Cognitive Music Listening Space: A Multivariate Approach

Brendon Mizener[1], Mathilde Vandenberghe[2], Hervé Abdi[1], & Sylvie Chollet[2]

[1] University of Texas at Dallas

[2] Junia, Univ. Artois, Université de Liège, Univ. Littoral Côte d'Opale, UMRT 1158
BioEcoAgro, F-62000 Arras, France

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Brendon Mizener: Stimuli creation, Survey design & creation, Data collection & processing, Statistical analyses, Writing - Original draft preparation; Mathilde Vandenberghe: Original concept, Survey design & creation; Hervé Abdi: Writing - Review & Editing, Statistical guidance; Sylvie Chollet: Original concept.

Correspondence concerning this article should be addressed to Brendon Mizener, 800 W. Campbell Rd., Richardson Tex. E-mail: bmizener@utdallas.edu

Abstract

French and American participants listened to new music stimuli and evaluated the stimuli using either adjectives or quantitative musical dimensions. Results were analyzed using Correspondence Analysis (CA), Hierarchical Cluster Analysis (HCA), Multiple Factor Analysis (MFA), and Partial Least Squares Correlation (PLSC). All except the HCA used Bootstrapping and Permutation testing for inferences. French and American listeners differed when they described the musical stimuli using adjectives, but not when using the quantitative dimensions. The present work serves as a case study in research methodology that allows for a balance between relaxing experimental control and maintaining statistical rigor.

Keywords: Music, Perception, Cognition, Multivariate Analyses

Word count: 5631

Cognitive Music Listening Space: A Multivariate Approach

We have a data collection problem: World events over the last year have shown that we need to be able to collect good data outside of the lab. In the lab, because we control error sources, we measure, on relatively small sets of observations, a few well-defined, quantitative variables, analyzed using standard techniques such as analysis of variance (ANOVA). But, with the labs closed (remember COVID?), how can we collect good data? Away from the controlled environment of the lab, quantitative variables are hard to measure, but we can collect, on large sets of observations, qualitative variables that can only be analyzed by newer multivariate techniques. In the present paper, we present a case study illustrating this tradeoff.

"Doesn't beer taste better in a bar? Or when you're listening to your favorite song?" The present study was designed to quantify a music listening 'space' that captures objective stimulus and cognitive dimensions for the sake of investigating cross-modal sensory mapping between beer drinking and music listening. It also addresses other questions: Are there differences in how people from different countries — and by extension musical cultures — perceive and describe music? What parallels exist between stimulus dimensions and cognitive dimensions of music?

For the present study, we have defined stimulus dimensions as quantitative musical qualities, such as tempo, range, and meter and cognitive dimensions as qualitative descriptions of music, such as "dark," "warm," and "round." These cognitive/qualitative dimensions are similar to the commonly investigated affective or emotional dimensions, but do not specifically assess affective quality. We designed three experiments to quantify individual and combined spaces for these concepts, using separate surveys. The first experiment included highly trained musicians and featured a simple multiple choice survey about the stimulus dimensions; the second included participants with any level of music training performing a check-all-that-apply task (CATA, Katz & Braly, 1933; Meyners & Castura, 2014; Coombs et al., 1956); the third experiment incorporated both surveys in a single analysis.

To analyze our data, we selected a set of multivariate analyses that allowed us to visualize answers to each of our questions. The mental spaces revealed by the individual surveys were calculated and visualized using Correspondence Analysis (CA), a method similar to Principal Components Analysis (PCA) that analyzes qualitative data. We used Multidimensional Scaling (MDS), a distance analysis, to visualize the differences between participants and participant groups. To find parallels between the surveys, we used Partial Least Squares Correlation (PLSC), a method that analyzes two data tables with different sets of variables measured on the same observations. We used a Multiple Factor Analysis to evaluate how French and American participants' responses differed. Each of these analyses provide different visualizations and interpretations of the data, which are discussed in more detail below.

Music Perception

Quantifying music perception is an interesting test case for this kind of data gathering and analytical paradigm. Most music or auditory perception studies have the inherent confound that small changes can affect listeners' perception, especially when the study involves timing, tuning, or sound localization. However, the experimental controls may be loosened slightly when investigating holistic music listening, because no single musical element is more important than the whole.

Quantitative and qualitative elements of music are theoretically distinct but practically inseparable (Bruner II, 1990). Listeners respond affectively to technical aspects of music using schemata informed by their individual musical experiences and personality traits (Kopacz, 2005), and composers use various musical and compositional techniques to convey the emotions they want to express (Battcock & Schutz, 2019; Bruner II, 1990). However, quantifying the perceptual interactions between more than one or two musical qualities is a challenge. One reason is that models like ANOVA and its variations are limited by how many variables a researcher can include while remaining coherent. Another is that asking participants to respond to multiple aspects of a stimulus taxes participants' perceptual capacity and is thus difficult to measure (W. F. Thompson, 1994).

One specific area that has attempted to capture a greater dimensionality is music emotion research. This is a well trod domain — see, for example Juslin and Sloboda (2010) — and the application of multivariate analyses to these questions is similarly well established. Early studies, including Gray and Wheeler (1967), Wedin (1969), and Wedin (1972) used MDS to capture the affective space of various musical stimuli. MDS continues to be used commonly in more modern studies (Bigand et al., 2005; Madsen, 1997; Rodà et al., 2014), with a narrow focus on valence and arousal, which were first proposed to be the most salient dimensions of perception by Osgood and Suci (1955).

A few studies have specifically investigated dimensions beyond those first two (for example Rodà et al., 2014), and there is recent conflicting evidence as to whether the valence-arousal plane represents the fundamental dimensionality behind music emotion perception (Cowen et al., 2020). Some hypotheses suggest there are "at least 13 dimensions" (Cowen et al., 2020) to the subjective music and emotion perceptual space (Juslin & Västfjäll, 2008). However, an important distinction between the present study and work in music emotion perception is that the adjectives we chose were informed by music composition and performance, rather than emotion (Wallmark, 2019).

There are many studies that evaluate the differences between trained and untrained musicians. The verdict as to whether trained musicians are better music listeners is still out, partially due to the fact that there is little standardization in how much training is required for a participant to be "highly trained" (Bigand & Poulin-Charronnat, 2006). There are, however, reported benefits with regard to sensitivity to the emotional content in music (Ladinig & Glenn Schellenberg, 2012) and familiarity with tonal systems (Bartlett & Dowling, 1980; Dowling, 1978). Recent works suggest that these benefits may be limited to specific technical aspects, and depend on the extent of training (Raman & Dowling 2017). We included highly trained musicians because they are sensitive to these technical aspects of music and will be able to accurately quantify the stimuli. Additionally, some of the response options to questions on the survey for Experiment 1 would only be familiar to participants with significant music training.

Intercultural music perception. There are a few common goals in intercultural studies of music perception. Some quantify the shared emotional experience between musical cultures (L. L. Balkwill et al., 2004; L. Balkwill & Thompson, 1999; Cowen et al., 2020; Darrow et al., 1987; Fritz et al., 2009; Gregory & Varney, 1996), and some ask participants to identify technical aspects of music from other cultures (Raman & Dowling, 2016, 2017). There are fewer studies that include semantics in their evaluation of music perception (Zacharakis et al., 2014, 2015), which makes this a prime area for research.

The research program presented in Zacharakis et al. (2014) and Zacharakis et al. (2015) deals specifically with timbre perception, and their use of adjectives is similar to the way they adjectives are used in the present study. In Zacharakis et al. (2014, 2015), Greek and English participants described timbre with adjectives from their native languages. These studies found that while there are some differences, overall, participants' descriptions of timbre do not differ much between languages (Zacharakis et al., 2014, 2015).

Present questions & methods of analysis

The primary question addressed in this study is: Can we quantify a cognitive space around music listening defined by both stimulus and cognitive dimensions of music. Secondary questions include whether French and American participants describe music differently, and whether those differences may arise from cultural differences or are purely semantic. To answer these questions, we employed a set of multivariate analyses that each offered a different perspective on the results of each experiment. We felt it may be useful to provide a quick overview of the data collection and analytical techniques for readers who may be unfamiliar.

Check-all-that-apply (CATA). The CATA technique — a method widely used in sensory evaluation — measures how participants describe a set of stimuli (Coombs et al., 1956; Katz & Braly, 1933; Meyners & Castura, 2014; Valentin et al., 2012). In a CATA task, stimuli are presented one at a time, and for each stimulus, participants are shown a list of descriptors and are asked to select the descriptors that describe the presented stimulus (Meyners & Castura, 2014). CATA easily assesses questions with multiple 'correct' responses (Coombs et al., 1956), and places little cognitive demand on participants because they do not have to generate responses (Ares et al., 2010).

CATA data are analyzed by 1) computing a pseudo contingency table that records the number of times descriptors were associated with stimuli and 2) analyzing this data table with Corresponence Analysis (CA) in order to visualize the patters of association between a) stimuli, b) descriptors, and c) stimuli and descriptors.

