Affective Analysis of Music Using the Progressive Exposure Method: The Influence of Bottom-Up Features on Perceived Musical Affect

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By
Joshua David Albrecht, M.M.
Graduate Program in Music
The Ohio State University
2012

Dissertation Committee:
David Huron, Advisor
David Clampitt
Robert Cudeck
Abstract

Existing paradigms for measuring the perceived affective content of music each present their own unique strengths and limitations. This dissertation describes a series of studies conducted to develop and implement a new paradigm called the *progressive exposure method*. This method presents a long passage in short, discrete segments and asks participants to rate the perceived affective content of those segments. This study uses the second movement of Beethoven’s *Pathétique* sonata (No. 8, Op. 13) as a case study for the method. The results provide a mosaic portrait of the perceived affective content of surface features in the movement. From this data, a model of perceived affective content in the movement is constructed and is tested for generalizability across many excerpts sampled from the Beethoven piano sonata corpus.
Acknowledgments

I would first like to thank my advisor, David Huron, for his personal guidance and investment through my time at Ohio State. His investment into me turned me into a researcher. I would also like to thank my other committee members, David Clampitt, for his eagerness to help me in my career and the thought-provoking conversations and classes we have had together; and Robert Cudeck, for going above and beyond the duties a statistics professor is called to perform and humoring countless statistical questions throughout my time at Ohio State. I would also like to thank Joseph Plazak and Daniel Shanahan, my readers who graciously read the document and provided valuable suggestions. Of course, I could not have done this work without the constant encouragements, challenges, and conversations of my lab mates, Yuri Broze, Gary Yim, Brandon Paul, Katelyn Horn, Niall Klyn, Randolph Johnson, Olivia Ladinig, and Laura Tiemann. Thanks above all to my gracious and patient wife Janelle, who endured long days and nights of writing. Finally, I would like to thank God my Father, whose grace to me has provided so much, including the strength to persevere through this work. May this work be for His glory.
Vita

May 2005 .................. B.M. Music History/Theory, The University of Wisconsin-Whitewater

May 2007 .................. M.M. Music Theory, The University of Texas at Austin

2009-2012 .................. Graduate Teaching Associate, Music Department, Ohio State University

Publications


Fields of Study

Major Field: Music

Specialization: Graduate Interdisciplinary Specialization in Teaching
Table of Contents

Chapter 1: Overture and Literature Review .......................................................1

Chapter 2: The Progressive Exposure Method ..................................................14

Introductions ........................................................................................................14

Existing Measurement Methods .......................................................................14

The Progressive Exposure Method ...................................................................17

Affective Analysis Using the Progressive Exposure Method .............................23

Chapter 3: An Initial Study to Ascertain Appropriate Affective Dimensions for Use in the Main Study .................................................................26

Introduction ........................................................................................................26

Participants .........................................................................................................28

Stimuli ................................................................................................................29

Procedure ...........................................................................................................35

Content Analysis ...............................................................................................37

Discussion .........................................................................................................41

Conclusion .........................................................................................................44

Chapter 4: The Procedure for the Main Study ...................................................45

Introduction ........................................................................................................45

Participants .........................................................................................................46

Stimuli ................................................................................................................47

Procedure ...........................................................................................................49

Conclusion .........................................................................................................53

Chapter 5: Reliability Measures for the Progressive Exposure Method ............55

Introduction ........................................................................................................55

Normalization of ratings ....................................................................................56

1. Intrasubjective reliability .............................................................................58

2. Intersubjective reliability ............................................................................64

Conclusion .........................................................................................................74

Chapter 6: Affective Analysis of the Second Movement of Beethoven’s Pathétique Sonata (No. 8, Op. 13) ................................................................. 77

Introduction .......................................................................................................77

Results ...............................................................................................................78
Affective Analysis.................................................................78
Parallel Passages.................................................................81
“Extreme” ratings.................................................................90
Excerpts with maximum and minimum ratings for each affect..........................96
Maximum and minimum affective dimensions .................................................99
Correlations between affective dimensions ....................................................103
Discussion..............................................................................106

Chapter 7: Building and Testing a Model of Perceived Surface Affective Content for
the Entire Corpus of Beethoven Piano Sonatas........................................115
Introduction........................................................................115
  Constructing a model of perceived affective content in the Pathétique ..........117
  Regression Analysis..........................................................124
Follow-up study: Testing the model ...............................................128
  Stimuli.................................................................................128
  Participants ........................................................................133
  Procedure.............................................................................133
  Results – Provided descriptions..................................................136
  Results – Participant Ratings..................................................143
  Excerpt effects....................................................................146
  surprisingly, predictions were not significantly correlated with ratings (p = 0.61),
  and was in fact -0.099. .............................................................168
  Summary.............................................................................168
  Affective dimension effects.....................................................170
  Conclusion............................................................................171

Chapter 8: Coda ..................................................................173
Selected Bibliography.........................................................................183

Appendix A: Complete Surface-Level Affective Analysis of the Second Movement of
Beethoven’s Pathétique Sonata Using the Progressive Method ......................191
Appendix B: Anova Summary Tables for Treatment Effect of Different Statements of
the Theme by Affective Dimension.......................................................206
Appendix C: The Distributions of Ratings for Each Affective Dimension..............218
Appendix D: The Regression Summaries for the Eleven Affective Dimensions Tested
in the Main Study.............................................................................230
Chapter 1: Overture and Literature Review

Joseph Addison’s poetic assertion of music’s power to provoke passion and evoke emotion is not unique. Many writers have spoken at length of the power of music to both stir the soul and express powerful emotions. Virtually any book of famous quotations that has a section on music contains numerous aphorisms about the emotional capacities of music.

Moreover, this emotional passion can be observed not only in the comments made by music lovers, but also as a motivator for musicians themselves. For example, several studies have implicated emotional experiences as an important reason to pick up an instrument or study music. Gellrich, et al. (1986) identified a sensual-aesthetic motive of playing certain pieces of music as one of the most important reasons for making music. In a study of piano players’ motivations, Persson (1993) described the hedonic drive of playing music as one of several motives for performing. Additionally, Nagel (1987), in identifying ways in which making music satisfied personal needs, found that music’s emotion-inducing capabilities were an important source of that satisfaction.
Nevertheless, emotion in music remains a controversial topic. Despite the apparent ubiquity of the belief that music is emotionally expressive, scholars have debated a wide range of issues regarding emotion in music: scholars differ as to what exactly is meant by the term emotion and whether it exists at all, including what processes or mechanisms underlie the experience or perception of emotion; they differ as to what qualifies as an emotion, whether emotions can be communicated at all, and if they can whether music can communicate emotion and whether it should; scholars debate how to recognize the signs of emotion, and the distinction between induced emotion in music and perceived emotion in music; they debate whether emotion can be measured, and if so, what is the best measurement method.

How is it that music, an abstract representation of dots on a page or an abstract acoustical organization of sound, can either express emotion to listeners or induce emotion in listeners? This is a question that has been puzzling thinkers since antiquity. In *The Republic* (1955), Plato discusses the expressive capacities of certain harmonies and modes, and specifically mentions that some harmonies are expressive of sorrow, some are soft and convivial, and some expressive of bravery. For Plato and for Aristotle, music has the power to imitate states of the soul in which emotions are expressed (Budd, 1985).

Perhaps the strongest criticism of music’s power to express emotion came through the voice of Eduard Hanslick. Writing in 1854, his book *Vom Musikalisch-Schönen* was a response to what he considered the prevailing view of aesthetics in his time, what he refers to as the “aesthetics of feeling” (1986). In his first chapter, Hanslick firmly
asserts that it is impossible for any definite emotions to be represented by music. This is because, for Hanslick, music cannot represent anything other than music and specific emotions require specific conceptual objects to be attached to. Susanne Langer (1957) likewise claims that music can not directly represent emotion, but only the dynamic aspects of emotion by using motion and rest, and tension and relaxation. Peter Kivy (1989), on the other hand, accepts that music can express particular emotions by mimicking the expression of human emotive gestures. However, these expressions of emotion must remain generalized, as there is no object to which to attach emotion in music. Moreover, Kivy claims that even when people listen to sad music they do not themselves become sad; rather, they are moved by the music and simply mistake this emotion for the feeling of sadness (1989). Perhaps most famously, Stravinsky (1975) declared that “music is, by its very nature, essentially powerless to express anything at all.”

