impressive and systematic work by Harrison and Pearce (2020), the present study is able to offer novel elements to this question and

also to expose some shortcomings in the state-of-the-art C/D research. We do this using an unprecedented amount of data that consists of three individual-level datasets and nine datasets with mean ratings encompassing 600+ stimuli which together span an excellent variety of chords and intervals. In the process, we offer several new predictors of C/D representing familiarity and a new category of predictor in the form of spectral envelope. The new predictors, namely `CorpPop` (familiarity) and `SpIrreg` (spectral envelope), made their way into the new model and offered substantial improvements over the past predictors in a rigorous series of statistical comparisons. We also release a new dataset, labeled the Durham Chord Dataset (DCD), which helped us to assess the underlying structure behind the large set of proposed predictors. We believe this dataset can also stimulate future research and we encourage its use by researchers interested in the research topic at hand.

In all of these datasets, the central problem for modeling C/D is the high collinearity between the predictors, mainly between roughness and harmonicity. When roughness was represented with the model by Hutchinson and Knopoff (1978) and harmonicity with Stolzenburg's model (2015), the correlations between these two variables was -0.65, -0.83, and -0.81 in Experiments 1, 2, and 3. The attempt to untangle their unique contribution to C/D ratings with partial and semi-partial correlations in Experiments 2 and 3 needs to keep in mind the caveat relating to the collinearity between the variables, which renders the interpretation of the unique contributions of the predictors rather volatile. In this study, the decision was to take the feature categories as given and analyze the feature category membership with hierarchical cluster analysis in Experiment 1, which preserved the separation of roughness and harmonicity mainly because the predictors in these two categories have the opposite signs in the correlations. In Experiment 2, any predictor that improved the prediction rate within the feature category was taken as a better alternative predictor of that feature category. In Experiment 3, an alternative feature reduction was carried out, which first identified three independent principal components that captured 69% in the covariance of the original 22 predictors. When these three components were used in linear regression, they explained 66% of the variance in C/D ratings. Most importantly, the component contributions in the regression analysis suggested that the component that captures familiarity accounts for 46.2% ($sr = 0.68$) of the variance, whereas roughness/harmonicity represented by a single component accounted for 19.4% of the variance ($sr = -0.44$). The results of this alternative analysis offer an interesting simplification of the feature categories that allows to keep the categories unassociated from each other. However, the model based on the principal component analysis is not too elegant, simple, nor is it the best model in this data, but it points to a possible way to eliminate redundant categories of C/D features in future analyses. Further research might identify other ways to segment C/D into meaningful elements.

While it is early to decisively conclude the exact anatomy of C/D, the current investigation has offered new perspectives to the topic. If future research wishes to pursue the independent feature categories, we may have come full circle in identifying the anatomy of C/D; as Johnson-Laird et al. (2012) point out, von Helmholtz drew the conclusion that the perception of C/D is dependent both on a psychoacoustic and on a cultural factor. Also, excluding spectral envelope from future models of C/D might be ill-advised, and there is still a host of additional factors to be explored, including loudness, which might influence dissonance ratings (see Kameoka & Kuriyagawa, 1969; Mashinter, 2006). We hope that the present findings inspire the field to investigate the topic with a more versatile set of stimuli (e.g., multiple registers, dynamics, timbres) and to continue to use open datasets and libraries to further tease apart the roles of acoustic and cultural predictors in the fascinating question of consonance and dissonance.

Author contribution

TE and IL conceived the study, TE compiled the datasets and extracted the model predictions, and carried out the analyses, and wrote the first draft of the manuscript. IL contributed to the writing of the manuscript, collapsed the chord inversions and encoded the revised chord corpus with Forte numbers, and created the Durham Chord Dataset piano stimuli.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Tuomas Eerola https://orcid.org/0000-0002-2896-929X

Action editor

David Meredith, Aalborg University, Department of Architecture, Design and Media Technology.

Peer review

Peter Harrison, Max Planck Institute for Empirical Aesthetics.
Andrew Milne, Western Sydney University, MARCS Institute for Brain, Behaviour and Development.

References

Abu-Bader, S. H. (2016). *Advanced and multivariate statistical methods for social science research with a complete SPSS guide*. Oxford University Press.