Correspondence Analysis. The primary analysis used on the data collected in the surveys is Correspondence Analysis (CA) (Benzécri, 1973; Escofier-Cordier, 1965; Greenacre, 1984). CA analyzes a contingency table, or any data structured similarly, and calculates the relationships between rows (observations) and columns (variables); in our case, musical excerpts and descriptors. The nature of the CA allows for observations and variables to be visualized in the same space using biplots.

Partial Least Squares Correlation. Partial Least Squares Correlation (PLSC) (Abdi & Williams, 2013; Tucker, 1958) analyzes two data tables that describe a single set of observations (rows) with different sets of variables (columns). PLSC computes a matrix of correlations between the sets of variables which is then analyzed to find latent variables with the largest covariance, i.e., the greatest amount of information common to the two data tables. This technique is commonly used in neuroimaging studies to evaluate correlations between matrices of imaging data and of behavioral or task data (Krishnan et al., 2011).

Multidimensional Scaling. Multidimensional Scaling (MDS) (Borg & Groenen, 2005) — a technique commonly used in music perception studies (Bigand et al., 2005; Madsen, 1997; Rodà et al., 2014; Wedin, 1969, 1972) analyzes a distance matrix computed between observations and visualizes them, positioning these observations on a map such that the distance between observations on the map best approximates their distance in the data table.

Multiple Factor Analysis. Multiple Factor Analysis (MFA) (Abdi et al., 2013) is an extension of PCA that analyzes and visualizes multiple tables or sets of variables that each describe the same observations. MFA visualizations are focused on the relationships between observations, and, for each observation, the relationships between the tables that contributed to that observation. Practically speaking, the most basic difference between MFA and PLSC is that PLSC extracts commonalities between two data tables, whereas MFA extracts similarities and differences between two or more data tables.

Inference Methods. Because the methods outlined above are not inferential methods, and do not inherently allow for hypothesis testing, we use permutation testing (Berry et al., 2011) and bootstrapping (Hesterberg, 2011). Permutation testing evaluates whether our data have a signal that is more salient than a random table and bootstrapping evaluates the stability of the result of an experiment with confidence intervals computed from resampling the original observations (Berry et al., 2011; Hesterberg, 2011).

Data Processing. Raw data were cleaned and processed in Excel and R. As described above, each of the analyses requires data in a different format. CA requires a contingency table or something like it, PLSC requires two tables with the same rows, MFA requires multiple tables with the same rows, and MDS requires a distance matrix. To transform the data into these formats, we first translated all French responses into English. Then the response data for each participant was transformed into a pseudo contingency table, with the stimuli, as observations, on the rows, and the responses, as variables, on the columns. This way, instead of a table with qualitative information in each cell, we had a table with the response options as variables and ones and zeros as counts in each cell. These individual tables were then compiled into three-dimensional arrays, or "bricks," of data, one for each survey, such that each "page" was an individual participant's responses. To get the contingency table for CA, we summed each array along the pages into a pseudo-contingency table, which contained the responses as count data. The CAs for the experiments were performed on each pseudo-contingency table separately, and the PLSC was performed using both at the same time. To get the distance matrix for the MDS, we computed a distance matrix for each of the bricks, so that the distance in each cell represented the distance between participants.

We only performed one MFA, on the results of the AS. To get the tables for MFA, we separated the AS brick into separate bricks for French and American participants and summed each along the pages, resulting

To establish groups for inferences and visualization, the variables from the QS were grouped by quality (e.g., levels of tempo, types of genre). However, since we did not use *a priori* grouping variables for the excerpts or adjectives, the pseudo-contingency tables were evaluated using hierarchical cluster analyses (HCA; see supplementary materials) to see what groupings arose during evaluation.

Experiment 1: Musical Qualities Survey

Methods

Participants. For the first experiment, we recruited highly trained musicians with a minimum of 10 years of formal music training to evaluate the stimulus dimensions or musical qualities of the stimuli, and to ascertain whether the stimuli truly reflected the composer's intent of varying on a wide range of musical dimensions (Raman & Dowling, 2017). Participants in the United States and in France were recruited by word of mouth and social media. There were a total of 84 responses to the survey, of which 51 were removed to incomplete data, leaving a total of 27 ($N_F = 9$, $N_A = 18$) for the analysis. All recruitment measures were approved by the UT Dallas IRB.

Stimuli. All stimuli were new, original musical excerpts, in various Western styles, composed by the first author specifically for this study (scores and audio files available upon request). The stimuli were all string quartets, designed to control for the confounding factor that different instruments are described in different ways (Zacharakis et al., 2014, 2015) and otherwise vary along a number of stimulus dimensions. The stimuli were composed to be coherent snippets approximately 30 s in length (actual range: 27 - 40 s, *M* = 32.4 s). Stimuli were composed using Finale composition software (Finale v25, MakeMusic, Inc.). Each stimulus was a wav file generated using the Finale human playback engine and embedded into a qualtrics question in that format.

Survey. Participants in Experiment 1 completed a survey (hereafter: Qualities Survey/QS) on Qualtrics that evaluated the musical stimuli on ten music stimulus dimensions such as tempo, range, and meter and gathered demographic data, including age, gender identity, nationality, occupation, and musical experience. The qualities assessed in the QS were selected from standard music-theoretical descriptors of western music. With the exception of styles of articulation, which were presented in traditional (Italian) musical terminology, all of the responses for each quality were presented in the vernacular, either French or English, depending on the nationality of the participant. For example, when rating the excerpts on tempo, participants were asked to rate the excerpt using the scale *Very Slow*, *Slow*, *Moderately Slow*, *Moderate*, *Moderately Fast*, *Fast*, and *Very Fast* (French: *Très lente*, *lente*, *moyennement lent*, *moyenne*, *moyennement rapide*, *rapide*, *Très rapide*). The full list of musical qualities and answer choices is listed in the supplementary materials.

Procedure. Participants were provided with a link to the survey and were instructed to listen to the excerpts presented in the survey either using headphones or in a quiet listening environment, but that was not controlled, nor was it assessed as part of the survey. After standard informed consent procedures, participants listened to 15 excerpts, presented one at time in a random order, and 10 questions per excerpt, one for each of the musical qualities being assessed. Demographic survey questions followed the experimental task.

Results

Participants. The scree plot in Figure 1 shows the eigenvalues for the MDS on the musical experts. The first two dimensions have $\lambda = 9.06$ and $\lambda = 7.52$, respectively. There was no permutation test performed for this analysis because it was exploratory in nature. There are no especially prominent dimensions, which suggests that all of the participants responded similarly to the excerpts. Visualizing the factor scores with means and confidence intervals revealed no significant differences between the experts based on any of the grouping variables used. Figure 2 show how the means of the factor scores, grouped by nationality and gender identity, respectively, are clustered on top of one another at the origin. The overlapping ellipses are the confidence intervals for the means. The individual data points are the same between the two plots, but the color of the dots, means and confidence intervals are different because they represent different grouping variables.

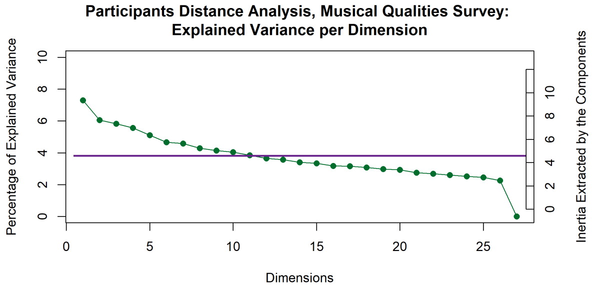


Figure 1

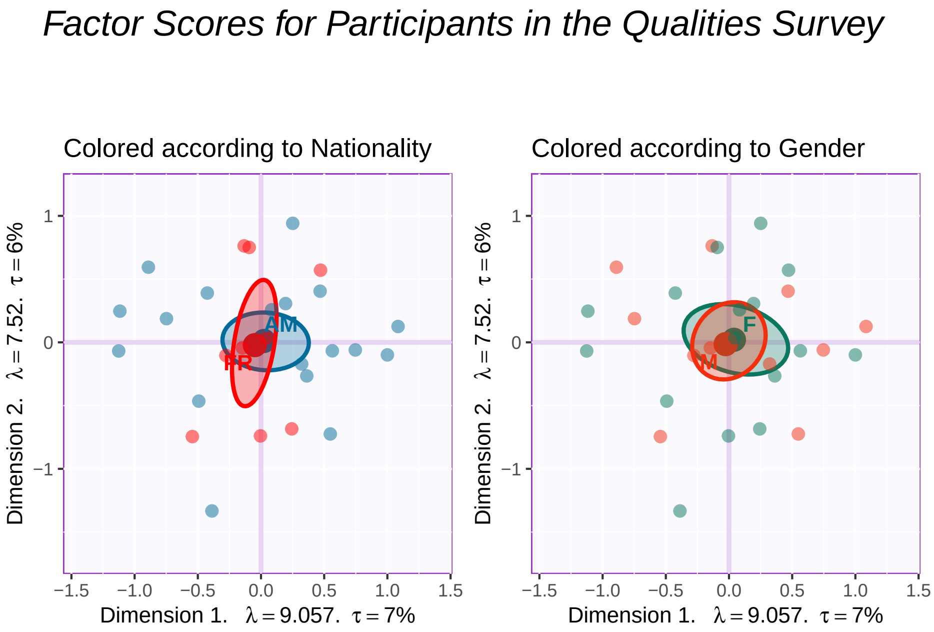


Figure 2

Excerpts.