The prevailing attitude toward musical emotion in the music theory literature remained that of skepticism until Leonard Meyer’s book Emotion and Meaning in Music (1956). In this work, Meyer challenges the dominant perspective of music analysis that focuses on objective structures and does not take into account the meaning and content of music. The meaning of music for Meyer depends upon the syntax of the music, but inevitably arouses affect or felt emotion when expectations are inhibited or blocked. This work was something of a crossover between psychology and music theory, but it opened the door for further work on music and emotion in music theory. His student, Wilson Coker, published Music and Meaning, A Theoretical Introduction to Musical
in 1972, in which he claimed that music used a gestural language to communicate both musical, congeneric meaning and referential, extrageneric meaning. Eugene Narmour (1990, 1992) also built a model of musical meaning, claiming that implications and their subsequent realizations were important for the communication of meaning and emotion in music. David Huron’s book *Sweet Anticipation* (2006) also looks at the role of musical expectation for the communication of musical emotion.

The debate about emotion is not unique to the field of music, however. There is a vibrant history of debate about theories of emotion in the psychology literature. Randolph Cornelius’s *The Science of Emotion: Research and Tradition in the Psychology of Emotion* (1996) chronicles the history of much of this debate. He identifies four contrasting schools of thought that approach the topic of emotion in different ways. Cornelius first identifies the Darwinian tradition (Darwin, 1872) in which the emotions, such as jealousy and hatred, are viewed as specific evolutionary adaptations that serve particular survival functions. Operating under this view, Plutchik (1984) suggests that all of the emotions can be tied to eight basic behavior patterns, such as acceptance or rejection (often of food), protection or destruction, and reproduction, for example. According to Plutchik, these behavior patterns map onto a small set of “basic” emotions.

The second approach to emotion is associated with the work of William James (1884). James famously contradicted the standard way we think about emotion. Traditionally, the idea is that emotion arises in response to external stimuli that motivate us to act. For example, the traditional view would say that seeing a bear causes us to feel fear and that motivates us to run. James suggested instead that seeing the bear
causes us to run, and then the act of running causes fear. According to James, emotions are a reaction of the body. We perceive our own physiological responses to stimuli and then we interpret these responses as emotions. This seemingly odd notion has a surprising amount of empirical verification. For example, Stepper and Strack (1993) conducted a study in which participants sat at either a high or low desk and were assigned a writing task. The resulting posture had an influence on the participants’ experience of emotion. For a review of the literature, see Laird’s work *Feelings: The Perception of Self* (2007).

As behaviorism began to fade as the dominant subfield of psychology in the 1970’s, a “cognitive revolution” spurred changes in the approach to explaining emotion (Baars, 1986; Sperry, 1993). The first work to focus on a cognitive explanation for emotion was Arnold’s *Emotion and Personality* (1960). In it, she maintained that cognitive appraisal was an integral part to experiencing any emotion. Under this view, emotion results as the decision for what would be best for the organism after an appraisal of the situation has been made. For example, though two people may both scream while riding a roller coaster, one person may be screaming with delight as they appraise the situation as posing no threat to themselves whereas another person could be screaming with terror as they think the ride may actually kill them.

Finally, Cornelius describes the social constructivist perspective in which emotions are viewed as reinforcing social rules that guide proper behavior in ways that are socially appropriate. For example, Averill (1980) found that there was a large gender difference in the propensity to weep in public, and that these differences were tied
primarily to a cultural influence rather than biological reasons. Spellman (1989) argues that the proper emotional reaction to being publicly insulted by a friend depends on the culture. In a Western culture, becoming angry is an appropriate response so that people would not think that you were a fool for letting someone take advantage of you. However, in Japan, a public insult would probably result in shame, and the proper response would be to smile and avoid a confrontation.

Assuming that musical emotion actually happens, and given the competing theories in the psychology literature, how are we to understand what emotion in music is? Is feeling deep emotion in music an evolutionary adaptation that enhances survival? Does listening to music produce physiological changes that are then (mis)attributed as emotion? In what way does our appraisal of the experience of music influence the emotional appeal of music? Or, is feeling emotion in response to music primarily a social phenomenon?

Before going further, it is prudent to make a distinction between two different types of emotional experience in music. The first type, often called induced or felt emotion, is when listening to music changes or reinforces a listener’s emotional state. This may happen, for example, when listening to upbeat music changes the listener’s mood from sadness to a more positive emotion. The second type, often called perceived emotion, is when music is recognized as expressing or portraying a particular emotion. For example, a piece may sound sad, even if it does not actually make you sad. While there is often quite a bit of overlap between the two types of musical emotion (Eerola & Vuoskoski, in press), research indicates that participants are much more reliable at rating
perceived rather than felt emotion (Francés, 1958/1988; Hampton, 1945; Swanwick, 1973). Additionally, research suggests that the different types of musical emotion can be empirically differentiated (Evans & Schubert, 2008; Kallinen & Ravaja, 2006; Zentner, Grandjean, & Scherer, 2008). For example, Wager et al. (2008) found using meta-analyses of PET and fMRI studies that perceived and induced emotions activate different areas of the brain.

Confusion also arises in that different authors will use different terminology to describe emotions in music. In their important compendium *Handbook of Music and Emotion* (2010), Juslin and Sloboda note that there are many different terminological uses across the field. For the purpose of clarity, they suggest the following terminological conventions: (1) *affect*, used as an umbrella term that covers all positive or negative states, including emotions, moods, preferences, and feelings; (2) *emotion*, referring to a brief but intense affective reaction, including feelings, arousal, expression, and action tendencies; (3) *mood*, referring to less intense affective states that do not have a clear object and last much longer than emotions; and (4) *feeling*, which refers to the subjective experience of emotions or moods.

A few important considerations arise from the way these terms are defined. It is important to note that, as defined above, the only thing that we have conscious access to are feelings. It is common to assume that our feelings are accurate representations of the emotions or moods that we experience. However, some researchers have investigated “unconscious affect” and “preattentive affective behavior” (Winkielman & Berridge,
2004; Öhman & Mineka, 2001), in which affective experiences are inaccessible to the consciousness.

What exactly constitutes an emotion is also contentious, both in the general psychological literature and in the music and emotion literature. If, as in the Darwinian perspective, emotions are motivational amplifiers that encourage behaviors leading to greater survival benefits, then feelings such as hunger or itchiness may be construed as emotions, because they encourage beneficial activities such as eating or protecting the skin. However, there has been a tendency in the literature to focus on a few emotions that appear “basic” to all of humanity. Ekman (1992), for example, conducted cross-cultural research and found that there were clearly recognizable cross-cultural facial expressions for emotions such as happiness, sadness, disgust, fear, anger, and surprise. As mentioned above, Plutchik (1984) identifies eight basic emotions tied to survival behaviors. Other authors identify four, eight, fourteen, or even sixteen basic emotions (Lazarus, 1991; Roseman, Spindel, & Jose, 1990). According to these authors, the broad and variegated range of emotions that we experience are all combinations of these basic emotions. Of course, the difficulty with this approach is that there is so much disagreement between authors as to what actually constitutes a basic emotion.

