Cognitive Music Listening Space: A Multivariate Approach

Brendon Mizener1, Mathilde Vandenberghe2, Hervé Abdi1, & Sylvie Chollet2

1 University of Texas at Dallas

2 YNCREA

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](mailto: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, Emotion, 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. 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? Picture this: You’re in a bar, and that new song that you like so much comes on. You take a sip of your drink, and it’s perfect. What drink did you just drink? What if a different song had come on? Would it have been as perfect? The present study was designed to quantify a music listening ‘space’ that captures perceptual 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? And: What parallels exist between qualitative and quantitative descriptions of music?

However, that doesn’t mean that online data collection is necessarily a worse option. Studies performed using remote data collection are not confined to the physical space of a lab, or the physical borders of a country. This means that sample population can be more diverse, even multinational. It’s also easier to recruit greater numbers of participants, and an increase in sample size helps to mitigate the issues caused by ‘dirtier’ data.

We can also reevaluate what kinds of questions we ask and how we ask them. Likert-type scales are commonly used in in-person studies, and translate well to online studies. In the right context - and with the right analysis - nominal responses may be just as useful, but those two aren’t the only options. The check-all-that-apply (CATA) technique (Coombs et al., 1956; Katz & Braly, 1933; Meyners & Castura, 2014), which we describe in more detail below, is an example of a data collection technique that features a number of benefits in this regard.

Another way to mitigate the effects of dirty data is to use multivariate analyses. There are a few reasons for this. Firstly, because we can’t enforce the same level of control over online studies, we should assume some of the assumptions associated with traditional univariate analyses, hypothesis testing, and inferences are violated. Secondly, expanding the scope of the analysis and the number of variables included helps to increase the signal-to-noise ratio. Whereas with univariate analyses, noise in the one target variable easily obscures the signal, in a solution to a system with multiple variables or dimensions, noise in one or two of those dimensions is less intrusive. This is because the multivariate solutions evaluate the total variance in all of the dimensions, instead of the variance for each individual dimension separately.

## 0.1 Music Perception

Quantifying music perception is a problem that presents an interesting test case for this kind of data gathering and analytical paradigm. Most music perception studies follow the standard practices in psychology described above, with regard to control and number of variables. Studies involving music or, indeed, any auditory perception, have the added difficulty that small changes can affect listeners’ perception, especially when the study involves timing, specific tuning, or sound localization. However, the experimental controls may be loosened slightly when investigating holistic music listening, because no individual facet is more important than the music as a whole.

In a holistic listening paradigm, listeners continuously evaluate incoming information on internal technical and affective schemata that are informed by their individual musical experiences and personality traits (Kopacz, 2005). While the the technical and affective aspects of music are theoretically distinct, in practice there is a great deal of interplay between the two. Listeners respond affectively to technical aspects of music, and composers use various musical and compositional techniques things to reflect the internal emotional states they want to express (**Battock2019?**). Individual technical aspects of music have been demonstrated to have specific effects on listeners’ affective perception (Bruner II, 1990; **Battock2019?**), but quantifying the interactions between more than one or two technical aspects is challenge. That challenge is made more difficult by individual associations and experiences that participants bring to the table (Kopacz, 2005).

One of the reasons these interactions have been difficult to pin down is that models like ANOVA which use only a few variables are limited by how many variables a researcher can include while remaining coherent. Thus, the many studies that follow standard experimental paradigms do not reflect the complexity inherent to music and music listening.

### 0.1.1 Music and Emotion.

The present study was not intended to investigate or quantify emotional responses to music, but it has many similarities with music emotion research. This is a well trod domain - see, for example Juslin and Sloboda (2010) - and we simply hope to bring a new perspective. Multivariate analyses have been used in this domain since at least Wedin (1969). These early studies (Wedin, 1969, 1972; **Gray1967?**) supported Osgood & Suci’s (1955) theory that valence and arousal are the two most salient dimensions in evaluating almost any stimulus, including music. Recent studies have asked participants to rate music only on these two dimensions, using sliders (Bigand et al., 2005; Madsen, 1997). This is useful, as it provides fine-grained detail on the level of arousal or valence a stimulus presents. And, depending on the specific application, it can capture real-time responses on these dimensions. However, evaluating only two dimensions is limiting. In addition to restricting the solution space to only two dimensions, this method provides no way to qualify that information. Participants may report similar levels of valence and arousal for different stimuli, but with nothing else to go on, it may be unclear as to why.

There have also been a few studies that have specifically investigated dimensions beyond those first two (for example Rodà et al. (2014)). There is also recent conflicting evidence as to whether the valence-arousal plane represents the fundamental dimensionality behind music emotion perception (Cowen et al., 2020). This alternative hypothesis suggests there are “at least 13 dimensions” (Cowen et al., 2020) to the subjective music and emotion perceptual space. This is especially relevant as it evaluates the dimensions that are common across musical cultures.

An important distinction between work in emotional perception in music and the current study is that the adjectives we chose were informed by music composition and performance (e.g., Wallmark (2019)) rather than emotion. We specifically intended to get at the semantics of music description. The present study, therefore, is in some ways more aligned with studies that evaluate timbre (for a full treatment see Wallmark and Kendall (2018)).

### 0.1.2 Intercultural Studies.

There are few common goals in intercultural studies. Some aim to quantify the shared emotional experience between musical cultures (Cowen et al., 2020; **Fritz2009?**; **Balkwill2016?**), and some ask participants to identify technical aspects of the intercultural music (Raman & Dowling, 2017; **Raman2016?**). There are far fewer studies that include semantics in their evaluation of music perception (Zacharakis et al., 2014; **Zacharakis2015?**), which makes this a prime area for research.

The research program presented in Zacharakis et al. (2014) and (**Zacharakis2015?**) deal specifically with timbre perception (as opposed to holistic music perception, as in the present study). However, their use of adjectives is similar. Zacharakis et al. (2014) and (**Zacharakis2015?**) asked Greek and English participants to describe timbre in their own languages, and found that, generally speaking, participants’ descriptions of timbre doesn’t differ much between languages.

## 0.2 Present questions & methods of analysis

The basic question was simple: how do French and American participants describe music? Our investigative paradigm, along with sample and size, are addressed in the methods section below, but we felt it may be useful to provide a quick overview of the data collection and analytical techniques for readers who may be unfamiliar.

### 0.2.1 Check-all-that-apply (CATA).

The CATA technique (Coombs et al., 1956; Katz & Braly, 1933; Meyners & Castura, 2014) is used to measure how participants evaluate a set of stimuli presented one at a time. For each stimulus, participants are shown a list of descriptors and are asked to select the descriptors that apply to the presented stimulus (Meyners & Castura, 2014). Katz and Braly (1933) provides an early example of the use of the CATA paradigm in the psychological sciences, but it is not terribly common in that domain anymore. However, it has been and continues to be used widely in sensory evaluation (Abdi & Williams, 2010a; Meyners & Castura, 2014). We alluded to a number of benefits to this technique above. The two most important benefits are that it easily assesses questions to which there is more than one ‘correct’ response (Coombs et al., 1956), and that it lowers the cognitive demand placed on participants because they do not have to generate their own responses (Ares et al., 2010). Both of these benefits make it very attractive to the psychological sciences.  
As with any data collection technique, it’s important to pair CATA with an appropriate mode of analysis. Because the data in a CATA survey can easily be transformed into count data, using the responses as variables, it pairs very well with Correspondence Analysis and Multidimensional Scaling. Transforming responses into counts has the added benefit of making this method robust to a wide variety of responses and violations.

### 0.2.2 Correspondence Analysis.

The primary analysis used on the data collected in the surveys is Correspondence Analysis (CA). CA has many names, and has been ‘discovered’ many times by many people. There are a number of excellent references that illustrate the calculative (Greenacre, 1984) and graphical or geometrical (Benzécri, 1973) aspects of CA. This technique is similar to Principal Components Analysis (PCA), except that it allows for the analysis of qualitative data. Data tables for CA are organized such that the value in a given cell represents the relationship between the observation and the variable symmetrically, that is, it is both the number of times a variable was associated with an observation, and the number of times an observation was associated with a variable, depending on perspective. Because of this, the variance of the table as a whole represents the variance associated with either the rows or the columns, depending on how it is analyzed. Therefore the axes of the factor plots of CA represent the variance extracted from both the the observations and the variables, so we can plot the factor scores for both rows and columns in a single space.