Arthurs, Y., Beeston, A. V., & Timmers, R. (2018). Perception of isolated chords: Examining frequency of occurrence,

instrumental timbre, acoustic descriptors and musical training. *Psychology of Music*, *46*(5), 662–681.

Aures, W. (1985). Ein Berechnungsverfahren der Rauhigkeit (A procedure for calculating auditory roughness). *Acustica*, *58*, 268–281.

Bowling, D. L., & Purves, D. (2015). A biological rationale for musical consonance. *Proceedings of the National Academy of Sciences*, *112*(36), 11155–11160.

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.

Burgoyne, J. A. (2012). *Stochastic processes and database-driven musicology*. PhD Thesis, McGill University, Montréal, Québec.

Churchill, C. (2004). Loudness and sharpness calculation (matlab scripts). URL http://hub.salford.ac.uk/sirc-acoustics/psychoacoustics/sound-quality-making-products-sound-better/an-introduction-to-sound-quality-testing/matlab-codes/

Cook, N. D. (2017). Calculation of the acoustical properties of triadic harmonies. *The Journal of the Acoustical Society of America*, *142*(6), 3748–3755.

Cuthbert, M. S., & Ariza, C. (2010). music21: A toolkit for computer-aided musicology and symbolic music data. In J. S. Downie & R. C. Veltkamp (Eds.), *Proceedings of the 11th international society for music information retrieval conference (ISMIR 2010)* (pp. 637–642). International Society for Music Information Retrieval.

Forte, A. (1973). *The structure of atonal music*. Yale University Press.

Friedman, R. S, Kowalewski, D. A, Vuvan, D. T, & Neill, W. T. (2021). Consonance preferences within an unconventional tuning system. *Music Perception: An Interdisciplinary Journal*, *38*(3), 313–330.

Gill, K. Z, & Purves, D. (2009). A biological rationale for musical scales. *PLoS One*, *4*(12), e8144.

Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, *84*(11), 2809–2815.

Harrison, P., & Pearce, M. (2018). An energy-based generative sequence model for testing sensory theories of Western harmony. In *Proceedings of the 19th international society for music information retrieval conference* (pp. 160–167). Paris, France, 23–27 September 2018.

Harrison, P., & Pearce, M. (2020). Simultaneous consonance in music perception and composition. *Psychological Review*, *127*(2), 216–244.

Helmholtz, H. L. F. (1875). *On the sensations of tone as a physiological basis for the theory of music*. Longman.

Herbst, J. P. (2018). Heaviness and the electric guitar: Considering the interaction between distortion and harmonic structures. *Metal Music Studies*, *4*(1), 95–113.

Huron, D. (1991). Tonal consonance versus tonal fusion in polyphonic sonorities. *Music Perception*, *9*(2), 135–154.

Huron, D. (1994). Interval-class content in equally tempered pitch-class sets: Common scales exhibit optimum tonal consonance. *Music Perception: An Interdisciplinary Journal*, *11*(3), 289–305.

Hutchinson, W., & Knopoff, L. (1978). The acoustic component of Western consonance. *Interface*, *7*(1), 1–29.

Järvinen, T. (1995). Tonal hierarchies in jazz improvisation. *Music Perception*, *12*(4), 415–437.

Jensen, K. (1999). *Timbre models of musical sounds*. PhD Thesis, Department of Computer Science, University of Copenhagen.

Johnson-Laird, P. N., Kang, O. E., & Leong, Y. C. (2012). On musical dissonance. *Music Perception: An Interdisciplinary Journal*, *30*(1), 19–35.

Johnston, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression analysis: A cautionary tale and an alternative procedure, illustrated by studies of british voting behaviour. *Quality & Quantity*, *52*(4), 1957–1976.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Kameoka, A., & Kuriyagawa, M. (1969). Consonance theory part I: Consonance of dyads. *The Journal of the Acoustical Society of America*, *45*(6), 1451–1459.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes. *II. Analyses acoustiques et quantification psychophysique. Le Journal de Physique IV*, *4*(C5), C5–625.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford University Press.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334–368.

Lahdelma, I., Armitage, J., & Eerola, T. (2020). Affective priming with musical chords is influenced by pitch numerosity. *Musicae Scientiae*, URL https://doi.org/10.1177/1029864920911127

Lahdelma, I., & Eerola, T. (2016). Mild dissonance preferred over consonance in single chord perception. *i-Perception*, *7*(3), 1–21.