A different approach, known as the dimensional model, is similar in spirit to the basic emotion model. In this approach, all of the emotions are essentially reduced to their principal components, represented by one (Duffy, 1941, who used arousal), two (Russell, 1980, who used valence and arousal), or three dimensions (Osgood et al., 1957, who used valence, activation, and power). By representing each dimension as an
orthogonal axis, any emotion can be plotted in the defined space. For example, anger, represented using the most common two-dimensional circumplex model, would be located highly negatively on the valence dimension and highly positively on the arousal dimension. Serene, on the other hand, would be plotted as highly positive on valence and fairly negative on arousal. One difficulty with this method is that different emotions that are experientially quite different, such as fear and anger, may both be plotted in similar locations.

Each of the approaches outlined above presumes some sort of a priori explanatory framework for emotions. Rather than prejudging different theories about emotions in music, it seems prudent to investigate musical emotions and allow the investigation to lead to theory. However, terms like emotion already bear heavy connotations from pre-existing theories. In order to not a priori exclude any possibilities for musical emotion, the remainder of this dissertation will use the term affect, because this term casts the broadest net.

How then does music express or convey affect to a listener? Drawing from communication theory, Kendall and Carterette (1990) propose a model of musical communication (Figure 1.1). They posit that for any communication to take place, there must be a transmission and reception of messages. In the most common case, a composer writes music (represented by C in the figure) using notational symbols that are then interpreted by a performer (P in the figure), who transmits the music via acoustical signals to the listener (L in the figure) who decodes those signals into meaning. For musical communication to be successful, the original communicative idea of the
Figure 1.1. A model of musical communication. C represents the composer, who encodes his/her ideas into musical notation. P represents the performer, who decodes the notation and recodes the message into an acoustic signal. L represents the listener, who receives the acoustic signal and finally decodes it (Kendall & Carterette, 1990).

composer would have to be similar to the received idea of the listener. A lens model specifically designed to discuss the transmission of emotion through music is presented by Juslin and Timmers (2010), shown in Figure 1.2. In this model, the listener receives an acoustic signal from the performer and decodes the perceived affective content of that signal. The signal is a representation of the intended emotional meaning encoded by the performer into cues such as tempo and articulations, and cues from the composer such as mode, pitches, and rhythm. Achievement can be measured through the accuracy with which the composer’s and performer’s affective intention are mirrored in the listener’s decoding.
Figure 1.2. The lens model of musical communication. In this model, the composer encodes the intended emotional expression into features like mode, pitch, and rhythm, the performer encodes intended emotional expression into features like tempo and articulations, and the listener decodes all of these features into perceived emotion (Juslin & Timmers, 2010).

As mentioned above, the measurement of affect in music can be difficult, for a listener’s feelings are the only thing that the listener has conscious access to. As a result, self-report is the most common means by which affect is measured in music (Västfjäll, 2002). This can take many forms, but always involves a listener or performer giving feedback about affective experiences in music. Common approaches involve having a participant listen to music and rate the affect of the music on a Likert scale (Mcnair, Lorr, & Droppleman, 1981) or check off appropriate adjectives from a checklist (Thayer, 1986). Alternatively, participants could keep a diary, in which they record their
daily emotion episodes (Sloboda & O’Niell, 2001) or simply provide free responses to
musical experiences (Gabrielsson & Lindström, 2003). With most self-report methods,
participants listen to music and then provide a retrospective rating of the affective
content of the music (Balkwill & Thompson, 1999; Gabrielsson & Juslin, 1996; Vieillard
et al., 2007). This method tends to be fairly straightforward to analyze and tends to
provide relatively clean data.

The retrospective response paradigm described, however, fails to take into
account how affect could change in music over time. A different type of self-report,
known as the continuous response paradigm, allows participants to rate the affective
content of music continuously throughout a listening experience. This paradigm,
originally used by Nielsen (1983) received considerable attention beginning in the
1990’s, most prominently from Schubert (1996, 1999, 2002, 2004). In this paradigm, a
listener will typically adjust a dial or a scale while listening to music. Another approach
requires listeners to move a cursor in a two-dimensional space. One difficulty with
continuous data collection is that it can be difficult to link changes in ratings with
musical features. This is because different musical features require different amounts of
time to process and respond to: changes in loudness can influence ratings within one
second, whereas changes in texture take two to three seconds to influence ratings
(Schubert, 2004). It may also be that responding in real-time puts a larger cognitive load
on participants, and thus reduces the intersubjective reliability.

Although self-report is the most common measurement method used, it poses
some problems. As mentioned above, not all affective experiences may be readily
accessible as feelings to the consciousness. Additionally, participants may not give completely honest answers in their responses. A study by Bradburn, Sudman and Warsink (2004) found that participants can be unwilling to accurately report their internal states if they think doing so is undesirable or socially unacceptable. Indirect measures of affective response provide a means of assessing affective states without requiring self-report. This can be accomplished by measuring changes in physiology (e.g. Scherer, 2001) or changes in behavior (e.g. Russell, 2003). For example, Clark (1983) found that participants who listened to music expressing negative affects took more time to count from 1 to 10 than participants who listened to music that expressed positive affects. Other studies have shown that listening to music expressing different affects can change blood pressure, heart rate, respiratory rate, muscular tension, and many other physiological responses (Bernardi et al, 2006; Harrer & Harrer, 1977; Haider & Groll-Knapp, 1981; Davis & Thaut, 1989).

There are many conceptual and methodological issues regarding understanding, measuring, and evaluating the perceived or felt affective content of music. Despite the many advances in understanding described above, it remains a controversial topic. Affect in music is certainly a topic on which intelligent adults can disagree. Nevertheless, the emotional impact of music remains one of its most cherished aspects. The following chapters offer one particular perspective of affect in music and explore a new way of investigating and measuring the perceived affective content of music.
Chapter 2: The Progressive Exposure Method

Introduction

The relationship between music and emotion remains a topic of perennial interest, evidenced by the flurry of scholarly and popular writing recently published (Tirovolas & Levitin, 2011; Juslin & Sloboda, 2010). Despite this interest, the exact nature of the relationship between composed music and the perception of the affective content of that music remains something of a mystery. In this chapter, different methods commonly used to measure the perceived affective content of music are reviewed and critiqued, and an alternative procedure – dubbed the progress *ive exposure method* – is introduced. Although the progressive exposure method has its own limitations, it also provides a number of advantages. After describing the progressive exposure method, a study is conducted that uses this new method (to be described further in the ensuing chapters).

Existing Measurement Methods

As noted in Chapter 1, there are currently many commonly-used approaches for the purpose of trying to measure or estimate the perceived affective content of music. Chapter 1 described two of the most commonly employed strategies in recent years: *retrospective response to excerpts* and *continuous data collection*. Each methodological paradigm has both advantages and limitations.
The classical paradigm in which (usually short) musical passages are played for participants and then they retrospectively rate the perceived affective content of the passage tends to produce very clean data (Zentner & Eerola, 2010). This paradigm also allows researchers to link measurable features of the performance or composition heard to the subjective affective evaluations of listeners. One of the most influential studies in recent years employs the lens model from classical communication theory to identify how performers “encode” different emotions into musical signals and how listeners “decode” signals into perceived emotion (see Figure 1.2 (Juslin & Timmers, 2010)). However, most applications of this paradigm only allow for one dominant affect in a given excerpt. This often takes the guise of a forced-choice paradigm in which participants are given a list of potential emotions being expressed and are asked to choose one that they feel is best represented by the music (Plazak & Huron, 2011). Also, by its very nature, this paradigm does not permit any investigation into how affect changes throughout a passage or excerpt.