### 0.2.3 Partial Least Squares Correlation.

Partial Least Squares Correlation (PLSC) (Abdi & Williams, 2013; **Tucker1958?**) analyzes two data tables that have the same information either on the observations (rows) or variables (columns). The PLSC extracts the covariance between two tables in the form of *latent variables*. 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). Additionally, we can which variables are responsible for creating or defining the primary axes of similarity between the two data sets. There are some criticisms of this technique that argue that it is overpowered, that it can ‘find’ spurious correlations, and to that end we would simply suggest caution when interpreting PLSC results. We used this technique to extract the covariance between two different surveys about the same set of stimuli.

### 0.2.4 Multidimensional Scaling.

Multidimensional Scaling (MDS) (Borg & Groenen, 2005) analyzes a square, symmetrical distance matrix in which each cell represents the distance, or the amount of difference, between the item on the row and on the column. Plotting the resultant factor scores shows the relative distance between all of the points. MDS has a history of being used in studies evaluating musical excerpts, at least since Wedin (1969) and Wedin (1972). It is commonly used to represent the distance between stimuli in a perceptual space. This is a useful application, but doesn’t use MDS to its full potential. It might be more effectively used to plot the participants and use bootstrapping to calculate differences between groups of participants, which is how it was applied in the present study.

### 0.2.5 Multiple Factor Analysis.

Multiple Factor Analysis (MFA) is the only unplanned analysis used in this study, and is also the newest (Abdi et al., 2013). We chose to run this analysis after finding significant mean differences between French and American participants using MDS. MFA is uniquely suited to analyze and visualize the relative contributions of multiple tables or groups of variables simultaneously, and allows for the disambiguation of the various contributions of either a population or a set of variables in a plot. The observations must all be the same for MFA, but analysis can either evaluate a single population, with variables grouped in ways that are useful or valuable to isolate, or with separate populations, using all the same variables for both groups. The number of tables (i.e., populations or groups of variables) that should be included in an analysis is limited by what makes sense, either mathematically in terms of planned analyses or visually on the plots of the partial factor scores. In any case, the visualization output for this plot provides the researcher with factor scores of the observations overall, and partial factor scores showing how each of the tables contributed to each observation; where each individual weighted table would fall in the factor space relative to the other/s. Because the tables for this analysis are weighted according to their overall inertia, with larger tables being weighted less than smaller tables, this is a very useful technique when dealing with unbalanced groups.

### 0.2.6 Inference Methods.

Because the methods outlined above are not inferential methods, and do not inherently allow for hypothesis testing, we need to also apply methods that help with that. To acheive this, we use permutation testing (Berry et al., 2011) and bootstrapping (Hesterberg, 2011).  
Permutation testing shuffles the data and recomputes the eigenvalues for each iteration. Because the eigenvalues extracted from the shuffled data tables are also an indication of how much variance is extracted by each dimension, random data should give us smaller eigenvalues, indicating a weaker signal. Therefore, if the observed eigenvalues are larger than a certain threshold, we can infer that the data we collected do, in fact, represent something real or important. Importantly, this is determined by the number of iterations that we permute, we can only infer to that degree. If we want to infer to the standard alpha level of .05, then we would need to run at least 100 permutations, and hope that the observed result was one of the largest five values.  
Bootstrapping, on the other hand, is resampling with replacement. We use this technique for two reasons: the first is to resample the factor scores to establish a confidence interval around the mean of the groups, the other is to resample with a focus on the loadings, to see which of the observations and variables load consistently on the dimensions we’re interpreting. Both give us an idea of the consistency of the data, and can once again give us an idea of the statistical significance of mean differences based on the number of iterations performed.

# 1 Methods

## 1.1 Participants

Participants (N = 604) were recruited similarly for both Experiments 1 and 2, and thus are discussed simultaneously here. Participants for this study were recruited in multiple ways. The participants in the United States (n = 292) were recruited using the traditional method of offering experimental participation credit, and also via social media. French participants (n = 312) were recruited by word of mouth, email, and social media. The only restrictions on participation were that the participant must have self-reported normal hearing. We recognize that although we suggest that data collected in this way have a much greater hypothetical reach, the data here represent a) a convenience sample, b) that is limited to participants that have access to the internet, and c) because of the nature of social media, many of the participants in the researchers’ social circles are themselves students, thus providing an additional confound. However, these specific limitations could be remedied when designing and implementing future research.  
The population we recruited was different for the two experiments. For Experiment 1, we specifically sought out highly trained musicians (n = 84) with ten years or more of music training. We recruited this population for two reasons: firstly, as a validation step, to ascertain whether the stimuli truly reflected the composer’s intent. Secondly, we had the goal of evaluating the perceptual effect of the stimuli as it relates specifically to the musical qualities. These perceptual evaluations were to then be correlated with the adjectives selected by those who participated in the adjectives survey. Participants were recruited for Experiment 2 (n = 520) without regard to level of music training.  
Of the responses to Experiment 1, 51 were removed to incomplete data (nF = 45, nA = 6), leaving a total of 33 for the analysis. Of the responses to Experiment 2, 160 were removed for not completing the survey (nF = 140, nA = 20), leaving a total of 360. Of the responses to the survey administered in the US, participants were excluded from analysis if they indicated a nationality other than American. “Asian-American,” for example, was included, but “Ghanian” was not. This left a total of 279 survey responses for Experiment 2 and 312 for analysis across both experiments.  
All recruitment measures were approved by the UT Dallas IRB.

## 1.2 Material

### 1.2.1 Stimuli.

All stimuli were original, novel musical excerpts, in various western styles, composed for this study. They were designed to evaluate a number of musical dimensions and control for others (e.g., timbre). The stimuli were all string quartets, in order to control for the confounding factor that different instruments are fundamentally described in different ways. All stimuli were between 27s and 40s long, with an average length of 32.4s. The intent was to have all stimuli be around 30s long while preserving musical integrity. All stimuli were composed using finale version 25.5.0.290 [cite finale] between April 13 and June 18, 2020. Stimuli were recorded as wav files directly from finale using the human playback engine and embedded into each question in qualtrics in that format.

### 1.2.2 Surveys.

There were two separate surveys presented to participants. The survey used in Experiment 1 (hereafter: Qualities Survey/QS) evaluated the musical stimuli on concrete musical qualities like meter and genre. The survey used in Experiment 2 (hereafter: Adjectives Survey/AS) asked participants to evaluate the stimuli using adjectives using the CATA paradigm. Both surveys also captured participants’ demographic data, including age, gender, nationality, occupation, and musical experience.  
The qualities assessed in the QS were selected from standard music-theoretical descriptors of western music. 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*. The full list of musical qualities and answer choices is listed in the supplementary materials. The words for the AS were selected using Wallmark (2019) as a guide and in consult with a French professional musician. Some words were initially selected in French and some in English. In all cases, words were selected for which there was a clear French (vis-à-vis English) translation. The words are listed in English and in French in the supplementary materials.

## 1.3 Procedure

Participants were provided with a link to either the AS or the QS. Both surveys were administered using Qualtrics. After standard informed consent, participants listened to 15 excerpts and answered questions. Participants were instructed to listen to the excerpts presented either using headphones or in a quiet listening environment, but that was not strictly controlled, nor was it part of the survey. Participants in Experiment 1 answered 10 questions per excerpt, rating the excerpts using the qualities and scales provided. Participants in Experiment 2 answered a single question per excerpt, in which they selected any and all adjectives that they felt described the excerpt. Demographic survey questions followed the experimental task.