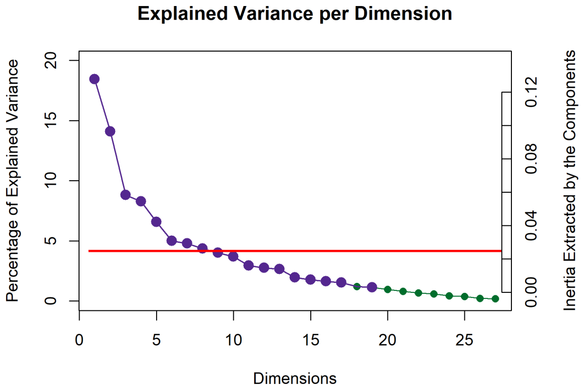


Figure 3

The scree plot for the analysis of the QS, Figure 3, shows the high dimensionality of this space, with the first four dimensions extracting a total of 49.63% of the variance. Significant dimensions (i.e., dimensions that represent something other than noise) are indicated in purple.

Graphing the contributions (see Figure 4) shows which excerpts and qualities contribute significantly — defined as more than the mean for each dimension — to the first two dimensions. The excerpts are grouped and colored according to the HCA (see supplementary materials). Contributing positively to dimension 1 are Excerpts 4, 13, 23, and 26, and contributing negatively are Excerpts 3, 7, 10, 24, and 27. Tempo, articulation, and dynamics variables seem to define the first dimension. The tempo variables contribute from low (tempo.F2 and tempo.F1) in the negative direction to high (tempo.F6 and tempo.F7) in the positive direction, with articulations from smooth (legato) to separated (marcato and staccato) and dynamics soft to loud trending the same way. Single variables from other groups also contribute: major harmony, triple meter, classical genre, undulating contour, and disjunct motion. The pattern for these variables and their groups may be clearer on the factor plots below.

The second dimension is dominated by a few groups of variables: harmony, meter, genre, dynamics. The single variables are slow tempo, ascending contour, and "no melody." The excerpts that contribute significantly in the positive direction are Excerpts 7, 12, 15, 27, and 19. In the negative direction are Excerpts 2, 3, 11, and 17. A full enumeration of contributions and bootstrap ratios is available at the URL in the author note.

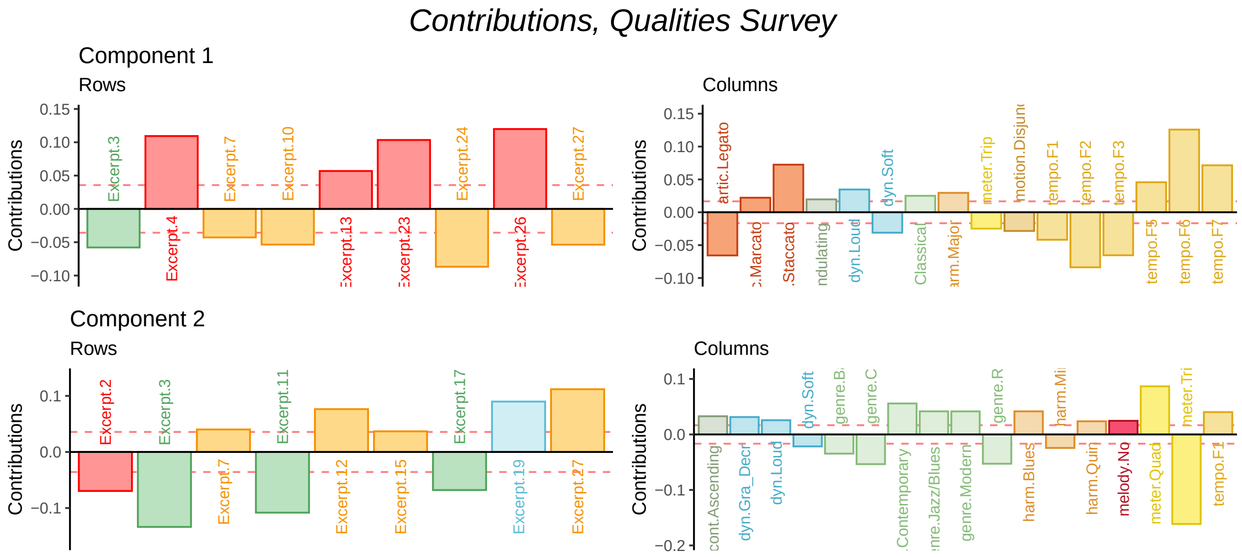


Figure 4

Discussion

The graph depicted in Figure 5 is a symmetric biplot depicting how excerpts and variables appear in the same mental space. The distance between the excerpts can be interpreted directly as similarity, and the distance between the musical qualities can be interpreted directly as similarity, but the distance between a quality and an excerpt cannot. Instead, the angle between an excerpt and a quality is indicative of their correlation. An angle of 0 or 180 degrees indicates a perfect positive or negative correlation, respectively, and an angle of 90 degrees indicates no correlation — the two items share no information. See Abdi and Williams (2010) for a more in-depth discussion.

Overall, this helps us to evaluate what qualities contribute to the excerpt groupings. The first dimension has a clear arousal trend — tempo, articulation, and dynamics all load from greater arousal to lesser on the first dimension. The second dimension is less clear, and does not seem to be tied to valence. Minor and major harmony, for example, both score negatively on dimension two. Both meter and dynamics are juxtaposed on the second dimension, with loud and quadruple meter scoring positively and soft and triple meter scoring negatively.

Preliminary visualizations (see supplementary materials) revealed six clusters grouped roughly by genre, with two notable outliers, Excerpts 6 and 14, in their own clusters, each the sole representative of their genres. Excerpt 6 is minimalist, à la Steve Reich, and Excerpt 14 is jazzy. These two were removed from the analysis and included instead as supplementary projections, essentially ‘out of sample’ elements.

There are a few musical connections revealed in this plot. Staccato articulations, correlated on this factor plot with high tempos, are played light and separate, and legato articulations, correlated with slow tempos, are played smooth and connected. In terms of performance practice, slow and long notes are played in a legato style to create a sense of continuity, and fast moving notes or phrases do not require the same technique. This plot also reveals the connection between genre and harmony (Cohn et al., 2001); the coordinate mapping of jazz/blues harmony and genre, which are stacked right on top of one another is the most extreme example of this. Other connections are also revealed, with older styles, such as baroque, classical, and romantic, and simpler harmonies of major and minor scoring negatively on the second dimension, while the newer styles, impressionist, modern, and contemporary, score positively on the second dimension, along with the more complex harmonies of chromatic, whole tone, and ambiguous. This follows historical practice: The Classical era had fairly structured rules for both harmony and voice leading, but the Romantic era relaxed those rules and introduced more complex harmonies (Cohn et al., 2001). The gradual devolution of those rules and the increase in complexity of harmony continued through the modern and contemporary styles (Kennedy et al., 2013). The whole tone scale, for example, wasn't used commonly until the impressionist era (Cohn et al., 2001).

Finally, we note that because of the nature of this survey, these results tell us more about the excerpts themselves rather than the behavior of the participants. The participants rated the stimuli similarly, validating the variety among the excerpts, indicating that they are different enough to create a large and varied factor space.