In contrast, the continuous data collection paradigm explicitly investigates how affect changes throughout an excerpt or passage (Schubert, 2004). Some approaches (Schubert, 1996) use a multi-dimensional model and therefore allow (at least in theory) for the investigation of complex mixtures of affect. Additionally, this approach has much more ecological validity in that listening to an entire work in one hearing more closely emulates most real listening experiences. However, the analysis and interpretation of the data resulting from this paradigm can be difficult. There are many problems that plague the data and introduce sources of noise. First, research has shown
that listeners require different amounts of time to hear, mentally process, and respond to changes in different aspects of the music (Schubert, 2004). For example, the response time observed in reaction to a change in loudness can take less than one second, whereas the response time to changes in texture require two to three seconds. If the response latencies were the same across all musical features, a simple correction would be possible. However, due to the differences in response latencies, it is difficult to attribute changes in listener responses to specific musical features. Another problem is that there tends to be a residual effect in participant response (Schubert, 2004). For example, participant responses have a tendency to begin to drift in a particular direction over time. The result is that the current rating tends to become a new baseline from which the participant makes subsequent judgments.

Finally, it is not clear that the responses are reliable between participants. Although valiant efforts have been made to generalize conclusions from this difficult data, it is often the case that there is a wide variance between participants (see Figure 2.1).

![Figure 2.1](image.png)

**Figure 2.1.** An example of continuous response data (Upham & McAdams, 2010), in which participants were asked to rate musical tension in Mozart’s Overture to *Le Nozze Figaro.*
from Upham & McAdams, 2010). It could be that this huge variance reflects a remarkable diversity in how different listeners respond to the same musical passage. However, it is also the case that there tends to be low test-retest reliability. That is, a given individual tends to respond differently when hearing the music a second time. A charitable interpretation might be that the low test-retest reliability reflects a wide range of personal experiences in listening to the same work. However, it could be that the continuous response task is too difficult. It could be, for example, that the task of continuously evaluating and responding to music places undue cognitive load on the participants. Participants may not be able to process the musical information quickly enough to provide reliable responses.

The Progressive Exposure Method

The progressive exposure method (PEM) offers a sort of compromise between these two classic paradigms. Essentially, the progressive exposure method divides a long work, movement, or excerpt into smaller discrete chunks. Like the retrospective response paradigm, the PEM polls participants after listening to each chunk. However, like the continuous response paradigm, many evaluations are collected over the course of the entire passage, and so there is the ability to observe differences in affective content in different portions of the music.

There are several ways in which the musical chunking can be performed. One approach would select chunks that cover a theoretically defined musical “unit,” like a phrase, a motive, a number of chords, or a section of music that ends in a cadence.
While this approach would have the advantage of being musically informed, it has a number of drawbacks. In the first place, even what appears to be the straightforward task of parsing music up into cadences or phrases might elicit differing interpretations by different theorists. Disagreement could arise for any number of theoretical reasons. Additionally, the result could end up producing excerpts spanning greatly differing lengths of music, which could be a confounding effect on the perceived affective content of those excerpts. For example, a longer segment of music that evades several cadences could exhibit a wider range of affective content than a shorter segment of music.

Moreover, the exclusion of segment boundaries within excerpts would systematically bias the excerpts – for example, affects related to closure would likely be higher if every excerpt ended with a segment boundary.

One alternate approach would employ excerpts of the same length. A benefit of such an approach would be the ability to control for excerpt length as a potentially confounding variable. Additionally, defining excerpt boundaries according to elapsed time rather than to musical features would not systematically bias the affective ratings by including or excluding certain musical features from the excerpts. Moreover, this approach does not rely on potentially contentious theoretical interpretations of the music in order to decide where the segment boundaries are. Nevertheless, using excerpts of the same length also has its drawbacks. Dividing excerpts by the elapsed time results in musically arbitrary segment boundaries, which may have an effect on ratings of perceived affect. However, this effect would be randomly dispersed among excerpts. It
therefore seems reasonable to use a specific length of seconds as the criterion for dividing up section boundaries.

After the work or passage is divided into discrete excerpts, each excerpt is presented to the participant one at a time, and the participant rates the amount of each tested affect perceived in each excerpt. Participants are permitted to listen to the excerpt as many times as deemed necessary and to introspect as long as deemed necessary to make an accurate judgment of the perceived affective content of that excerpt. For each excerpt, each affective dimension is presented one at a time to avoid overloading the participant with decisions to make. In this way, over the course of the study, a listener is eventually exposed to each excerpt from the longer passage and so provides responses for the entire work.

An additional concern is in what order to present the excerpts of the movement. One approach to choosing an order of excerpt presentation would be for listeners to be exposed to all of the excerpts in their original order. A benefit of this approach is that it maintains the order of the music originally intended by the composer. While something of the deeper structure of the music could be maintained using this approach, the experience of listening in this manner would nevertheless be quite different from that of listening to the movement uninterrupted.

On the other hand, presenting the excerpts in a particular order to every participant could introduce other biases into their responses. For example, listening to short excerpts and rating the perceived affective content is probably not a common experience for participants. There could be a maturation effect in which the participants
become more skilled over time. Conversely, if the experience is boring, participants could provide less reliable responses toward the end of the study. There are many additional potential sources of error that could result in heteroscedasticity or collinearity related to the placement of an excerpt within the study.

Both the strategy of presenting the excerpts in the composed order, or a diachronic presentation, and the strategy of presenting the excerpts in random order, or a mosaic presentation, are valid ways of using the progressive exposure method. The latter provides access to lower-level information of the musical surface, whereas the former includes some elements of higher-level structures (though chopped up into discrete segments). If studies using both methods were conducted, the results from the two studies could be compared. By canceling out the effects from the low-level mosaic presentation from the high-level diachronic presentation, the residual should reflect the effects of the higher-level contextual musical features. While acknowledging the weaknesses of using the mosaic presentation of excerpts in mitigating higher-level musical structures, the decision was made to present the excerpts in random order.

After the data is collected, it can be amalgamated for each affective dimension across the movement. The resulting mosaic portrait can be represented as a line graph of the perceived level of each affective dimension throughout the work or passage. In an analysis of continuous data, it is difficult to determine how the ratings relate directly to the music being heard due to the problem of lag. With the progressive exposure method, however, direct comparisons can be made between the ratings and the corresponding excerpts in the score that elicited those ratings. The interpretation of such
data is much easier, as the music corresponding to each rating can be analyzed using more traditional theoretical methods. The result is an affective narrative of the work in which specific moments in the music can be related to ratings of perceived affective content.

As noted earlier, the progressive exposure method is something of a compromise between the goals and advantages of the retrospective response paradigm and the continuous data collection paradigm. As mentioned above, by presenting the excerpts in discrete chunks, direct comparisons can be made between the structure of the musical features in each excerpt and the ratings of the perceived affective content of that excerpt. The progressive exposure method also shares with the retrospective response paradigm the approach to letting participants listen to the excerpt as many times as desired and introspect as long as deemed necessary. It is therefore expected that the data resulting from the PEM will be clean in a similar way.

However, the mosaic portrait of the selected passage reflects something of the goals of continuous data collection. Although not truly continuous, individual snapshots of small segments of the studied passage reveal differences in the perceived affective content of various segments of the passage. By amalgamating the data, an affective narrative of how each affect changes throughout the passage can be constructed. Also, because the PEM allows participants to rate the amount of perceived affective content of more than one affective dimension per excerpt, the method also allows the investigation of complex mixtures of affect in each excerpt.
Despite these benefits, the progressive exposure method has its own unique set of limitations. In any experimental paradigm, there are trade-offs between ecological validity and experimental control. Listening to a work in randomly-presented discrete excerpts in a laboratory environment while making affective judgments between each excerpt is not a typical listening situation. It seems likely that this artificially-constructed musical experience would differ in important ways from listening to a work continuously in a more natural environment. These differences may limit the generalizability of the data from this study.