### 1.3.1 Data Processing.

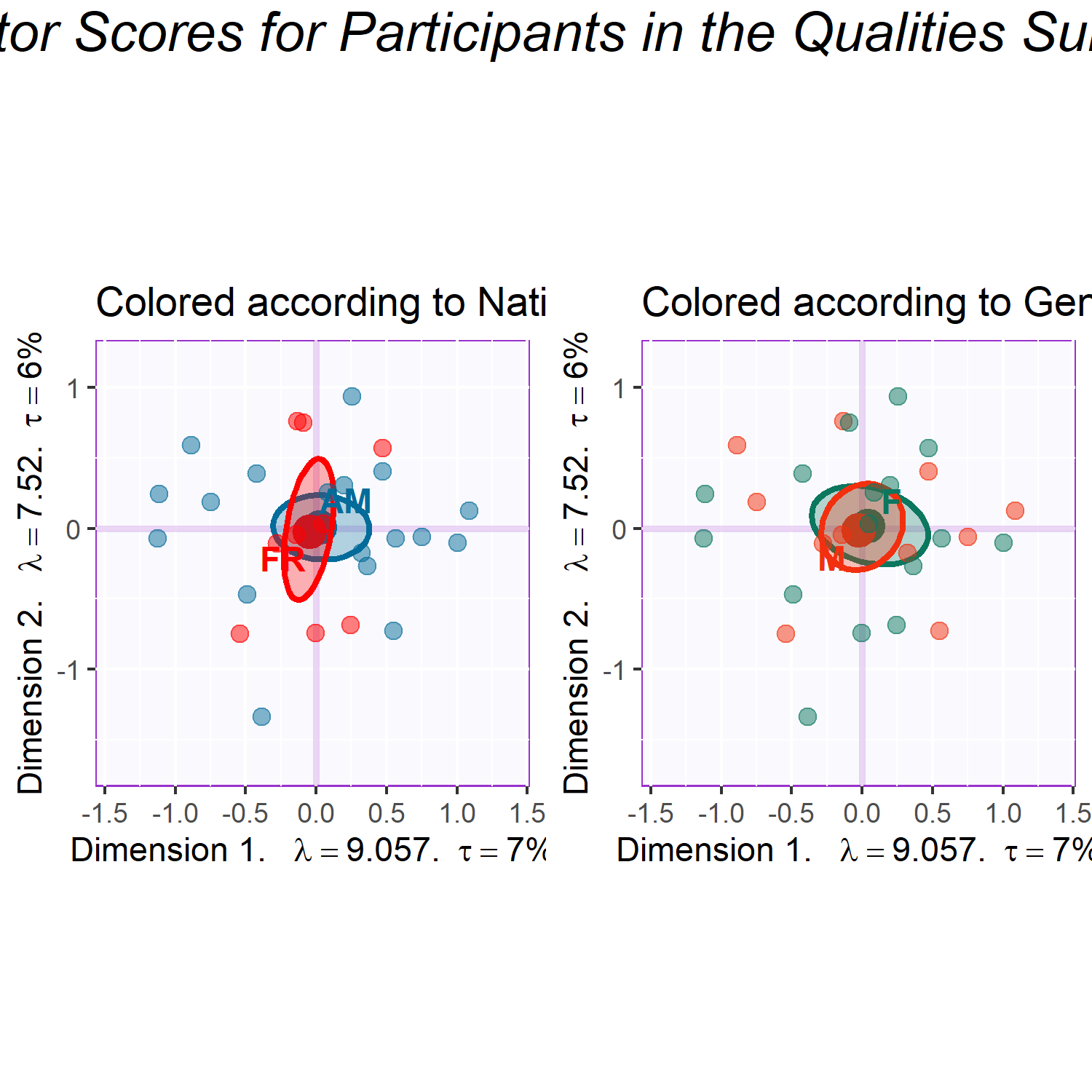
Raw data were cleaned and processed in Excel and R. This included translating all French responses to English for ease of analysis. Data were cleaned and transformed into a pseudo contingency table for each participant, with the stimuli, as observations, on the rows and the responses as variables on the columns. In these individual tables, a one (1) at the intersection of each row or column indicates that the participant selected that adjective or musical quality for that stimulus. A zero means that they did not. These individual tables were all compiled into into two ‘bricks,’ or three-dimensional arrays of data with the same structure for the rows and columns, and the participants on the third dimension, which we will refer to as ‘pages’ here. Each array was then summed across pages into a single, two dimensional, summary pseudo-contingency table, so that any given cell contained the total number of times a participant selected a given adjective or quality for a given stimulus.  
Since we did not use *a priori* grouping variables for the excerpts or adjectives, the summed tables were evaluated using hierarchical cluster analyses to see what groupings arose during evaluation. Hierarchical cluster analyses, included in supplementary materials, captured groupings of the excerpts when rated by the adjectives and when rated on musical qualities. The musical qualities were grouped by quality (e.g., levels of tempo, types of genre). These groupings were used for coloring on the plots and for statistical inferences.

# 2 Results

## 2.1 Experiment 1: Musical Qualities Survey

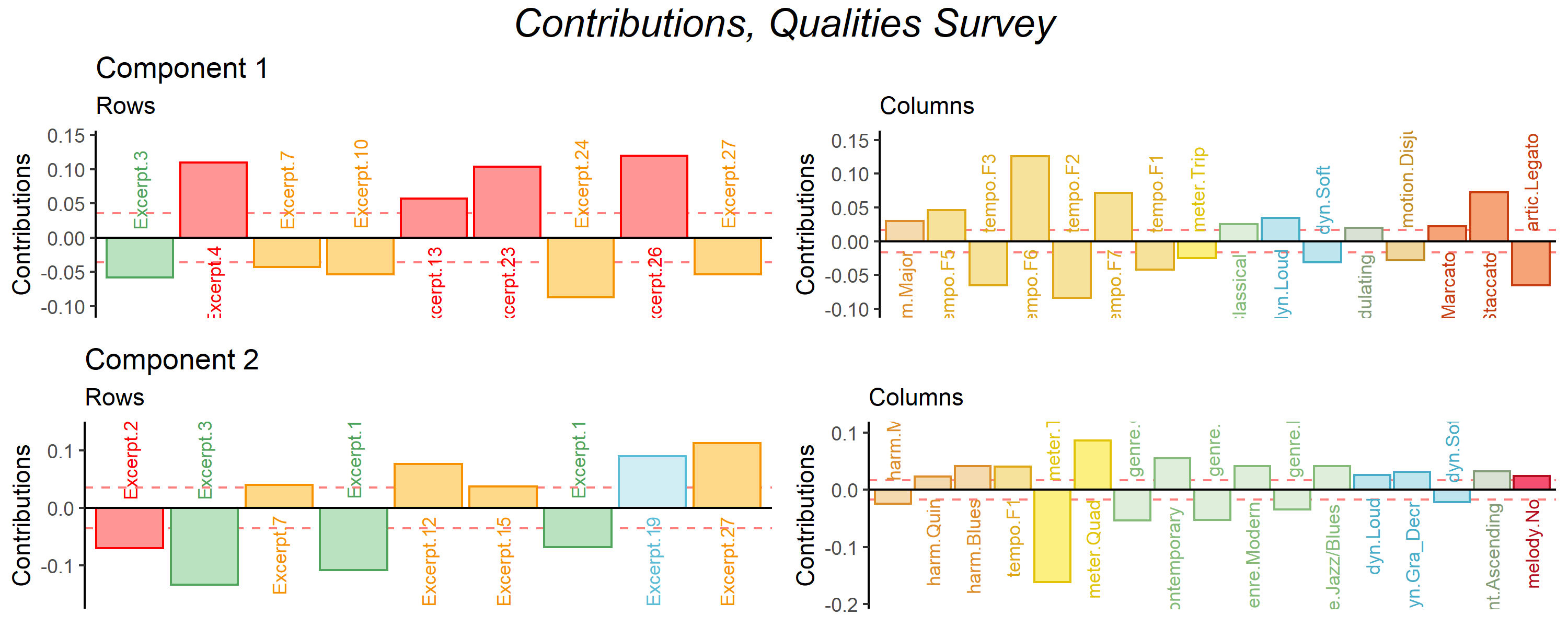
### 2.1.1 Participants.

The scree plot in Figure ?? shows the eigenvalues for the distance analysis between musical experts. The usual guideline of analyzing only dimensions with eigenvalues greater than one seems prohibitive here, as all dimensions except the last have > 1. For the purposes of this case study, we’ve opted to focus on the first two dimensions, with = 9.06 and = 7.52, respectively. This scree plot suggests that each of the participants is contributing similarly to the dimensionality of this analysis. To evaluate this, we ran a Multidimensional Scaling (MDS) analysis on a double-centered cross product symmetric distance matrix calculated from the pages of the brick. This analysis revealed no significant difference between the experts based on any of the grouping variables used. The factor plots in Figure ?? show how the means of the factor scores, grouped by nationality and gender identity, respectively, show the means clustered on top of one another, right at the origin. The overlapping ellipses are the confidence intervals for the means.