Experiment 2: Musical Adjectives Survey

Methods

Participants. Participants with self-reported normal hearing were recruited for Experiment 2 without regard to level of music training. Participants in the United States were recruited through the UT Dallas Psych Research Sign-up System (SONA) and by word of mouth and social media. French participants were recruited by word of mouth, email, and social media. Only participants who signed up via the SONA portal were compensated with research participation credit, other participants were not compensated. Of a total of 520 survey responses, 160 were removed for not being incomplete, leaving a total of 360. Participants from the US who indicated a nationality other than American were excluded from analysis. "Ghanian," for example, was not included, but responses such as "Asian-American" were. This left a total of 279 ($N_F = 108$, $N_A = 171$) survey responses

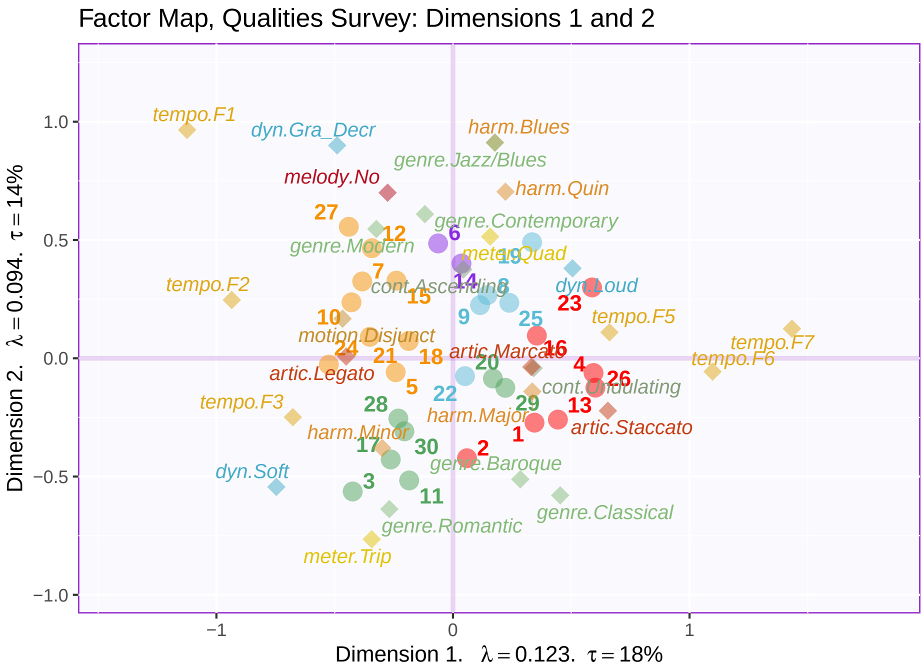


Figure 5

for analysis. All recruitment measures were approved by the UT Dallas IRB.

Stimuli. The stimuli used for Experiment 2 are the same as those used for Experiment 1.

Surveys. Participants in Experiment 2 completed a survey (hereafter: Adjectives Survey/AS) via qualtrics in either English or French, depending on their location, that consisted of a CATA task in which participants evaluated the stimuli using thirty-three adjectives such as 'dark,' 'warm,' and 'colorful' (French: 'sombre,' 'chaleureux,' and 'colore'). The adjectives for the AS were selected using Wallmark (2019) as a guide and in consult with a French professional musician. Some adjectives were initially selected in French and some in English. In all cases, adjectives were selected for which there was a clear French (vis-à-vis English) translation. The adjectives are listed in English and in

French in the supplementary materials. The survey also gathered demographic data, including age, gender identity, nationality, occupation, and musical experience.

Procedure. Participants were provided with a link to the survey and were instructed to complete it in a quiet listening environment or using headphones, but that was not controlled, nor was it assessed as part of the survey. After standard informed consent procedures, participants performed a CATA task on 15 excerpts presented one at time in a random order. All adjectives were presented, in a random order, for each excerpt. Demographic survey questions followed the experimental task.

Results

Participants. The scree plot depicted in Figure 6 shows the explained variance per dimension for the distance analysis of participants on the AS. The first five dimensions all have $\lambda > 1$: 1.66, 1.27, 1.13, 1.09, and 1.06, respectively, but because we had many participants and thus a high dimensionality in this analysis, the first dimension extracts only ~3% of the overall variance. Again, as above, for the purposes of this case study, we're focusing on the first two dimensions.

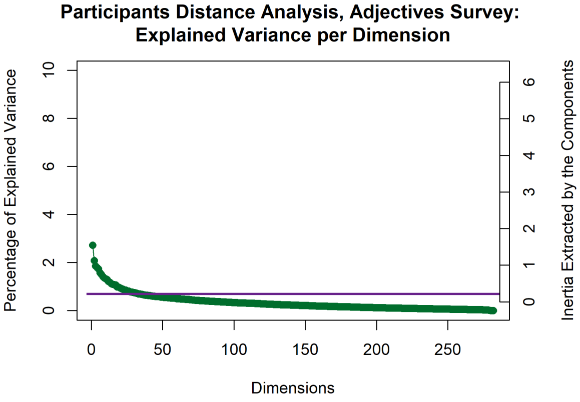
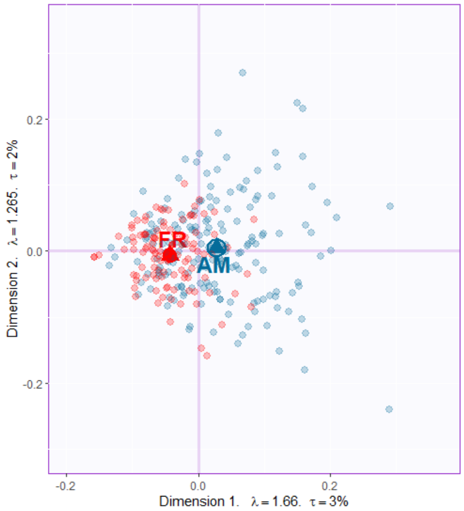


Figure 6

An MDS analysis of a distance matrix calculated from the pages of the brick revealed significant differences in how French and American participants described the excerpts, *p.* < .01. The factor scores of the participants are plotted in Figure 7, with with group means and bootstrapped confidence intervals shown for those means. We also analyzed the data using two other participant groupings as factors, gender identity and level of music training, neither of which were significant.

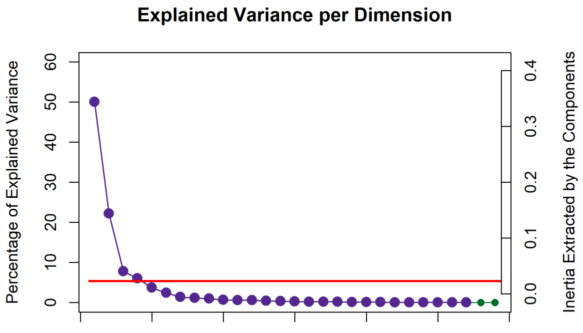
Figure 7. R_V Analysis of Participants in the Adjectives Survey



Note. Group means are indicated with triangles and labled with AM and FR. The ellipse around the group mean indicates the confidence interval, after bootstrapping 1000 iterations. The fact that there is a clear separation between the group means and the confidence intervals suggests that there is a significant difference between the groups, $p > .001$.

Excerpts. The plot in Figure 8 shows the explained variance per dimension in the analysis of the AS contingency table. Although there are no components with $\lambda > 1$, there are two strong dimensions that extract a majority of the variance. The first two dimensions extract 72.25% of the variance, with the first dimension extracting a majority: 50.05%, and the second dimension extracting almost a quarter of the overall variance: 50.05%. Although excerpts 6 and 14 are outliers in the musical qualities survey, for reasons detailed above, they were not outliers in this analysis. We therefore included them in all of the analyses for Experiment 2.

The contributions to the first two dimensions are depicted in Figure 9, colored according to the clusters extracted using HCA. Contributing



significantly — more than the average — to the first dimension are Excerpts 3, 5, 7, 10, 18, 24, and 27, in the positive direction and Excerpts 4, 13, 19, 23, 26, and 29 in the negative direction. Adjectives that contribute positively to the first dimension are "Sad," "Dark," "Melancholy," "Slow," "Mysterious," "Solemn," and "Disturbing," and those contributing negatively are "Fast," "Happy," "Dancing," "Colorful," and "Bright." On the second dimension, excerpts that contribute in the positive direction are 1, 6, 7, 16, and 25, and those contributing in the negative direction are Excerpts 10, 11, 20, 28, and 29. The columns contributing strongly in the positive direction are "Aggressive," "Fast," "Disturbing," "Mysterious," "Surprising" and "Complex," and those contributing negatively are "Warm,"Soft","Happy","Slow","Round", and"Light".