Another consequence of listening to a work in short, randomly-presented excerpts is the attenuation of the effect of deep structure on the perception of affective content. One might expect that one consequence of presenting the material in this way would be a refocusing of the mechanisms by which the affective content of the movement is determined. In other words, whereas typical listening by participants that are familiar with the style of this type of music may rely more on top-down processing, understanding the syntax of the music, and discerning deeper relationships in the implications of longer-range musical events, a focus on shorter excerpts may encourage more bottom-up processing on affective ratings. Aspects of the music that may typically go unnoticed in more common listening environments may come to the fore in determining the affective content of a short excerpt. This type of bottom-up processing may increase the effect of low-level musical elements like register, texture, density, articulations, and dissonance on affective ratings and decrease the effect of implications and subsequent realizations, large-scale harmonic structure, Schenkerian lines,
modulation, or contrasts in formal sections of the movement. A possible critique of presenting a movement in small chunks in this manner might argue that the entire experience is less musical and more ephemeral.

**Affective Analysis Using the Progressive Exposure Method**

As described above, the progressive exposure method can be used to perform an empirically-driven surface-level affective analysis of an entire composition. The results from such an affective analysis offer the potential of new insights into the work to be analyzed. Whereas traditional musical analysis has historically not focused on emotional content, but rather the structures, syntax, and form of the music, the progressive exposure method offers the ability to analyze the perceived low-level affective content of a work and connect it to some of the more traditional insights of theoretical analysis.

For this reason, the second movement of Beethoven’s *Pathétique* sonata (No. 8, Op. 13) was chosen as the work to be analyzed using this method. The movement is a well-known and well-analyzed work, and many insights have already been elucidated by careful readings of the piece. In order to illustrate the types of insights that may be added to the literature about a piece using the PEM, this well-known piece was chosen. Additionally, it is hoped that using the PEM will be able to provide converging evidence about well-established interpretations and intuitions of this work.

There are, nevertheless, disadvantages of using a popular work for the progressive exposure method. For example, familiarity could interfere with the ratings
of excerpts. If the movement is autobiographically important to a participant, they might have emotional reactions based on memories rather than on the music heard. Additionally, if listeners know the work well, they may know the context both before and after the excerpts heard, and that contextual knowledge may affect ratings of perceived affective content. Studying a newly-composed work, or a work no one was familiar with, would mitigate these effects.

There are five goals for the following study using the progressive exposure method:

1) To use an eclectic list of affects that are not only relevant to music, but are specific to Beethoven’s *Pathétique* (to be described in detail in the following chapter).

2) To use the progressive exposure method to analyze the perceived surface-level affective content throughout the movement, building a mosaic portrait that tracks changes in ratings of the perceived affective content throughout the movement, and constructing an affective narrative for each affective dimension.

3) To test the inter-rater and intra-rater reliability of the progressive exposure method and to demonstrate a high level of reliability. Additionally, to evaluate the affective dimension scales used and to eliminate scales that are unreliable.

4) To investigate complex mixtures of affect in the movement

5) To examine the relationship between subjective ratings of the perceived surface-level affective content of music and measurable musical features, building a model of perceived affective content and testing the generalizability of that model to other works.
In the analysis of this work, Chapter 3 will describe the difficulties in selecting affective dimensions or categories to be used in performing an affective analysis of any work; it will also describe an initial study that was conducted to determine which affective dimensions are appropriate for an affective analysis of the movement. Chapter 4 will describe the procedure used in conducting the affective analysis study. Chapter 5 will test the measurement scales used in the main study to determine if the progressive exposure method provides a reliable means of collecting data about the perceived affective content of music. Chapter 6 presents the results of the main study, analyzing the affective content of the *Pathétique* sonata, second movement. Finally, Chapter 7 looks in depth at the relationship between the surface musical features of the movement and the ratings provided by participants. A model of perceived surface affective content in music is built, based on the relationship between measurable musical features of excerpts of music and ratings of the perceived affective content. Finally, the chapter describes a follow-up study conducted to test whether the models are generalizable to be able to predict the perceived affective content of excerpts from other piano sonatas by Beethoven.
Chapter 3: An Initial Study to Ascertain Appropriate Affective Dimensions for Use in the Main Study

Introduction

Before using the progressive exposure method to analyze the surface-level affective content of the second movement of Beethoven’s *Pathétique* sonata, appropriate response categories had to be determined. That is, for the purpose of the study, it was important to use affects that would be appropriate to the piece. As mentioned in Chapter 1, there are many different approaches to picking affective terms or dimensions. One approach uses a geometric space to map any emotion onto one or several dimensions. The most common approach, associated with the work of Russell (1980), uses arousal-sleepiness as one of its dimensions and pleasure-misery as a second dimension; however, other models have been used, such as the model of Schimmack and Grob (2000), which uses the dimensions valence, energy, and tension. An alternate approach has been to use affective terms taken from “the basic emotion” literature (Ekman 1972). These terms were derived from emotions recognized cross-culturally through facial expressions and may not be appropriate to emotional expression in music.

Both the basic emotions and dimensional-model approaches appear to prejudge what might be appropriate language for describing the affective aspects of music. An
alternative strategy might pursue a bottom-up approach in which listeners themselves provide affective terms that seem appropriate for some musical experience. This approach was followed by Zentner, et al (2008). In a series of studies, 1,393 participants reduced an initial list of 515 affective terms to a list of terms deemed appropriate for music. The result, termed the *Geneva Emotional Music Scale (GEMS)*, consisted of 66 affective terms. These terms might be considered to be appropriate for the analysis of the perceived affective content of music.

Nevertheless, the *GEMS* model also did not seem appropriate for the purposes of this study. One reason for this is that it is impractical to test 66 different affective terms using the progressive exposure method. In a five-minute work, there would be 100 five-second excerpts. If the rating of each affective dimension takes 10 seconds for each excerpt and there are 66 affective terms, it would take eleven hours to collect data for one work. Additionally, the model using 66 terms was developed to cover a broad range of styles, including such disparate musics as “classical music” and “Latin American” music. Therefore, because the model was built to be generalizable to many types of music, some of these terms are probably not pertinent to any one particular piece. On the other hand, the distillation of terms to form the *GEMS* came from a much larger number, and so some terms that may be pertinent to a given piece may have been eliminated because the effect may have been too small across all of the music tested.

Rather than prejudging which affective terms would provide useful descriptors for an affective analysis of Beethoven’s *Pathétique*, second movement, it seemed appropriate to build a new bottom-up model of eclectic terms that would be relevant for
this piece. Therefore, the movement itself was used as stimulus for the derivation of affective terms for further analysis. A panel of experienced musicians listened to excerpts from the movement and described their perceptions of the affective content of the excerpts. An informal content analysis was performed on these descriptions resulting in a list of fifteen affective dimensions or categories.

Participants

The purpose of this study was to solicit terms that would be musically appropriate for describing the perceived affective content of the second movement of Beethoven’s Pathétique sonata. However, not all listeners are the same. Cultural familiarity with the music heard may play a role in the perceived affective content of that music. Therefore, while it may not be important that participants are familiar with this specific work, it is important that they are familiar with the “language” of the Pathétique. Accordingly, in selecting participants, it was important to select listeners who were very familiar with a wide range of music written in the style of the Pathétique. If a listener is unfamiliar with music within a particular style, it is likely that many of the cues by which emotional meaning is communicated by the composer or performer will be misinterpreted or misunderstood by the listener (Kendall & Carterette, 1990). Another important consideration was that the participants would be experienced and comfortable with thinking and speaking about music; participants who are not accustomed to speaking about music may be unable to provide thoughtful responses or may be terse.
A useful tool for assessing the musicality of participants, which includes both one’s knowledge about music and one’s ability to understand and respond to music, is provided by the Ollen Musical Sophistication Index (OMSI) (Ollen 2006). A participant’s score is calculated from responses to a ten-item questionnaire. The score represents a probability measured in percentage points (times 10) that a music expert would categorize the respondent as “more musically sophisticated” (Ollen 2006). Participants with scores above 500 are considered more musically sophisticated. In order to ensure a high level of musical sophistication, any participant whose OMSI score was less than 850 was therefore excluded.

Five experienced listeners were recruited from Ohio State University’s School of Music. The participants included two graduate students in music theory, two graduate students in composition, and one professor in performance. Their scores on the Ollen Musical Sophistication Index were 999, 993, 988, 959, and 886. Four of the participants were male and one was female.