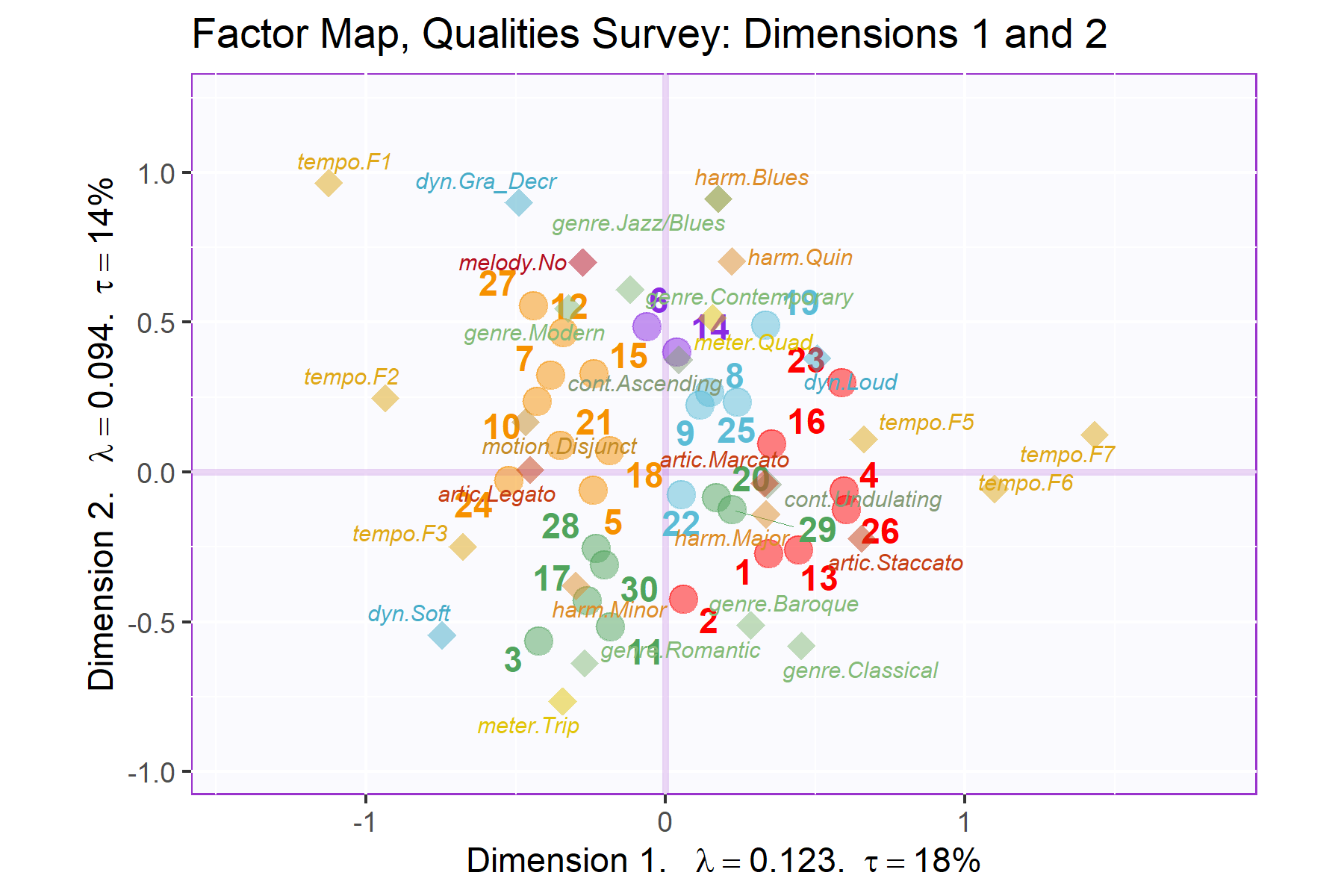
### 2.1.2 Excerpts.

The scree plot for the analysis of the musical quality ratings survey, Figure ??, shows the high dimensionality of this space, with the first three dimensions extracting a total of 18.44%, 14.09% and 8.81% respectively, totaling only 41.34% of the variance. It isn’t until we get to the 11th dimension that we see >80% of the variance explained. However, given that the assumption in an analysis like this is that the sample is random, it’s important to take these numbers with a grain of salt. Music itself is not random, and in a single excerpt of music of the type that was presented in this study, repetition is common, and some musical qualities are inextricably linked, for example some stylistic elements with genre. Graphing the variable loadings (see Figure ??) of the musical qualities shows which ones contribute the most to the first two dimensions. Because of how CA is calculated, we know that the excerpts that load on the same dimension and direction as the musical qualities are the excerpts that are most associated with those qualities. The contributions shown here are only those that contribute significantly to the first two dimensions. There are some obvious groups of variables, especially tempo and articulation in the first dimension, with fewer contributions from the dynamics group. The tempo variables, which are a continuum, load from high (tempo.F6 and tempo.F7) in the positive direction to low (tempo.F2 and tempo.F1) in the negative direction. Other contributions are one-off: major harmony, triple meter, classical genre, undulating contour, and disjunct motion. The excerpts that load positively, and are therefore associated with the qualities that load in the positive direction, are all from group 2: Excerpts 4, 13, 23, and 26. The ones that load in the negative direction are from mostly from group 4: Excerpts 7, 10, 24, and 27, with one from group 3, Excerpt 3.  
The second dimension seems to dominated by a few groups: harmony, meter, genre, dynamics. The one-offs are slow tempo, ascending contour, and “no melody.” The excerpts that load significantly on this dimension are from all four groups. In the positive direction, it’s Excerpts 7, 12, 15, and 27 from Group 4, and Excerpt 19 from Group 1. In the negative direction it’s Excerpts 2, 3, 11, and 17. All are from group 3 except for Excerpt 2, which is from Group 2. A full enumeration of contributions, loadings, and boostrap ratios is available at the github url in the author note.



### 2.1.3 Discussion.

The graph depicted in Figure ?? is a biplot depicting how excerpts and variables plot in the same space. This biplot is possible because of the nature of correspondence analysis. Because the rows and columns of the contingency table X by definition have the same variance, the eigenvalues extracted from any matrix are the same as . Thus the axes on which the factor scores are plotted are the same for both the rows and the columns. However, interpretation requires some discernment. 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 indicates a correlation of 1, an angle of 90 indicates a correlation of 0, and an angle of 180 indicates a correlation of -1.  
Overall, this helps us to evaluate what qualities contribute to the excerpt groupings. These first two dimensions suggest that the hierarchical cluster analysis (see supplementary materials) revealed groupings roughly according to genre. However, there are two notable outliers. Excerpts 6 and 14 are unique in that they are each the only representative of their respective genres. Excerpt 6 is minimalist, à la Steve Reich, and Excerpt 14 is jazzy. Preliminary versions of this analysis showed that they dominated the 2nd and 3rd dimensions, respectively (see supplementary materials for visualizations). In the plot below, they are included instead as supplementary projections, essentially ‘out of sample’ elements. Their placement on the plot below alludes to the fact that the dimensionality of this space may in fact be related to musical genre or family. Although they dominated the space when included in the sample, they are much closer to the barycenter of the plot when included as out of sample. Were they to fall exactly on the origin, that would suggest that they shared no information whatsoever with the other excerpts included in the analysis. The disparity between their placement on the graph below and their placement on the graphs in which they are included in the main sample suggests that they share some information, but there is still a large amount of information that is not accounted for in the factor space depicted in Figure ??.  
One perceptual element that is revealed here is that tempo and dynamics seem to contribute, intensity-wise, similarly to the first dimension. This points to two specific things. Firstly, it highlights possible bias in the compositional process. The excerpts were not intentionally composed with those characteristics being similar in mind, but it’s entirely possible that the high or low arousal levels of the various excerpts that participants respond to also drove some of the compositional process, and that turned up in the results. Secondly, it’s possible that the level of arousal was conflated between various musical qualities. For example, the intensity and therefore tempo of a stimulus may have been affected by the volume or dynamics (Kamenetsky et al., 1997). Perception of tempo is also affected by note rate or event density, which is also tied to arousal. In two pieces played at the same tempo, the one with more notes per unit time is more likely to be judged faster than one with fewer (Drake et al., 1999). There are also a few musical elements revealed from the associations. The term staccato means short, or light and separated, and the term legato means smooth and connected. The participants in this experiment didn’t have access to the notation, so they would be judging the excerpts aurally only. Between faster and slower excerpts, notes of the same rhythmic value take up less time in the faster excerpts, and may be more likely to be judged as light and separate, regardless of what the actual articulation was. Slow tempo and legato are associated differently. In terms of performance practice or pedagogy, slow notes are often intended to be connected as smoothly as possible, in order to create a sense of continuity. In terms of genre and harmony, many genres have harmonies associated with them (Kennedy et al., 2013), and the coordinate mapping of jazz/blues harmony and genre (on the third dimension) is the most extreme example of this. A glance back at the factor scores plot shows us more detail: the older styles, baroque, classical, and romantic, are negative on the 2nd dimension, as are the simpler harmonies of major and minor. Likewise the newer western styles, impressionist, modern, and contemporary, load positively on the 2nd dimension, along with the more complex harmonies of chromatic, whole tone, and ambiguous. A brief historical survey of the development of western harmony provides an interpretation for this. The classical genre has fairly structured rules for both harmony and voice leading, but the romantic era relaxed those rules and introduced more complex harmonies. The gradual devolution of those rules and the increase in complexity of harmony continued through the modern and contemporary styles (Kennedy et al., 2013). Historically speaking, the whole tone scale wasn’t used commonly until the impressionist era. It is worth remembering, however, that because of the nature of this survey, these results tell us more about the perception of the excerpts themselves rather than the behavior of the participants. Because the excerpts were composed with the intent of varying across all of these musical dimensions, what we see is a sort of validation that there is, in fact, that variety among these excerpts, and that they are different enough to create a large and varied factor space.