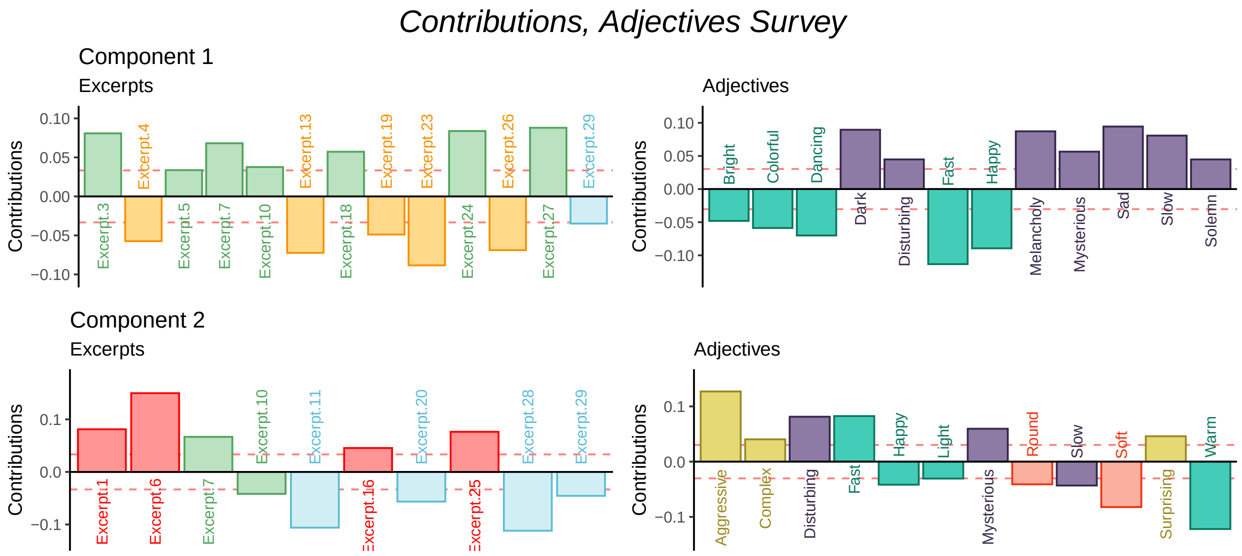


Figure 9

The barplots in Figure 10 show the bootstrap ratios calculated for the rows and columns. All rows and columns are included to compare what clusters are consistently distributed in the space. This barplot was not included for Experiment 1 because it is less informative given that Experiment 1 evaluates the nature of the excerpts not the behavior of participants. There are more significant bootstrap ratios than there are significant contributions, which means that the model seems to be stable, and the few excerpts and adjectives that are not significant are likely significant on higher dimensions. All excerpts are stable on one dimension or the other, likewise all adjectives except 'sparse.'

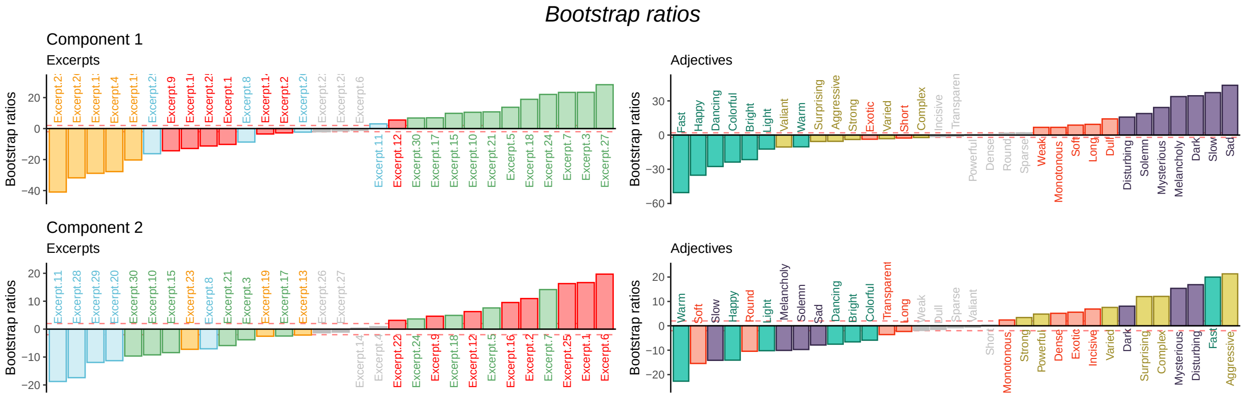


Figure 10

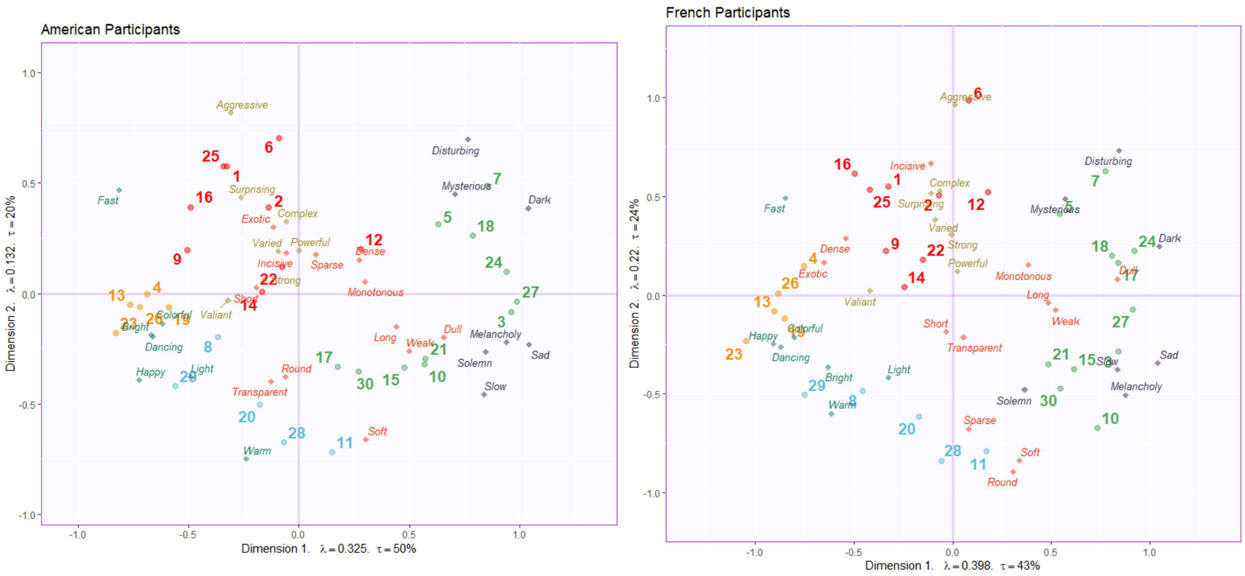
Discussion

The factor maps below show the row and column factor scores for the American and French participants separately to illustrate the differences observed between the groups of participants. These are once again symmetric biplots, interpretation is the same as the factor plot for the musical qualities. There's a clear valence-arousal plane apparent for both, with valence on the first dimension and arousal on the second dimension. However, the French data show a weaker first dimension but a stronger second dimension relative to the Americans, both in terms of variance extracted (τ), and effect size (λ). This suggests that French participants responded with less affect but greater arousal than the American participants. There are also differences in how the adjectives and the excerpts are distributed in the space. One clear example is Excerpt 6, which is in the low valence/high arousal quadrant in the American plot, and the high valence/high arousal quadrant in the French plot, suggesting that the two groups tended to assign different valence to this excerpt. For the adjectives, 'bright' and 'dancing' are directly on top of one another in the American plot, but there is some space between the two in the French plot. This reflects shared meaning but differences in semantics or associations between languages.

Although this experiment was designed to evaluate the cognitive response to music, and not the emotional response, there is significant overlap in the results observed here and the results of the work investigating music and emotion, specifically in the appearance of the valence-arousal plane. Studies on music and emotion have used this construct as a way of measuring emotion, but the original proposal was that this plane is simply a measure of "meaning" (Osgood & Suci, 1955). The adjectives selected for use in Experiment 2 reflect this original proposal, as they were selected specifically to represent a cognitive space, not an emotional one.

The results of the MFA are displayed in Figure 12. In these plots, the differences in behavior between the groups is more clear. The triangles represent the compromise between the mental spaces of the French and American participants, and the lines extending from them indicate the scores of the groups separately. Excerpts and adjectives that were rated similarly have shorter lines extending from them, but excerpts that were rated differently have much longer lines. Examples of excerpts that were rated differently are numbers 6, 8, 12, and 17. Adjectives that were used differently include "Disturbing," "Round," "Solemn," and "Bright." The valence-arousal plane revealed by the CA is also present here, and provides a framework for interpreting the differences between groups. Excerpt 17 is perhaps the most extreme example. American participants rated this excerpt with much lower arousal and slightly less negative valence than the French participants, so much so that for the American participants, the excerpt landed in the low arousal/negative valence quadrant, and for the French participants it landed in the high arousal/negative valence quadrant. Another interesting case is for Excerpt 8, which lands in the same quadrant for both groups, but much further from the origin for the French participants than the Americans.