**Stimuli**

There exists a large number of commercial recordings of this work, raising the question of which recording to select. The expressive content of a performance of classical music might be regarded as a combination of at least two different intentions: (1) the compositional intent of the composer and (2) the interpretation of the performer. Presumably, a listener’s description of the affective content in a sound recording would be influenced by both of these expressive intentions (Kendall & Carterette, 1990).
There exists a wide array of performance interpretations for any given work. Some interpretations might be understated, whereas others might exaggerate or amplify the latent expressive content implied in the score. Other interpretations might intentionally diverge from the latent expressive intentions of the composer – such as playing a passage in a sarcastic or ironic manner. Recall that the purpose of this study is to analyze the affective content of the composition. In one sense, this type of analysis resembles traditional theoretical notions of analysis, in which the nuance of performance, while essential to the work’s musicality, is not the focus of attention. However, the interpretation of the performer is inescapably a component of any recording.

The ratings of participants in this study will therefore be based not only on the composer’s intent, but also on the nuances of the specific performance chosen. While the effect of different performances on affective ratings is a worthwhile topic of study, the purpose of this study is to analyze the affective intent of the composer.

Therefore, while recognizing the dependency on performance, an attempt was made to minimize the influence of individual performance styles. In other words, the focus of the study was on the compositional structures rather than on the effect of idiosyncratic performances. One approach to reducing or eliminating the effect of performance style would be to have a computer manage the performance through the use of MIDI. The result would be a mechanical realization of the work with exact articulations, and precise rhythmic and dynamic content. While this approach would successfully provide listeners with only the musical structures written by the composer,
it seems likely that such a recording would radically impact or diminish the perceived emotional content of the piece. A strong case could be made that the resulting “performance” would be less faithful to the composer’s original intent.

A better approach would be to select a recording that is not extreme in interpretation, but which reflects the interpretive decisions of a large proportion of the performers of this work. The ideal recording would reflect a sort of “mean” from the “population” of performances of the composition in question. This recording would represent the most common interpretations of the composition without biasing the participants by too extreme of interpretations in one dimension or another.

Bruno Repp (1998) used a creative approach in a study to determine the “average” inter-onset intervals between successive notes in the beginning of Chopin’s Etude in E major, op. 10, No. 3. He recorded three performances each for nine experienced piano players on a MIDI keyboard. He was able to arrive at an average of the expressive timing decisions by averaging across the three trials for each participant, and then averaging across the nine performers. What resulted was an average performance of micro-timing deviations in performance.

Unfortunately, the method was not known to exist prior to this study for Beethoven’s Pathétique sonata, second movement. Therefore, in order to find an appropriate recording of the movement, two researchers listened to all of the compact disc recordings of this movement that were available from the Ohio State University music library. There were 23 different recordings of this movement and two researchers listened to each recording. From this large list, they individually compiled a short list of
recordings that they felt best reflected a sort of “mean” of recordings. The short-listed recordings were amalgamated into one list and then both researchers listened to the these recordings together and agreed on the recording that was perceived to be most representative of the average interpretation of the movement. The selected performance by John O’Conor was on the album Breathe®: relaxing piano for lovers. The length of the recording was 4’47”.

As mentioned in Chapter 2, it was decided to use excerpts of a fixed length for the progressive exposure method. An important issue is the duration of the excerpts used in the progressive exposure method. A handful of studies have examined the temporal processing involved in music listening. Gjerdingen and Perrott (2008) and Schellenberg, et al. (1999) studied the duration of musical excerpts required for style identification. Perrott and Gjerdingen likened the task of style identification to “scanning the radio dial” where a listener decides whether to remain at the current station or move to the next station in a very short period of time. In a formal study, they found that listeners begin processing information pertaining to musical style within 250 ms from stimulus onset, and that listeners are 65% accurate in identifying specific genres from a list of possible choices within this timeframe. A study by Plazak and Huron (2011) suggested that listeners are confident with their ability to characterize the mood or affective content of excerpts that are about 1 second in duration. It may be that such judgments reach ceiling at some length of exposure.

In light of this research, sound excerpts on the order of 1-5 seconds might be considered suitable for the progressive exposure method. Shorter time periods would
lead to a practical constraint for the main study. For example, the 4’47” of this work would require 287 one-second excerpts. If, in an experiment, a listener requires 30 seconds to respond to each excerpt, then using one-second excerpts would require more than two hours to characterize all of the excerpts. The duration of such an experiment would be excessive.

After listening to the work in question, it seemed that 2-3 seconds seemed too short, while 6-second excerpts seemed too long. That is, 6-second excerpts were more likely to span passages exhibiting more than one affective character. Accordingly, 5-second excerpts were used in this study. This duration may not be appropriate for other works. For example, works with faster tempos may very well exhibit a faster affective flux.

The O’Conor recording was divided into 56 five-second excerpts. One approach to conducting the study would have participants listen to the entire movement. However, for the purposes of this study, it did not seem necessary for participants to discuss every excerpt. The form of the movement is a Rondo, and so there are many passages that are repeated. One might expect the perceived affective content of such parallel passages to be quite similar. An alternate approach would have participants listen to just part of the piece. A concern to this approach is that there may be a tendency to make judgments relative. For example, moderate dynamics can appear quite loud in a quiet piece. If listeners just heard these moments, the descriptions of perceived affective content might be skewed. Several excerpts from the work that captured representative samples of the affective expression were therefore chosen. Fifteen
Example 3.1. The fifteen excerpts chosen for the initial study. These excerpts were intentionally chosen to be representative of a wide array of affects in the movement.
excerpts spanning a total of 1’15” were chosen from among the 56 excerpts. These excerpts were specifically chosen because they were deemed by the experimenter to be representative of the expressive variety of the movement. These fifteen excerpts are shown in Example 3.1. In order to avoid abrupt onsets and offsets, each excerpt was edited so as to begin with a 500-ms fade-in and end with a 500-ms fade-out.

Procedure

Each participant was tested individually in an Industrial Acoustics Corporation sound attenuation room. The participants listened to the excerpts at a comfortable listening level in open-field conditions rather than with headphones. The experimenter read the following instructions aloud while the participant read along.

INSTRUCTIONS:

The purpose of this experiment is to gather information about music and emotion. At the end of the experiment, I’ll say more about our specific goals.

In this study, you will be listening to brief excerpts from a slow movement from a piano sonata by Beethoven. After each excerpt, I’ll ask you to describe your feelings, and what emotions or moods you think the passage expresses or conveys. I’ll prompt you with a few questions to try to get you to say as much as you can about the music. I will be transcribing your remarks on a lap-top computer, so I might ask you to slow down or repeat what you said. There will be fifteen brief excerpts, and the whole procedure shouldn’t take any more than about 30 minutes.

I may ask you some questions, but the purpose of the questions is simply to get you to talk about what you hear. Ideally, I wouldn’t ask you any questions at all.

You can talk about any aspect of the sound – whatever catches your attention, whatever you think, whatever it reminds you of. Once again, I want you to describe your feelings, and what emotions or moods you think the passage expresses or conveys. My preference is for you to simply talk about what you are feeling or what you think the passage conveys without my prompting.

Do you have any questions about this?

Participants were also told they could listen to each passage as many times as desired.
Participants heard 15 selected five-second excerpts of music, presented in random order, and their task was to talk about the emotional aspects of each excerpt (i.e., open subjective report). In order to encourage the participants to speak at length about their perceptions, the experimenter transcribed their remarks using a laptop computer. After each remark, the experimenter prompted the participant for further observations. That is, there was constant pressure on participants to come up with additional thoughts or observations pertaining to a given five-second excerpt.