## 2.2 Experiment 2: Musical Adjectives Survey

### 2.2.1 Participants.

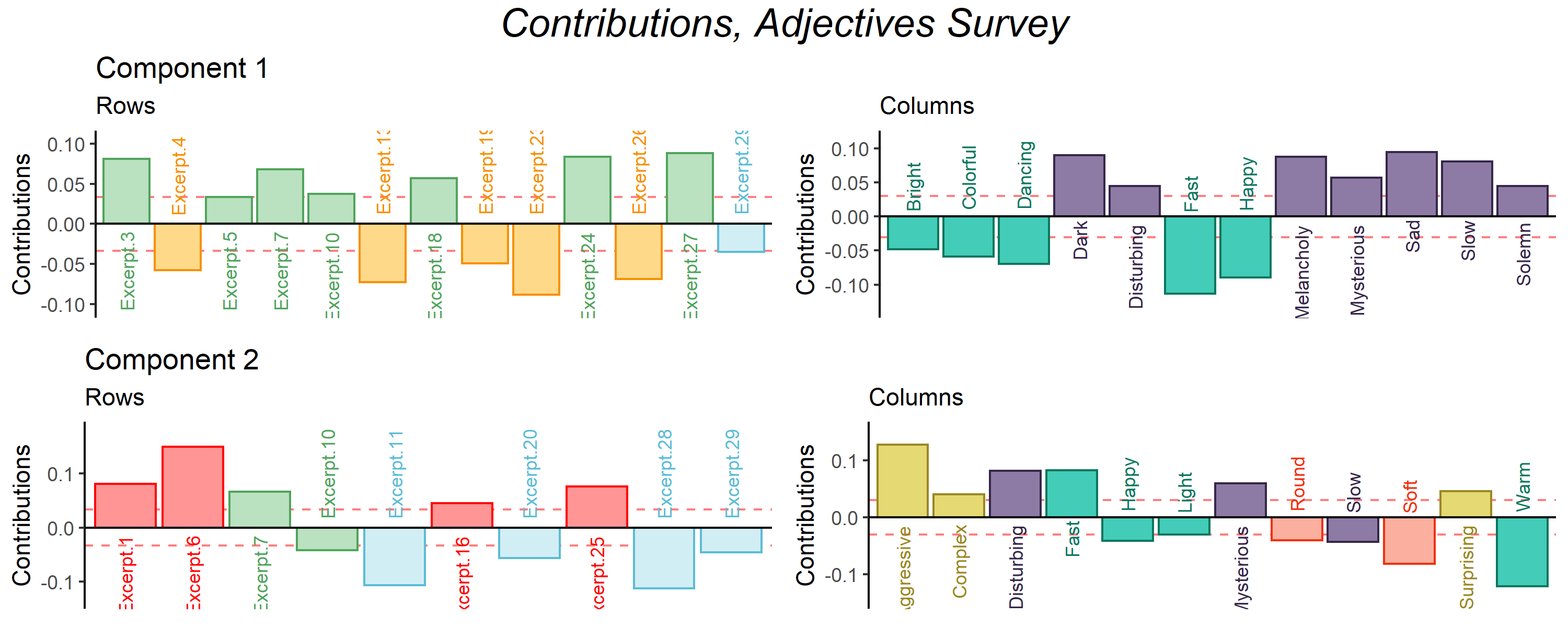
The scree plot depicted in Figure ?? shows the explained variance per dimension for the distance analysis of participants in the adjectives survey. Again, having a high number of participants means that the dimensionality is high, and each dimension is only extracting a little variance. The first five dimensions all have > 1: 1.66, 1.27, 1.13, 1.09, and 1.06, respectively, but because of the high dimensionality here, 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.  
An MDS analysis of a distance matrix calculated from the pages of the brick revealed significant group differences in how French and American participants described the excerpts, . < .01. The factor scores of the participants are plotted in Figure ??, with with group means and bootstrapped confidence intervals shown for those means. The bootstrapping resampling was performed with 1000 iterations. We also analyzed the data using two other participant groupings as factors: gender identity, with three levels: Male, Female, or Non-Binary, and level of music training, with three levels: < 2 years, 2-5 years, and >5 years. Neither of these analyses revealed any significant differences between groups.

### 2.2.2 Excerpts.

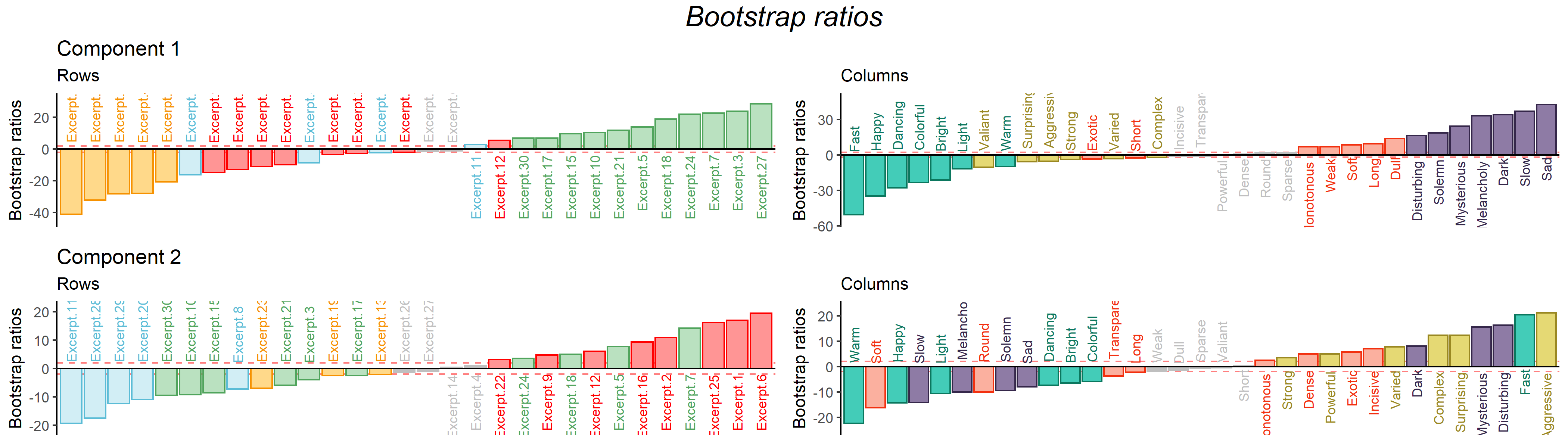
The plot in Figure ?? shows the explained variance per dimension in the analysis of the excerpts contingency table. Although there are no components with > 1, there are two strong dimensions that extract a majority of the variance. The first two dimensions extract 72.25% of the variance, with the first dimension extracting a majority: 50.05%, and the second dimension extracting almost a quarter of the overall variance: 50.05%.

This plot also suggests that there are multiple ‘elbows,’ at the 3rd, 5th, and 7th dimensions, respectively, with the third and fourth dimensions forming an ‘eigen-plane,’ of two dimensions which extract similar amounts of variance and should be considered together. For this analysis, however, we’re focused on the two first dimensions. Additionally, 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 ??. Contributing significantly to the positive end of the first dimension are excerpts from group three (green) and to the negative end are excerpts from group one (yellow). Strong contributions on the positive end of the dimension from the adjectives “Sad,” “Dark,” “Melancholy,” “Slow,” “Mysterious,” “Solemn,” and “Disturbing.” The negative end of the first dimension is defined by the adjectives “Fast,” “Happy,” “Dancing,” “Colorful,” and “Bright.” The second dimension is dominated by excerpts from group 4 (red) in the positive direction and group 2 (blue) in the negative direction. Two excerpts from group 3 also contribute significantly, excerpts 7 in the positive direction and excerpt 10 in the negative direction. The columns contributing strongly in the positive direction are “Aggressive,” “Fast,” “Disturbing,” “Mysterious,” “Surprising” and “Complex.” The columns contributing in the negative direction are “Warm,”Soft“,”Happy“,”Slow“,”Round“, and”Light".