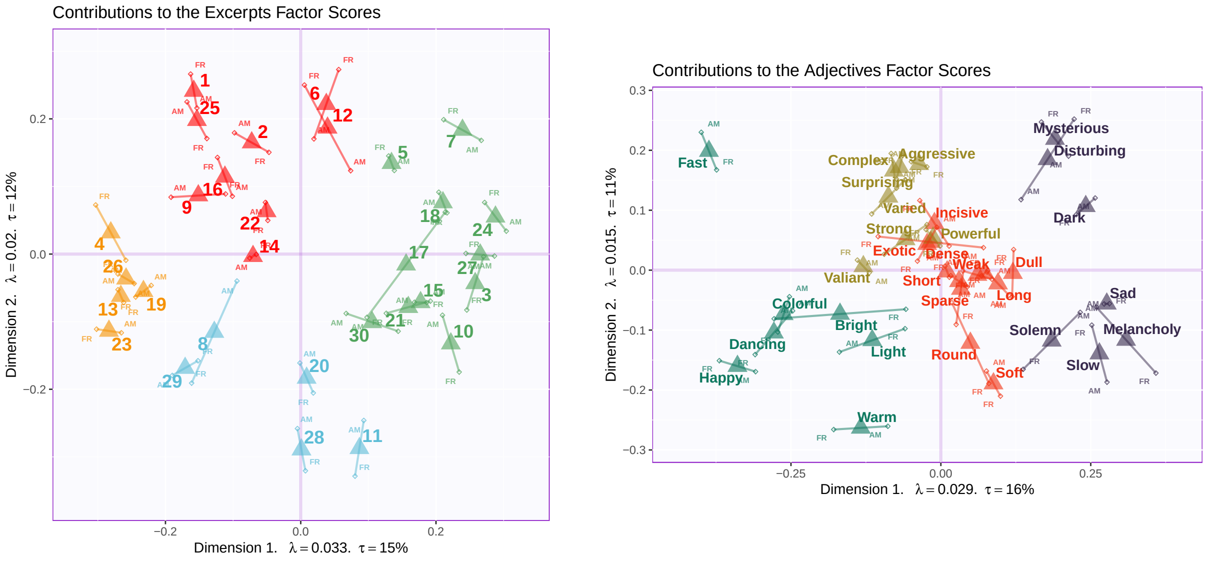
Figure 11. Symmetric Plots for Rows and Columns of the Adjectives Surveys, by Participant Nationality



Note. For these plots, the survey responses were split by nationality and analyzed separately. Note the differences in variance extracted by each of the first two dimensions.

Some examples of differences in the use of adjectives includes "disturbing" (inquiétant) seems to be more extreme for the French participants than the Americans. In English, "Solemn" (solennel) carries more valence, and in French it carries more arousal; similarly, "bright" (brillant) seems to carry much more positive valence in English than in French. In English, "melancholy" (melancolique) and "sad" (triste) were used the same, but in French they were used very differently.

Figure 12. Partial Factor Scores Plots from the MFA



Contributions to the Excerpts Factor Scores

Contributions to the Adjectives Factor Scores

Note. In each plot, the triangles represent the combined factor scores and the small circles represent the partial factor scores contributed by each of the groups.

Experiment 3: Combined Surveys

Experiment 3 used the pseudo-contingency tables from both Experiments 1 and 2. Since excerpts 6 and 14 were excluded from analysis for Experiment 1, those rows were also removed from the contingency table for Experiment 2. This is so that the dimensions of the two tables for this PLSC would be conformable. The goal of this experiment was to identify the strongest shared signal between the two tables. Remember that PLSC shows what is common between two different sets of data — how often an excerpt was associated with *both* a musical quality and an adjective. The visualizations below show which adjectives are associated with which musical dimensions. Even though both individual tables have their own factor spaces, plotting the common factor space between the two should allow us to see which excerpts are separated from one another using data from both surveys.

Results.

This analysis revealed two dimensions that extracted the majority of the variance (83.60%), with the first dimension extracted 64.35% and the second 19.26%. The scree plot in Figure @fig:screePLSC indicates there are two elbows in this graph, at the 3rd and 5th dimensions, suggesting that further analysis is possible.

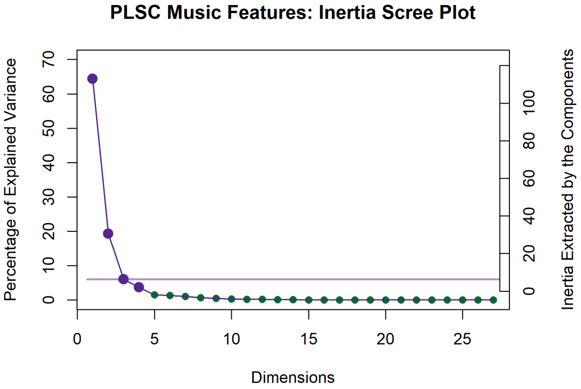


Figure 14

The plot below shows which variables from each data table load the most on the first and second dimensions. For the purposes of this visualization, only the variables for which 70% or more of the variance is explained are shown. The nature of the PLSC also suggests that these are the variables that are most associated with one another between the two tables. The strongest signal on the first dimension juxtaposes the slow and legato musical qualities in the positive direction with the fast, staccato, marcato, and conjunct musical qualities in the negative direction. The adjectives associated with the qualities in the positive direction are “Dark,” “Dull,” “Long,” “Melancholy,” “Sad,” “Slow,” “Solemn,” and “Weak.” The adjectives associated with the negative direction are “Bright,” “Colorful,” “Dancing,” “Fast,” “Happy,” and “Light.”

The second dimension identified in the positive direction major harmony and medium dynamics, associated with “Light,” “Round,” “Soft,” and “Warm.” The negative direction is driven by the impressionist genre being associated with “Aggressive,” “Complex,” “Dense,” “Disturbing,” “Powerful,” and “Surprising.”

Figures 16 and 15 show us that there are more variables that contribute significantly to these dimensions than for which a significant portion of the variance is explained. There are similar groups, however: on the first dimension, the tempo variables are contributing

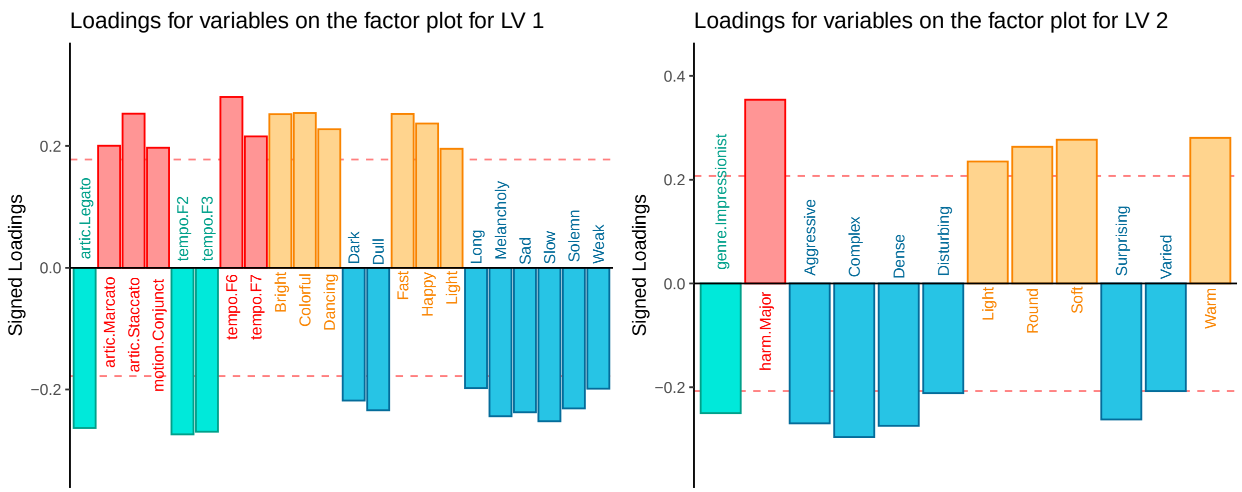


Figure 15

significantly, along with some from harmony, density, genre, dynamics, motion, range, and articulation. The adjectives contributing significantly are Bright, colorful, Dancing, Fast, Happy, Light, and Valiant in the positive direction, and Dark, Dull, Long, Melancholy, Monotonous, Sad, Slow, Solemn, and Weak in the negative direction. This juxtaposes some negatively and positively valenced adjectives, and identifies which musical qualities contribute to the valence dimension. Even though some of these variables did not contribute significantly in their plots above (see Figures 11 and 5), their appearance here indicates that they are part of the shared signal between the tables. One-third of the musical qualities contributing to the second dimension are harmony and genre. Also contributing are the dynamics and contour groups, while contour, articulation, motion, and range show only one or two variables. The adjectives contributing negatively are Aggressive, Complex, Dense, Disturbing, Incisive, Mysterious, Powerful, Surprising, and Varied, and those contributing positively are Light, Round, Soft, Transparent, and Warm.