Lahdelma, I., & Eerola, T. (2020). Cultural familiarity and musical expertise impact the pleasantness of consonance/dissonance but not its perceived tension. *Scientific Reports*, *10*, 8693.

Lartillot, O., & Toiviainen, P. (2007). *A matlab toolbox for musical feature extraction from audio* (pp. 237–244). International conference on digital audio effects. Bordeaux.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 261–268). Springer.

Mashinter, K. (2006). Calculating sensory dissonance: Some discrepancies arising from the models of Kameoka & Kuriyagawa, and Hutchinson & Knopoff. *Empirical Musicology Review*, *1*(2), 65–84. https://doi.org/10.18061/1811/24077

McLachlan, N., Marco, D., Light, M., & Wilson, S. (2013). Consonance and pitch. *Journal of Experimental Psychology: General*, *142*(4), 1142–1158.

Milne, A. J. (2013). *A computational model of the cognition of tonality*. PhD Thesis, The Open University.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 20170213.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690.

Parncutt, R. (1988). Revision of Terhardt's psychoacoustical model of the root(s) of a musical chord. *Music Perception*, *6*(1), 65–93.

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

Parncutt, R. (2006). Commentary on Cook & Fujisawa's "The psychophysics of harmony perception: Harmony is a three-tone phenomenon". *Empirical Musicology Review*, *1*(4), 204–209.

Parncutt, R. (2014). The emotional connotations of major versus minor tonality: One or more origins? *Musicae Scientiae*, *18*(3), 324–353.

Parncutt, R., & Hair, G. (2011). Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies. *Journal of Interdisciplinary Music Studies*, *5*(2), 119–166.

Parncutt, R., Reisinger, D., Fuchs, A., & Kaiser, F. (2019). Consonance and prevalence of sonorities in Western polyphony: Roughness, harmonicity, familiarity, evenness, diatonicity. *Journal of New Music Research*, *48*(1), 1–20.

Parncutt, R., & Strasburger, H. (1994). Applying psychoacoustics in composition: "harmonic" progressions of "nonharmonic" sonorities. *Perspectives of New Music*, *32*(2), 88–129.

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

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*, 1070.

Rameau, J. P. ([1722] 1971). *Treatise on harmony*. Dover Publications.

Schwartz, D. A., Howe, C. Q., & Purves, D. (2003). The statistical structure of human speech sounds predicts musical universals. *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.

Sethares, W. A. (2005). *Consonance and dissonance of harmonic sounds. Tuning, timbre, spectrum, scale* (pp. 77–95). Springer.

Smit, E. A., Milne, A. J., Dean, R. T., & Weidemann, G. (2019). Perception of affect in unfamiliar musical chords. *PLos One*, *14*(6), e0218570.

Stolzenburg, F. (2015). Harmony perception by periodicity detection. *Journal of Mathematics and Music*, *9*(3), 215–238.

Stumpf, C. (1898). Konsonanz und dissonanz (Consonance and dissonance). *Revue Philosophique de la France Et de l'Etranger*, *46*, 184–188.

Temperley, D., & Clercq, T. d. (2013). Statistical analysis of harmony and melody in rock music. *Journal of New Music Research*, *42*(3), 187–204.

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.

Terhardt, E. (1982). Die psychoakustischen grundlagen der musikalischen akkordgrundtöne und deren algorithmische bestimmung. In C. Dahlhaus & M. Krause (Eds.), *Tiefenstruktur der Musik, volume Tiefenstruktur, chapter Die psychoakustischen Grundlagen der musikalischen Akkordgrundtöne und deren algorithmische Bestimmung* (pp. 23–50). Technical University of Berlin.

Terhardt, E. (1984). The concept of musical consonance: A link between music and psychoacoustics. *Music Perception: An Interdisciplinary Journal*, *1*(3), 276–295.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Vassilakis, P. N. (2001). *Perceptual and physical properties of amplitude fluctuation and their musical significance*. PhD Thesis, University of California, Los Angeles.

Wang, Y., Shen, G., Guo, H., Tang, X., & 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.

Zajonc, R. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, *10*(6), 224–228.

Zwick, W., & Velicer, W. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442.

Zwicker, E., & Fastl, H. (1990). *Psychoacoustics: Facts and models*. Springer-Verlag.