One primary goal was to encourage introspection and to solicit a wide range of responses. Therefore, participants did not type their responses directly into a computer, since the work involved in typing might discourage lengthy commentaries or reduce the number of observations. Similarly, participants did not simply record monologues. By having the experimenter present, and persistently prompting the participant, it was assumed that participants would be more likely to think about the excerpt and to continue speaking until they found an apt characterization of the passage. The experimenter typed the responses, sometimes asking the participant to slow down or repeat comments in order to ensure proper transcription.

Previous research suggests that participants are more reliable when asked about perceived rather than induced affect (Francés, 1958/1988; Hampton, 1945; Swanwick, 1973). However, in a survey of music and emotion studies, Vuoskoski and Eerola (in press) suggest that there is often a fair amount of overlap between perceived and induced emotion and that listeners often conflate the two when reporting their experiences. In order to encourage a wide range of responses from listeners, the
instructions were intentionally left open-ended, incorporating both the language of felt and perceived emotions. Synonyms were used for both the concept of affect (“emotion,” “mood,” “feeling”) and for the experience of the affect (“what...the music expresses or conveys,” “whatever catches your attention,” “whatever you think,” “whatever it reminds you of,” “what you are feeling”) to encourage the participants to speak freely about anything related to the experience or perception of affect in the excerpts.

Content Analysis

Each participant spoke at length for most of the excerpts they heard. Five examples of responses are provided in Table 3.1, one from each participant. For the purposes of an informal content analysis, the responses from the participants were parsed into discrete comments. That is, the transcribed monologue for each excerpt was partitioned into complete thoughts, as judged by the researcher. Some complete thoughts spanned multiple sentences whereas some sentences were deemed to contain multiple complete thoughts. In Table 3.1, vertical lines are used to indicate the partitions as decided by the researcher. Each discrete comment was printed on a separate slip of paper for the purpose of carrying out an informal content analysis. In total, there were 592 discrete comments.

The slips of paper were manually sorted according to whatever categories seemed appropriate. The categorization was done twice, by one researcher involved in the project, and by one independent researcher not involved with the project. After sorting, both researchers provided descriptive labels for each of their categories.
<table>
<thead>
<tr>
<th>Participant</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Evokes feeling of resolve in the sense we can be confident about the beauty being expressed.</td>
</tr>
<tr>
<td>2</td>
<td>Bouyant, change of the inner voice.</td>
</tr>
<tr>
<td>3</td>
<td>Just a great sense of settling.</td>
</tr>
<tr>
<td>4</td>
<td>It sounds fuller.</td>
</tr>
<tr>
<td>5</td>
<td>I think, especially the way the ascending melodic scalar figure slowed down and stopped for a second.</td>
</tr>
</tbody>
</table>

Table 3.1. An example of a free response from each participant. Each commentary is in response to the same excerpt, measures 55-56a. The pipes (|) indicate where the comments were broken into discrete segments.

One of the researchers divided the comments into 22 categories, whereas the second researcher assembled the same comments into 32 categories. Recall that the aim of this study was to identify a set of pertinent affective categories or dimensions that can be used in the main study to characterize different musical moments. The aim is to use these analyses to find a common semantic core pertinent to the specific musical work.

One approach would simply amalgamate the two analyses to form 54 categories.
However, there were several categories that appeared to be synonymous in the two lists. The content categories for both analysts are shown in Table 3.2. The labels for the categories for the first researcher are shown in the left-most column, and the categories for the second researcher are shown in the center column.

In the process of reconciling the two content analyses, the principal goals were to identify the shared or common categories, and to reduce the total number of categories into a manageable set, suitable for the main study. Accordingly, in reconciling the two content analyses, four criteria shaped the choice of categories: (1) Some categories appeared to describe structural aspects of the music rather than the affective content. For example, a category labeled “intrinsic musical qualities” was discarded. These categories are marked with an asterisk (*) in Table 3.2. (2) Ten categories between analysts appeared to be similar or synonymous, and so were combined. These categories are marked with italics in Table 3.2. (3) In five cases, both analysts had assembled nearly identical comments for a given category, although the category labels provided by the two analysts were not identical. An example of such a category was “suspense” and “anticipation,” which were simply combined together. These categories are marked in bold in Table 3.2. (4) Finally, some of the categories were small, representing or containing a small number of participant comments. These categories were deemed to not be representative of the affective content of many excerpts, and were simply eliminated. These categories are marked with a dash (–).

Following this process, a final list of 15 affective categories was established. The right-most column of Table 3.2 shows the 15 combined category labels, and the
<table>
<thead>
<tr>
<th>Analyst 1</th>
<th>Analyst 2</th>
<th>Combined</th>
<th>Analyst 1</th>
<th>Analyst 2</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moodiness</td>
<td>Emotional</td>
<td>Emotional/Moody</td>
<td>-Triumphant/</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Emerging-from-difficulty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sincerity</td>
<td>True/Truthful</td>
<td>Sincerity/Truthful</td>
<td>Unsettled/Anxious</td>
<td>Threatening/Unsettling</td>
<td>Unsettled/Anxious</td>
</tr>
<tr>
<td>-In sincerite</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive/Happy</td>
<td>Happy/Joyful</td>
<td>Happy/Joyful</td>
<td>Sad/Tragic/Depressed</td>
<td>Sad/Depressed/Desperate/Depression</td>
<td>Sad/Depressed/Tragic</td>
</tr>
<tr>
<td>(Happy/Joy/Glee/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Celebration/Goodwill)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dark</td>
<td>Dark</td>
<td>Dark</td>
<td>Important/Majestic/Serious</td>
<td>Importance/Graceful/Majestic/Royal</td>
<td>Important/Serious</td>
</tr>
<tr>
<td>-Obnoxious</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calm/Serene</td>
<td>Pastoral/Peaceful/Relieved</td>
<td>Calm/Serene</td>
<td>Anticipation/Building-to-something</td>
<td>Tenseness/Suspense</td>
<td>Suspense/Anticipation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*Accomplishment/Arrival/Closure</td>
<td>-Moving</td>
<td></td>
</tr>
<tr>
<td>Cheeky/Impudent/Sassy</td>
<td>Humorous/Cheeky</td>
<td>Cheeky/Sassy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lonely/Isolated</td>
<td>Isolated/Alone</td>
<td>Lonely</td>
<td>*Intrinsic Musical Qualities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Intimate</td>
<td></td>
<td></td>
<td>-Cartoonish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Buoyant/Bouncy</td>
<td></td>
<td></td>
<td>-Nostalgia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contentment/Settled/Resolved</td>
<td>Contentment/Complacent</td>
<td>Contentment</td>
<td>*Realization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carefree/Lighthearted</td>
<td>Carefree</td>
<td>Carefree</td>
<td>*Uncertainty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighty/Heavy</td>
<td>Grounded/Weighty</td>
<td>Weighty</td>
<td>*Superficial/meaningless</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal struggle/Striving/Yearning</td>
<td>Striving/Yearning</td>
<td>Striving/Yearning</td>
<td>-Unpleasing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Unnerving</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.2.** The results of the content analysis. The far-right column indicates the labels used in the main study. Labels in *italics* were deemed synonymous categories and so were combined. Labels in *bold* contained similar content, though the label chosen was different. A label capturing both affective terms was created for the main study. Labels with asterisks (*) described categories related to musical structures instead of affect, and so were discarded. Labels with a dash (-) had small representations and so were discarded.
researcher-defined categories from which they were derived. Combined categories are horizontally aligned with the final amalgamated label indicated in the “combined” column. These combined categories were then used in the main study.