Discussion. The factor score plots for this analysis shows that the first two sets of latent variables extracted by the analysis effectively separate the groups of excerpts into the clusters defined in the HCA for the adjectives survey. The strongest correlated signal between the two data tables separates Excerpts groups 2 and 3 and the second latent

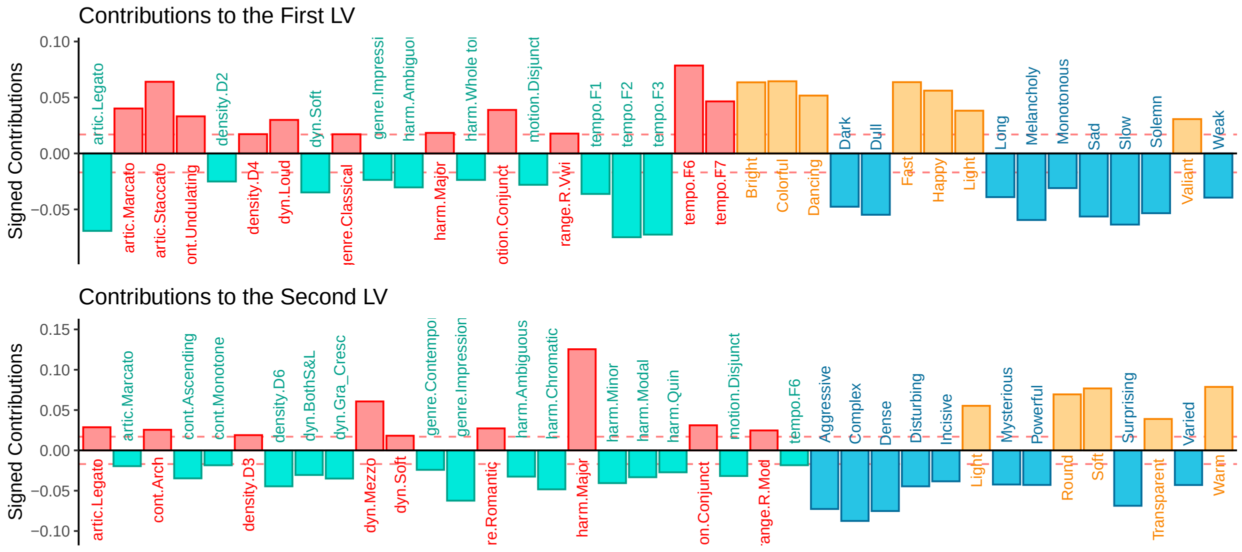


Figure 16

variable separates groups 1 and 4. Although there are no factor plots for the variables in this analysis, the valence-arousal plane created by the first two dimensions is still apparent. This suggests that the excerpts that are more distant from the origin in the first LV plot are defined more by valence than arousal, and those clustered further from the origin on the second LV plot are defined more by arousal than valence. For example, Excerpt 26 is one of the most extreme examples of positive valence, but is much closer to the origin in the second LV plot similarly with Excerpt 27, but with negative valence. This is contrasted with Excerpt 7, which is one of the most negatively valenced stimuli, but also scores very high on arousal, although the confidence interval for that group circles the origin of that plot.

General Discussion

We used multivariate analyses to explore the musical and cognitive spaces created by participants from France and the United States when responding to a set of new string quartets. These results revealed a clear valence-arousal plane that was common to participants from both countries, but significant differences in the behavior of French and

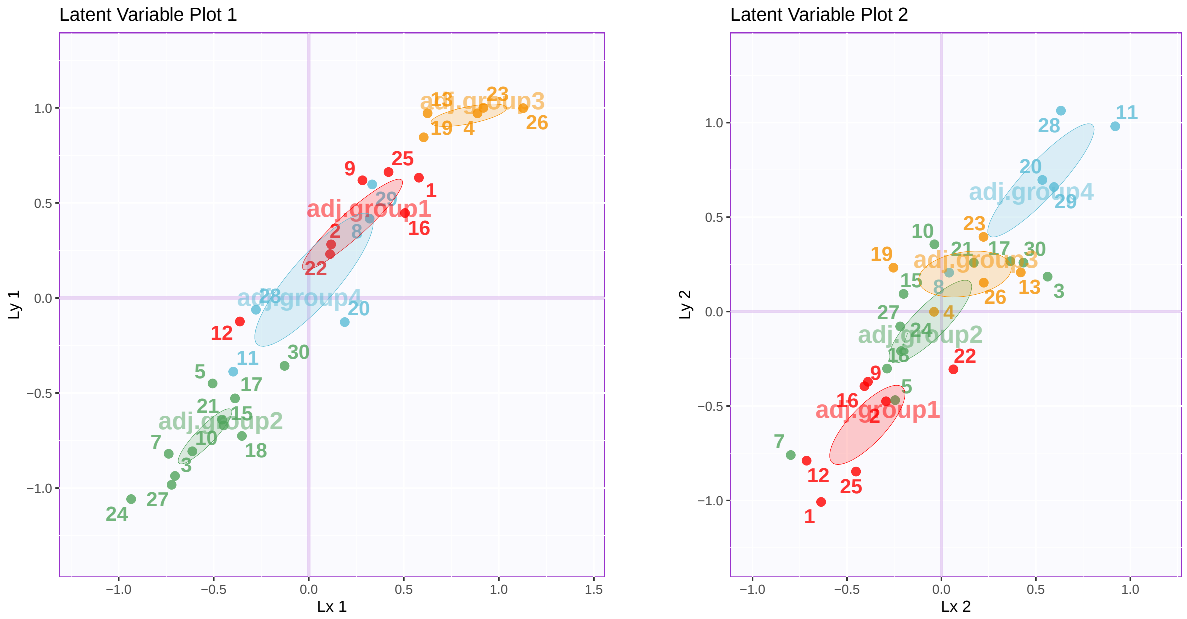


Figure 17

American participants when responding to the stimuli using adjectives, and a space largely defined by genre when evaluating the stimuli using musical qualities. The combined musical and cognitive spaces identified which musical qualities were associated with which descriptors in terms of valence and arousal. One effect observed in Experiment 1 which was not observed in Experiment 2, in which two individual excerpts — numbers 6 and 14 — dominate the factor space, requires more explanation. This effect is due to the nature of CA, which is to find the average observation. In a CA, information that is common falls towards the origin — the center of mass — of the factor plot, while information that is further from the average, or more rare, ends up further away (Abdi & Williams, 2010). Therefore, if any individual item on a survey is rated significantly different from the rest of the items, that item can dominate the factor space. In this case we have two such examples: Excerpt 6 was written as a Steve-Reich-esque minimalist, ostinato based excerpt, and excerpt 14 was written to be jazzy. This effect does not appear in Experiment 2 because the AS was designed to evaluate the excerpts more generally on holistic qualities, not to separate the excerpts along specific musical dimensions. Excerpt 6 still appears near the edge of the mental space in Experiment 2, indicating that it is somewhat of an outlier, but does not dominate the space the way it does in the results Experiment 1. In order to interpret the space without the effects of Excerpts 6 and 14, those two excerpts were removed from the dataset for the initial analysis and then included as *supplementary projections*, sometimes also referred to as *out of sample observations*. This allowed us to evaluate what information is shared by the outliers and the other elements in the dataset without having the outliers dominate the visualization of the factor space. The fact that this was a necessary step supports our interpretation that the factor space is defined by genre. The supplementary projections' location in the factor space shows that there is some shared information between genres. If the supplementary observations had projected onto the origin or very close to it, that would indicate that they shared no information with the other variables. One takeaway from these results is that a deep understanding of the stimuli may help to predict the approximate dimensionality of the solution factor space, and when designing surveys or stimuli, multiple items per group, or presumed dimension, are needed. The outliers distorting the factor space that were observed in the present study are analogous to the single noisy dimension described in the introduction. The noise contributed by Excerpt 6 is also present in the results of Experiment 2, possibly because untrained participants are less likely to be familiar with minimalism than the trained participants in Experiment 1, but the results are robust to that noise because the participants were not asked to rate the excerpts on any explicit dimensions or qualities.

The significant results for the participants in Experiment 2 that were not observed in Experiment 1 suggest that this experimental paradigm works as intended. Significant differences between the experts' ratings of the stimuli would have indicated that the experts were inconsistent and thus not reliable raters of the music. The significant differences between the French and American adjectives surveys indicate that language connotations do have a significant impact on how participants rate the stimuli. MFA revealed what those differences were and highlighted further possibilities for analysis.

Limitations & future directions

Although we evaluate the scores and ratings of participants from different countries, we recognize that the issue of multiculturality is not addressed to a significant degree in this study, because France and the United States are both western countries that share Western musical culture. To truly address this question, an experiment would need to include participants from multiple, contrasting musical cultures, with languages that are more distant than English and French. However, specific musical qualities, like harmony, may not apply or translate well to other musical cultures, because the concepts of melodic and harmonic material are not the same across all musical cultures (Cohn et al., 2001; Raman & Dowling, 2017). Therefore the specific questions included in the QS would need to be adjusted.

Although we suggest that data collected in this way have a much greater hypothetical reach, we recognize that the data collected for these experiments represent a convenience sample, and many of the participants were students. However, these limitations could be easily remedied in future studies. Another question that fell beyond the scope of this study what the source of the semantic drift between languages is. Although illustrated in Figure 12, the source of the differences between French and American participants is not entirely clear. These differences may not come from cultural aspects of music listening, but linguistic sources, including the adjectives' frequency of use in either language or the cultural associations with the words (B. Thompson et al., 2020), or even the physical characteristics of the words themselves (Reilly et al., 2012). Diving more into those questions would be a fascinating future study. Another interesting study would be to use adjectives from specific domains, to see how music maps onto different sensory spaces, for example textural words, like 'moist,' 'slimy,' 'dry,' 'puckered,' 'smooth.'

Finally, the results of this study and possible extensions, in conjunction with studies that have already evaluated music perception non-verbally, may provide insight into the way in which people without language react to music, such as nonverbal autistic people.