The barplots in Figure ?? show the bootstrap ratios calculated for the rows and columns. Here we’ve included all of the rows and columns, because it’s useful to see both which are significant and which are not. This is an inferential method that tells us is how consistently each of the observations and variables load on the first two dimensions. The threshold in this case is *p* < .05. From this we get an idea of which of the rows and columns are stable, in other words, which ones tended to be rated in a certain way consistently across all participants, and also how likely these are to be observations reflective of the population as a whole. In this plot, the more extreme value of the bootstrap ratio, the more likely that it is a reflection of the ‘real’ value. The values in the center of each plot that are grayed out identify the rows or columns that are not consistently loading on the dimensions. With the observations and variables ordered like this, it makes it easy to see how the consistently the clusters are distributed in the space. This plot was not included for Experiment 1 because it would be less informative given what the survey in Experiment 1 was assessing. Experiment 1 doesn’t evaluate the behavior of participants, but the nature of the excerpts. Note that there are far more significant bootstrap ratios than there are significant contributions. That just means that while not everything is contributing, overall the model seems to be stable. Fewer significant bootstrap ratios would suggest that there was a greater amount of variance in the observations and variables than were accounted for, at least in the first two dimensions. Looking at the nonsignificant values for the adjectives may inform our understanding of the participants’ use of the adjectives. ‘Incisive,’ ‘transparent,’ ‘poweful,’ ‘dense,’ ‘round,’ and ‘sparse,’ are all nonsignificant on the first dimension, and ‘weak,’ ‘dull,’ ‘sparse,’ ‘valiant,’ and ‘short’ are all nonsignificant on the second dimension. All but ‘sparse’ are significant on one dimension or the other. Looking at the column sum for ‘sparse’ tells us that it was used, so this isn’t an effect of participants not using this word. It’s more likely that ‘sparse’ doesn’t really fit into the Valence-arousal plane. It’s a neutrally valenced word that could describe excerpts that fall anywhere within that plane. ‘Weak’ and ‘transparent’ give us another important perspective. These were the two least commonly used adjectives, but the fact that they are consistently loading on one dimension or the other suggests that when they were used, they were used in the same way.



### 2.2.3 Discussion.

The factor maps below show the row and column factor scores for the American and French participants. These are once again symmetric plots, interpretation is the same as the factor plot for the musical qualities. There’s a clear valence-arousal plane apparent for both, and in both cases valence seems to define the first dimension and arousal defines the second dimension. However, the difference in the amount of variance extracted by the first two dimensions between the French and American participants is notable. The French data show a weaker first dimension but a stronger second dimension relative to the Americans, both in terms of variance extracted (tau), effect size (lambda). This tells us that French participants were less affected by the excerpts than the American participants, but they responded more to the arousal of the excerpts. There are also differences in how the adjectives and the excerpts are distributed in the space. One clear example is that Excerpt 6 is in quadrant two in the American plot, but quadrant one in the French. This is a small change, but it suggests that the French participants were more likely to assign negative valence to this excerpt, and American Participants were more likely to assign positive valence. 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. It’s possible that this reflects the idea that although the meaning is shared between languages, there are semantic or associational differences between the words.

Another way to visualize the relative contributions of the groups to the factor space is to use an MFA, the results of which are displayed in Figure ??. In these plots, we can see the differences in behavior between the groups more clearly. A few examples of excerpts that were rated differently are Excerpts 6, 8, 12, and 17. Words that were used differently include “Disturbing,” “Round,” “Solemn,” and “Bright.” It appears that the valence-arousal plane uncovered in the CA is also present here, and this provides a framework for interpreting the differences in behavior between the 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. The way in which the two groups used the words is also curious. For example, Disturbing seems to be more extreme for the French participants than the Americans. On the other hand, “Solemn” seems to be more a function of arousal in French and valence in English. “Bright” is another example of a word that seems to have the same intent but different extremity between cultures. For American participants, “Bright” seems to carry much more positive valence than for French participants.