Conclusions

Expanding the collection and analytical paradigms, and thus expanding scientific scope and perspective, has the added benefit of increasing reach. By expanding the ways in which we collect data, including developing investigative paradigms that are accessible on mobile platforms and that reduce participant demand while maintaining rigor and integrity, we are able to more readily and consistently reach a broader participant population, that might normally be excluded from everday research paradigms, specifically racially and ethnically diverse populations, poorer populations, those with limited access to transportation, or who have a disability, or are immunocompromised. Pairing this kind of data gathering with appropriate analysis will help maintain scientific integrity. The number of ways that exist to analyze data from a single set of experiments is considerable, and the results of each analysis illuminate different parts of the story behind the data. Not every form of analysis is appropriate in every context, but understanding how, and perhaps more importantly when, to apply an analytical technique is vital to uncovering new perspectives or insights.

References

Abdi, H., & Williams, L. J. (2010). Correspondence Analysis. In N. Salkind (Ed.), *Encyclopedia of research design*. Sage.

Abdi, H., & Williams, L. J. (2013). Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression. In B. Reisfeld & A. N. Mayeno (Eds.), *Methods in molecular biology: Computational toxicology volume II* (Vol. 930, pp. 549–579). Springer Science+Business Media, LLC. https://doi.org/10.1007/978-1-62703-059-5

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*, 149–179. https://doi.org/10.1002/wics.1246

Ares, G., Deliza, R., Barreiro, C., Giménez, A., & Gámbaro, A. (2010). Comparison of two sensory profiling techniques based on consumer perception. *Food Quality and Preference*, *21*(4), 417–426. https://doi.org/10.1016/j.foodqual.2009.10.006

Balkwill, L. L., Thompson, W. F., & Matsunaga, R. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. *Japanese Psychological Research*, *46*(4), 337–349. https://doi.org/10.1111/j.1468-5584.2004.00265.x

Balkwill, L., & Thompson, W. F. (1999). A Cross-Cultural Investigation of the Perception of Emotion in Music : Psychophysical and Cultural Cues. *Music Perception: An Interdisciplinary Journal*, *17*(1), 43–64. https://doi.org/10.2307/40285811

Bartlett, J. C., & Dowling, W. J. (1980). Recognition of transposed melodies: A key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(3), 501–515. https://doi.org/10.1037/0096-1523.6.3.501

Battcock, A., & Schutz, M. (2019). Acoustically expressing affect. *Music Perception*, *37*(1), 66–91. https://doi.org/10.1525/MP.2019.37.1.66

Benzécri, J.-P. (1973). *L'analyse des données.* Dunod.

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 527–542. https://doi.org/10.1002/wics.177

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

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion*, *19*(8), 1113–1139. https://doi.org/10.1080/02699930500204250

Borg, I., & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling* (2nd ed., Vol. 36). Springer Science+Business Media, Inc.

Bruner II, G. C. (1990). Music, Mood, and Marketing. *Journal of Marketing*, *October*, 94–104.

Cohn, R., Hyer, B., Dahlhaus, C., Anderson, J., & Wilson, C. (2001). *Harmony.* Oxford University Press.

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement*, *16*(1), 13–37. https://doi.org/10.1177/001316445601600102

Cowen, A. S., Fang, X., Sauter, D., & Keltner, D. (2020). What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(4), 1924–1934. https://doi.org/10.1073/pnas.1910704117

Darrow, A. A., Haack, P., & Kuribayashi, F. (1987). Descriptors and Preferences for Eastern and Western Musics by Japanese and American Nonmusic Majors. *Journal of Research in Music Education*, *35*(4), 237–248. https://doi.org/10.2307/3345076

Dowling, W. J. (1978). Scale and Contour: Two Components of a Theory of Memory for Melodies. *Psychological Review*, *85*(4), 341–354. https://doi.org/10.1037/0033-295X.85.4.341

Escofier-Cordier, B. (1965). *L'analyse des correspondances* [Doctoral Thesis]. Université de Rennes.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A. D., & Koelsch, S. (2009). Universal Recognition of Three Basic Emotions in Music. *Current Biology*, *19*(7), 573–576. https://doi.org/10.1016/j.cub.2009.02.058

Gray, P. H., & Wheeler, G. E. (1967). The Semantic Differential as an Instrument to Examine the Recent Folksong Movement. *Journal of Social Psychology*, *72*(2), 241–247. https://doi.org/10.1080/00224545.1967.9922321

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press.

Gregory, A. H., & Varney, N. (1996). Cross-cultural comparisons in the affective response to music. *Psychology of Music*, *24*(1), 47–52. https://doi.org/10.1177/0305735696241005

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 497–526. https://doi.org/10.1002/wics.182

Juslin, P. N., & Sloboda, J. A. (Eds.). (2010). *Handbook of music and emotion: Theory, research, applications.* Oxford University Press.

Juslin, P. N., & Västfjäll, D. (2008). All emotions are not created equal: Reaching beyond the traditional disputes. *Behavioral and Brain Sciences*, *31*, 559–621. https://doi.org/doi:10.1017/S0140525X08005554%20Patrik

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *Journal of Abnormal and Social Psychology*, *28*(3), 280–290. https://doi.org/10.1037/h0074049

Kennedy, J., Kennedy, M., & Rutherford-Johnson, T. (2013). *Harmony* (6th ed.). Oxford University Press.

Kopacz, M. (2005). Personality and music preferences: The influence of personality traits on preferences regarding musical elements. *Journal of Music Therapy*, *42*(3), 216–239. https://doi.org/10.1093/jmt/42.3.216

Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, *56*(2), 455–475. https://doi.org/10.1016/j.neuroimage.2010.07.034

Ladinig, O., & Glenn Schellenberg, E. (2012). Liking unfamiliar music: Effects of felt emotion and individual differences. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(2), 146–154. https://doi.org/10.1037/a0024671

Madsen, C. K. (1997). Emotional Response to Music as Measured by the Two-Dimensional CRDI. *Journal of Music Therapy*, *34*(3), 187–199. https://doi.org/10.1093/jmt/34.3.187

Meyners, M., & Castura, J. (2014). Check-All-That-Apply Questions. In *Novel techniques in sensory characterization and consumer profiling* (pp. 271–306). CRC Press/Taylor & Francis. https://doi.org/10.1201/b16853-12

Osgood, C. E., & Suci, G. J. (1955). Factor analysis of meaning. *Journal of Experimental Psychology*, *50*(5), 325–338. https://doi.org/10.1037/h0043965

Raman, R., & Dowling, W. J. (2016). Real-Time Probing of Modulations in South Indian Classical (Carnātic) Music by Indian and Western Musicians. *Music Perception*, *33*(3), 367–393. https://doi.org/10.1525/MP.2016.33.03.367

Raman, R., & Dowling, W. J. (2017). Perception of modulations in south indian classical (carnatic) music by student and teacher musicians: A cross-cultural study. *Music Perception*, *34*(4), 424–437.

Reilly, J., Westbury, C., Kean, J., & Peelle, J. E. (2012). Arbitrary symbolism in natural language revisited: When word forms carry meaning. *PLoS ONE*, *7*(8). https://doi.org/10.1371/journal.pone.0042286

Rodà, A., Canazza, S., & De Poli, G. (2014). Clustering affective qualities of classical music: Beyond the valence-arousal plane. *IEEE Transactions on Affective Computing*, *5*(4), 364–376. https://doi.org/10.1109/TAFFC.2014.2343222

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*(10), 1029–1038. https://doi.org/10.1038/s41562-020-0924-8

Thompson, W. F. (1994). Sensitivity to combinations of musical parameters: Pitch with duration, and pitch pattern with durational pattern. *Perception & Psychophysics*, *56*(3), 363–374. https://doi.org/10.3758/BF03209770

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, *23*(2), 111–136. https://doi.org/10.1007/BF02289009

Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: a review of new descriptive methods in food science. *International Journal of Food Science & Technology*, 1–16. https://doi.org/10.1111/j.1365-2621.2012.03022.x

Wallmark, Z. (2019). A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, *47*(4), 585–605. https://doi.org/10.1177/0305735618768102

Wedin, L. (1969). Dimension Analysis of Emotional Expression in Music. *Swedish Journal of Musicology*, *51*, 119–140.

Wedin, L. (1972). Evaluation of a Three-Dimensional Model of Emotional Expression in Music. *The Psychological Laboratories*, *54*(349), 1–17.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An Interlanguage Study of Musical Timbre Semantic Dimensions and Their Acoustic Correlates. *Music Perception: An Interdisciplinary Journal*, *31*(4), 339–358. https://doi.org/10.1525/MP.2014.31.4.339

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2015). An Interlanguage Unification of Musical Timbre: Bridging Semantic, Perceptual, and Acoustic Dimensions. *Music Perception: An Interdisciplinary Journal*, *32*(4), 394–412. https://doi.org/10.1525/MP.2015.32.4.394