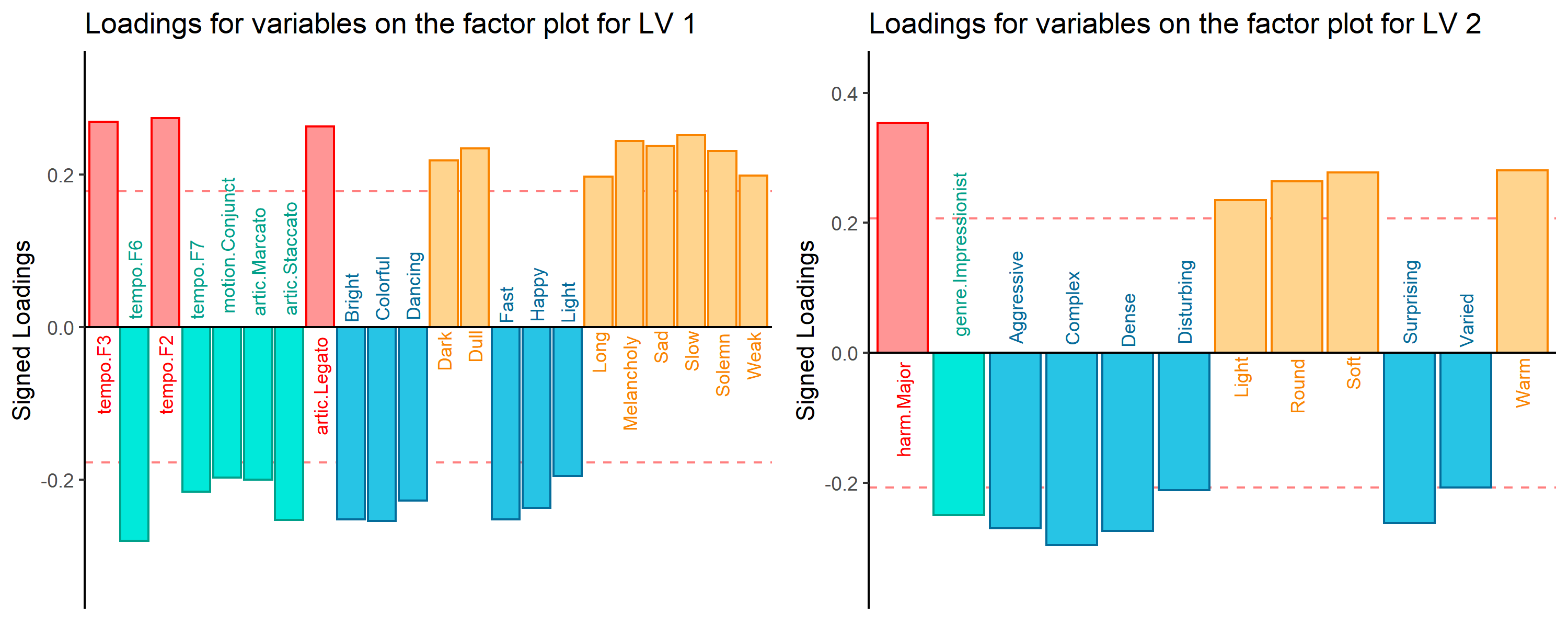
## 2.3 Experiment 3: Combined Surveys

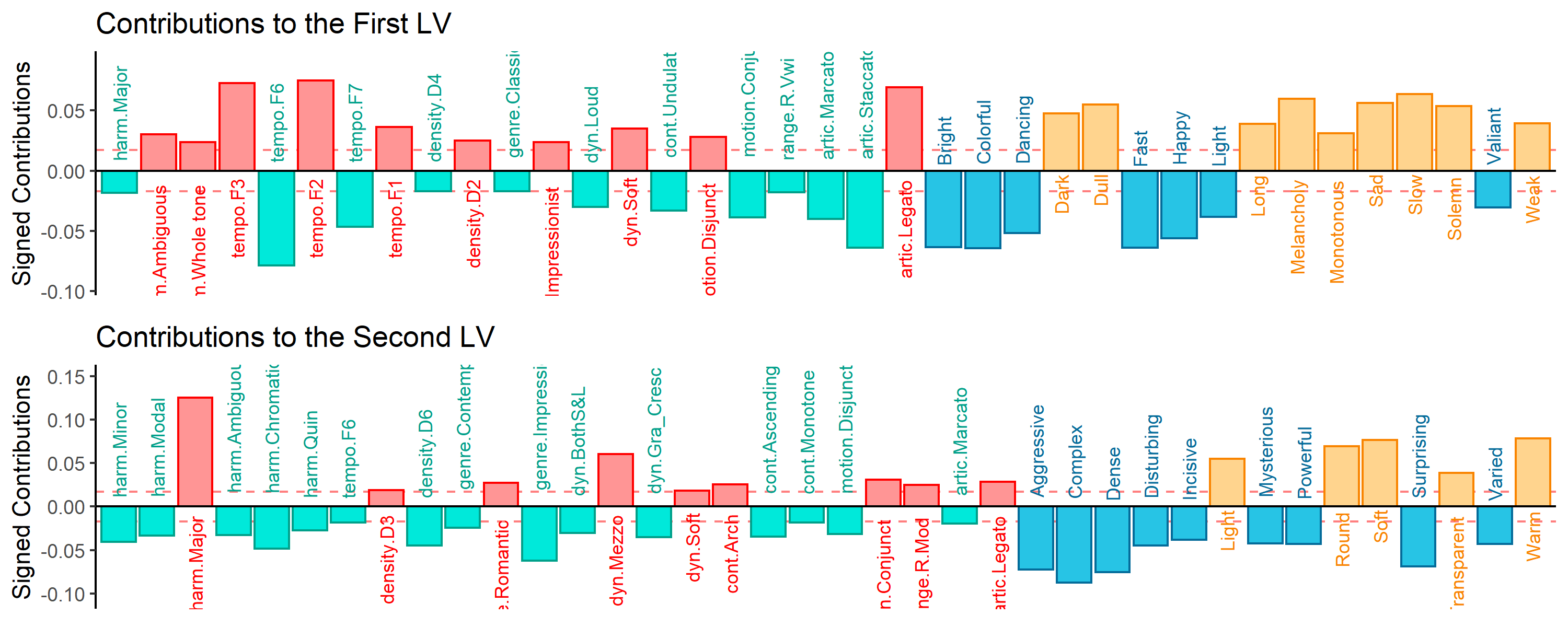
Experiment 3 used the pseudo-contingency tables from Experiments 1 and 2 together. Since excerpts 6 and 14 were excluded from analysis for Experiment 1, we also removed those rows from the contingency table for Experiment 2. This is so that the dimensions of the two tables for this PLSC would be conformable (remember that we need the same rows or columns in both tables for this analysis). The point of this experiment is to identify the strongest covariance, or the strongest shared signal, between the two tables. Now, this is not to say that these two tables are evaluating the same thing. Instead it allows us to see what is most common between two sets of different information - how often an excerpt was associated with *both* a musical quality and an adjective. The visualizations below allow us to see which variables from each of the two tables correspond with one another; 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.

### 2.3.1 Results.

This analysis revealed two dimensions that extracted the majority of the variance (83.60%). Of that total extracted by the first two dimensions, the first dimension extracted 64.35% and the second dimension extracted 19.26%. The scree plot below shows that it’s possible that there are two elbows in this graph, at the 3rd and 5th dimensions. The 3rd and 4th dimensions are also significant, extracting 6.02% and 3.67% of the variance, respectively. Interpretations of the third dimension and beyond is beyond the scope of this paper, but seeing that there are multiple significant dimensions beyond the second suggests possible future analyses and interpretations using this method.

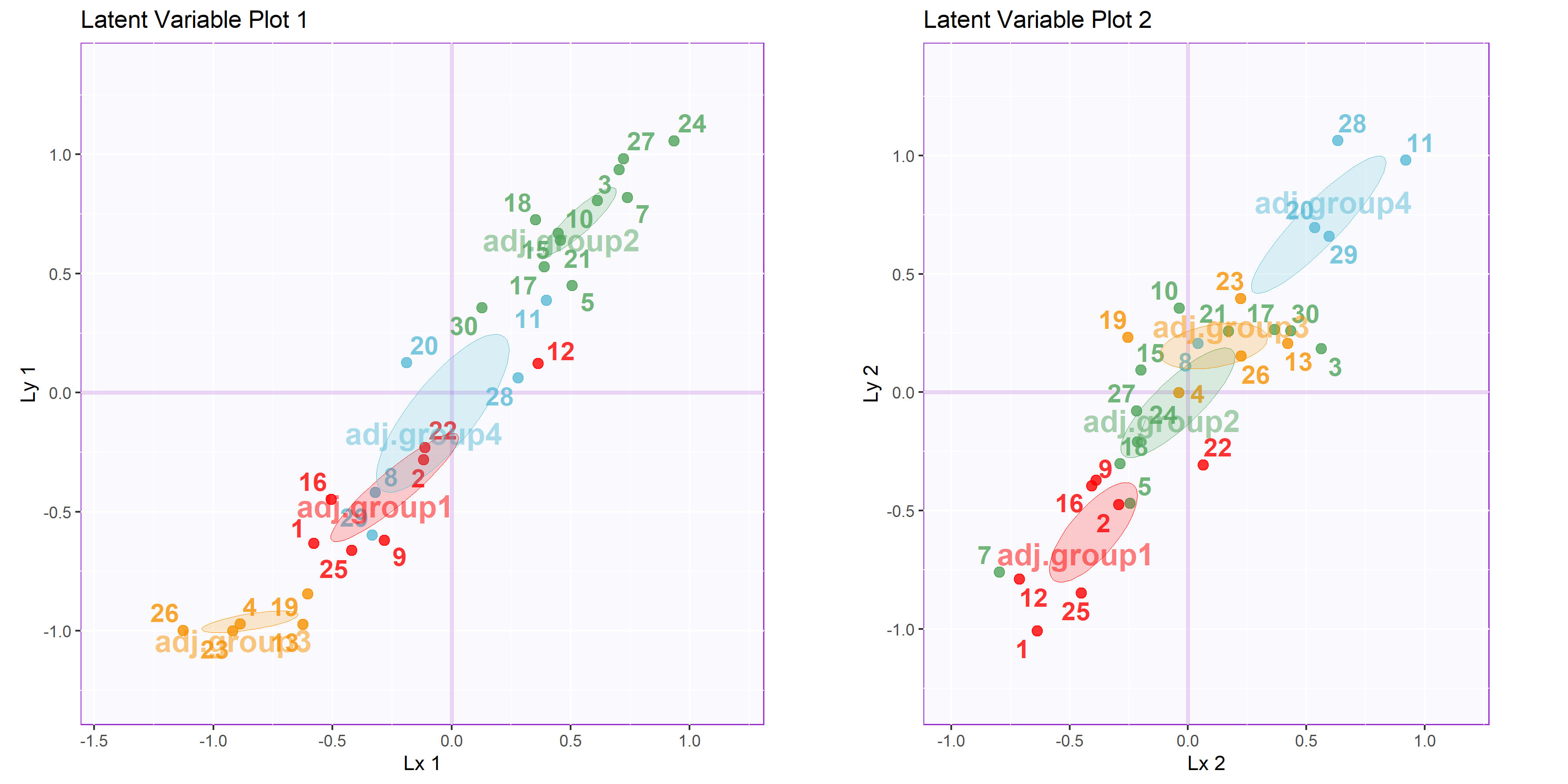
The plot below shows which variables from each data table load the most on the first and second dimensions. For the purposes of this visualization, we are showing only the variables for which 70% or more of the variance is explained. The nature of the PLSC also suggests that these are the variables that 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 mezzo 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.”



Contributions and loadings are similar, but not exactly the same. A variable’s contributions to a dimension are the ratio of the squared factor score to the eigenvalue representing that dimension (Abdi & Williams, 2010b), and loadings are the correlation between a variable and a component, effectively indicating the shared information between the two. For a more complete disambiguation, see (Abdi & Williams, 2010b). Figures ?? and ?? show us that there are quite a few more variables that contribute significantly to these dimensions than for which a significant portion of the variance is explained. We do see similar groups, however: on the first dimension, the tempo variables are contributing 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 negative direction, and Dark, Dull, Long, Melancholy, Monotonous, Sad, Slow, Solemn, and Weak in the positive direction. What’s notable here is that while some of these variables did contribute significantly in the plots above (see Figure ?? and Figure ??), some didn’t contribute much at all and fell near the barycenter of the factor plot. We also see that this juxtaposes some negatively and positively valenced adjectives, which allows us to identify which of the musical qualities contributes to the valence dimension. The second dimension tells us a similar story. Here we see more of the harmony variables, along with one tempo variable, some density, genre, a few dynamics, contour, motion, range, and articulation. 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. Again we see similar effects of variables that may not have contributed significantly to their respective plots above, but are contributing significantly here. Also, this second latent variable seems to be defining the arousal dimension.  


### 2.3.2 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. This factor plot shows us how the strongest correlated signal between the two data tables separates Excerpts groups 2 and 3, but groups 1 and 2 didn’t contribute much to this dimension, instead contributing to the 2nd latent variables. The second latent variable separates Groups 1 and 4, with Groups 2 and 3 more barycentric. This suggests that, generally speaking, the excerpts that were clustered in groups 2 and 3 are those that could be defined by positive and negative valence, respectively, and those in groups 1 and 4 would be defined more by high and low arousal. That being said, these excerpts are not defined exclusively along these dimensions, but rather more by one than the other. For example, Excerpt 26 is characterized by being one of the most extreme examples of positive valence, but doesn’t score as highly on the arousal dimension, similarly with Excerpt 27 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 barycenter for that group is near the origin of that plot.



# 3 General Discussion

Although this study was designed to evaluate the sensory or cognitive response to music, and not specifically the emotional response, there is significant overlap in the results observed here and the results of the work investigating music and emotion. The appearance of the valence-arousal plane in the results of Experiment 2 was not unexpected, even though the adjectives we selected were not intended to be explicitly emotional. This goes to show difficult it is to avoid any emotional content when selecting descriptors, and from another perspective, how much emotional contagion the musical examples carry. Overall, this supports the idea that the first two dimensions on which music is judged holistically are valence and arousal. Some of the results discussed in Experiment 1 require more explanation. In that experiment, there was an issue of having two individual excerpts dominate the factor space, numbers 6 and 14, which did not happen in Experiment 2. One of the differences between CA and PCA is that PCA is distorted by raw values, so bigger numbers have more influence on the overall factor space, but CA finds the average observation. Information that is common, therefore, falls towards the center of the plot, while information that is further from the average, in other words, more rare, ends up further from the center of the factor plots (Abdi & Williams, 2010a). Therefore, if a survey like the one used in Experiment 1 includes a item that is wildly different than the others in the set, the ratings will be very different, and that item will dominate the factor space. In this case we have two such examples: excerpts 6 and 14. Excerpt 6 was written as a Steve-Reich-esque minimalist, ostinato based excerpt, and excerpt 14 was written to be jazzy. The reason this effect occurs with the first survey and not the second is that the musical qualities on which the excerpts were rated were explicit and designed to separate the excerpts along the various musical dimensions, while the adjectives survey was designed to evaluate the excerpts more generally on holistic qualities. Excerpt 6 still appears as a minor outlier in the visualizations for the second survey, but does not dominate the space the way it does in the results of the first. What we did to mitigate that is to use those two excerpts as *supplementary projections*, sometimes also referred to as *out of sample observations*. This allows us to evaluate what information is shared by those outliers with the other elements in the dataset without having them dominate the visualization of the factor space. If, when we projected those values into the factor space, they projected onto the origin or very close to it, we would know that those observations shared no information with the other variables. The fact that they are where they are offers support to the idea that the first survey separates the excerpts approximately by genre. Because the ‘genre’ information isn’t shared with the other observations, they are being projected onto the space sharing only the information that does not deal with genre, like tempo or range. What this tells us is that musical qualities surveys captured a result that may have characterized by 4-6 factors, each approximating genre and the qualities associated with that genre and the general affective space captured an entirely different set of information about the stimuli and the perception of the stimuli.  
The hierarchical cluster analyses revealed different groupings in how the stimuli were rated between the two surveys. The PLSC then showed that when including both sets of data, there was a coherent interpretable factor space on which the excerpts were plotted. Another way to look at the results of the two surveys together would be to run a MFA, similar to the one above that plotted the difference between French and American raters on the adjective survey. This would allow us to calculate a common factor space for the two surveys without separating the first and second dimensions of each. We should note that the results of an MFA would provide us with a picture of the data that is fundamentally different from that provided by the PLSC, as it would be a true ‘common factor space’ instead of a space defined by the covariance. The important question here is simply which question is more important. In the case of these experiments, the PLSC more effectively answered our question about what musical information is commonly associated between the technical and affective qualities.  
An important overall takeaway from this is that with a deep understanding of the stimuli, we may be able to predict the approximate dimensionality of the solution factor space. In the first survey, the solution was that the first two dimensions separated the stimuli along genre or stylistic lines. Because we used only one stimulus from the minimalist and jazz genres, we had a factor space that was distorted by outliers. To have a solution in which we don’t see these specific excerpts as outliers, but as coherent members of a factor space, we would need more examples of those styles. This suggests that when creating surveys or designing stimuli, we should keep in mind that we need multiple items per group, or presumed dimension. This is not to say that we will always be able to a priori predict the factor space of the solution. For example, Experiment 2 may also have benefitted from more minimalist or jazz examples. In a system in which the overall structure is obtained by evaluating the stimuli holistically, having a single outlier will necessarily distort the space, either because it is an outlier in sensory terms or because it is the only stimulus against which there is no direct reference. This in a way embodies the issue described in the introduction, in which a single dimension is noisy. The noise, specifically in Experiment 2, comes from the fact that those participants were likely to be less familiar with mimalism and/or jazz than the trained musicians who took the QS, but the reason the results are overall robust to that noise is that the participants were not asked to rate the excerpts on any explicit dimensions or qualities.

## 3.1 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. The sample was still largely students, and France and the United States are both western countries that share western musical culture. To truly address this question, it would be very interesting to include participants from multiple, contrasting musical cultures, with languages that are more distinct than English and French. This presents new problems, however, as the specific musical qualities included in the surveys may not all apply to or translate well to other musical cultures. One specific example is harmony. The idea of the scale as the basis for both melodic and harmonic material may not be the same across all musical cultures in which scales exist. For example, in western music, the scale represents the notes that are used in a composition to form both melodies and chords, which are the harmonic material, but also specifically infer an order (Cohn et al., 2001). In carntic music, the notes that make up the ragas are not used to create chords, and do not imply a specific order (Raman & Dowling, 2017). Therefore asking participants to identify the harmonic material may not make sense in all situations, and a better question would be to ask what kind of scale or scales are used in a given excerpt. Another question that fell beyond the scope of this study is the concept of semantic drift between languages. Although illustrated in Figure ??, the source of the differences between French and American participants is not entirely clear. We humbly hazard to guess that some of the sources of the difference include aspects of perception that extend beyond the musical. These could be linguistic sources, such as the physical characteristics of the words themselves (Reilly et al., 2012), the cultural associations with the words, or the frequency of use in either language (Thompson et al., 2020). Diving more into those questions of linguistics and semantic drift between languages would be a fascinating future study. Another interesting study would be to repeat this study using adjectives from specific domains or that that avoid explicit emotional or musical content, to see how music maps onto different sensory spaces. For example, ‘moist,’ ‘slimy,’ ‘dry,’ ‘puckered,’ ‘smooth.’ Although some of these adjectives may carry musical weight, in the context of other words that all relate to haptic sensation, it may provide some interesting feedback regarding how the music maps into other sensory domains. Finally, using these studies may provide pilot work for the way in which people without language react to music, nonverbal autistic people, for example. Whereas this study explicitly uses language as an interlocutor for music perception, it offers insight into ways to better communicate with people who do not have that ability.

# 4 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, we are able to more readily and consistently reach participants who 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. Developing investigative paradigms that are accessible on mobile platforms and that reduce participant demand while maintaining rigor and integrity will similarly enable us to reach a much greater subset of the population. If we are able to pair this kind of data gathering with appropriate analysis, we can maintain the standards of scientific integrity that we as a community expect from traditional hypothesis testing. The literature to date in the music cognition domain has focused on a fairly small subset of the multivariate analyses available to investigate these questions. As presented here, the number of ways that exist to analyze the data from a single set of experiments is considerable, and the results of each analysis illuminate different parts of the story the data are telling. Not every form of analysis is appropriate in every context, but understanding how, and perhaps more importantly when, to apply a technique or type of analysis is an important to uncovering new perspectives or insights.

# 5 References

Abdi, H., & Williams, L. J. (2010a). *Correspondence Analysis* (N. Salkind, Ed.). Sage.

Abdi, H., & Williams, L. J. (2010b). Principal component analysis Tutorial Review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(August), 1–16.

Abdi, H., & Williams, L. J. (2013). *Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression: Vol. II* (B. Reisfeld & A. N. Mayeno, Eds.; pp. 1453–1454). 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*(2), 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>

Benzécri, J.-P. (1973). *L’analyse des données.* (p. 615). 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., 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, pp. 1–614). 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>

Drake, C., Gros, L., & Penel, A. (1999). How fast is that music? The relation between physical and perceived tempo. *Music, Mind, and Science*, *101*, 190–203.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis* (pp. 1–376). Academic Press.

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.* (pp. xiv, 975–xiv, 975). Oxford University Press.

Kamenetsky, S. B., Hill, D. S., & Trehub, S. E. (1997). Effect of tempo and dynamics on the perception of emotion in music. *Psychology of Music*, *25*(2), 149–160. <https://doi.org/10.1177/0305735697252005>

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>

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>

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>

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>

Wallmark, Z., & Kendall, R. A. (2018). *Describing Sound* (pp. 1–37). <https://doi.org/10.1093/oxfordhb/9780190637224.013.14>

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>