Discussion

The initial study produced an eclectic list of fifteen discrete affective terms. These terms were distilled from the free responses provided by participants. It is worthwhile to compare the affective terms produced from this study with the three emotion models discussed so far. Of the basic emotions, only “happy” and “sad” appear embedded within the labels “happy/joyful” and “sad/depressed/tragic.” There is no appearance of “anger” or “fear.” Additionally, while many of the derived terms can be placed on the two dimensional model, it is not clear how some labels would fit into the dimensional paradigm. For example, both “important/serious” and “cheeky/sassy” seem to be ambiguous or neutral on both the valence and arousal dimensions while nevertheless seeming to have strong characters and being quite distinct from one another. Finally, most of the affective labels derived from the content analysis are missing from among the 66 terms used in GEMS. In fact, the only terms with direct corollaries are “happy,” “joyful,” “sad,” “calm,” and “serene.” The presence of many affective terms not included in any of these existing models suggests that these models may not be appropriate tools with which to measure the perceived affective content of Beethoven’s Pathétique, second movement.
The method used resulted in a list of fifteen affective dimensions that were given as relevant descriptors of the perceived affective content of the movement, distilled down from the 592 discrete comments given by the participants. Moreover, many of these affective dimensions were not found in other emotion models. Like the GEMS, the list is admittedly a distillation from a large list of affective terms; nevertheless, this experimentally-derived list seems appropriate to the movement to be analyzed.

There are nevertheless some concerns with the terms derived. Though the excerpts were specifically chosen to cover a broad range of sections within the movement, they were chosen from a set of excerpts defined by arbitrary cut-offs. The resulting excerpts were therefore somewhat artificial in their presentation of material, sometimes beginning or ending in musically awkward places. This artifact may have impacted the way that the participants discussed the excerpts. In post-experiment interviews, some of the participants explicitly mentioned the excerpt boundaries as contributing to the way they described the affective content of the excerpts. Some examples of comments that reflect this artifact are: 170. “driving towards a point, but cut off.” 198. “Unsettled stopping in the middle of the phrase.” 505. “It’s like an emotion cut off (it has to do with where the excerpt cuts off).” These sorts of comments seem to have increased the prevalence of terms like “unsettled,” “anxious,” and “cheeky.”

Another related concern has to do with the mechanical noise which can be heard from the piano in some of the excerpts. In post-experiment interviews, some participants reported that the act of focusing on short excerpts highlighted some of the background noises in the recordings that they would not have otherwise noticed. For
example one participant said of the final cadential material, “it seems obvious to me that the performer is being very careful about articulation. Without knowing what came before or what’s coming after, I wonder why he did it that way.” Speaking of the same excerpt, another participant said, “Obnoxious, but of the performance...musical accomplishment and relaxation that should be achieved washed away by obnoxiousness or nasally that I’d want to slap.”

Another interesting consequence of using experienced musicians was a deep familiarity with this specific movement. Of the five participants, four knew it was a Beethoven sonata, and three knew which sonata it was. However, all five reported being familiar with the piece. This familiarity led the participants to sometimes comment not only on the excerpt they heard, but the contextual material immediately preceding and following this excerpt. For the purposes of this study, this type of implicit context was not a problem, as the goal was simply to arrive at a list of affective labels appropriate to the piece.

Another troubling result from the methodology of the current study relates to some of the compound labels attained. Using multiple terms in one label that are roughly synonymous, such as “happy/joyful” or “sad/depressed/tragic” could impact how a listener perceives the affective content of an excerpt. Other compound labels that were chosen because the contents of a category were summarized by the two researchers using different summative terms, such as “unsettled/anxious,” “striving/yearning,” or “emotional/moody,” are more distant in their meaning. Though these labels may accurately portray the content of the free responses, the conflict from these non-
synonyms may present difficulties in the main study. Yet other labels, like “cheeky / sassy” or “emotional/moody” may be unclear to all participants in the main study, or may potentially be defined in very different ways by different participants.

Conclusion

In this study, fifteen affective labels were distilled from 592 comments that arose directly from participants’ experience of excerpts from this piece. This approach was motivated by a desire to build a new bottom-up model to attain affective labels that were appropriate for an affective analysis of this movement. The fifteen affective dimensions were subsequently used in an exploratory study in which every excerpt from the movement was rated by participants along these dimensions. The methodology of this study is described in the following chapter.
Chapter 4: The Procedure for the Main Study

Introduction

In the initial study, participants were exposed to selected five-second excerpts from Beethoven’s *Pathétique* sonata, second movement. Participants were asked to provide comments about the perceived affective content of the excerpts that they heard. In total, there were 592 discrete comments that were elicited from participants. These comments were distilled down to fifteen affective dimensions using content analysis.

In the current study, these fifteen affective dimensions are used to perform a surface-level affective analysis of the same recording of the movement. Participants are asked to judge the perceived affective content of the movement along these dimensions using the progressive exposure method. As mentioned earlier, many researchers have approached the investigation of affect in music by examining only one affect at a time. In this study, fifteen affects are simultaneously investigated. This approach allows an examination of the potential interactions between multiple affects.

By presenting the movement in small, discrete excerpts, the relationship between perceived affect and the musical structures that comprise these short excerpts is more easily discerned. Collecting the data in this way also allows an investigation of how each individual affective dimension changes throughout the movement, by
amalgamating the ratings gathered in the correct time-sequence. This chapter will
describe the procedure for this study and explain the decisions made in formulating the
procedure.

Participants

For this study, the goal was to gather data on how listeners perceive the surface-level
affective content of five-second excerpts from Beethoven’s *Pathétique* sonata, second
movement. For the initial study, expert participants provided spontaneous unguided
musical descriptions of the perceived surface-level affective content of selected excerpts
from the *Pathétique*. It was therefore appropriate to use musicians who were experienced
with listening to and intelligently describing emotional expression in early Romantic
piano music. For this study, affective dimensions were provided *a priori* for the
participants, and they simply rated the perceived affective content of excerpts along
these dimensions.

Judging a stimulus according to predefined categories is usually easier than
having to spontaneously provide the categories yourself. This makes it an inherently
simpler task. It was therefore deemed unnecessary to use musical experts. However, it
may be the case that listeners who are completely unfamiliar with the style and syntax of
early Romantic piano music may misperceive the intended emotional expression of the
music and would therefore differ significantly from the population of listeners familiar
with the music. It was therefore appropriate to recruit participants who had some
degree of musical training and familiarity with early Romantic piano music.
For this study, 110 music majors were recruited to participate. 51 students were recruited from the Ohio State University school of music subject pool. This was one of several experiments that could be selected by participants in order for them to receive course credit for sophomore-level Aural Skills. In addition, 59 students were recruited from the student population of Westminster Choir College. The mean age for participants was 22.1 years (standard deviation = 6.4), and the mean number of years of musical training was 14.3 (sd = 6.2).

Stimuli

The stimuli used in this experiment were five-second excerpts derived from the same recording used in the initial study. Specifically, 56 five-second excerpts were used that spanned the entire duration of the 4’47" John O’Conor recording of the second movement of Beethoven’s Pathétique sonata (No. 8, Op. 13). As in the first study, each excerpt was edited with a 500-ms fade-in and fade-out to avoid abrupt onsets and offsets.

Each excerpt was exactly five seconds in duration, and so the beginning and endpoints for the excerpts were musically arbitrary. There is a danger that some of the participant responses may be artifacts of the position of the excerpt due to editing. In order to minimize this possible artifact, two complete sets of edited excerpts were created that were offset by 2.5 seconds. In short, one set of excerpts spanned the time sequence: 0”-5”, 5”-10”, 10”-15”, etc; a second set of excerpts spanned the time sequence: 2.5”-7.5”, 7.5”-12.5”, 12.5”-17.5”, etc. Half of the listeners heard the first set, and half
heard the second set. Using this method, 112 sound excerpts were created in two sets of 56 each.

Using dovetailed excerpts that overlap in content has the advantage of providing a higher degree of resolution in the data. By way of illustration, consider Example 4.1, spanning 140”-150”. Here, A represents 140”-145”, B represents 145”-150”, and D represents 142.5”-147.5”. If this passage were divided into just excerpts A and B, there would be two resulting points of data – each one representing five seconds of music. However, excerpt D is composed of the last 2.5” of A and the first 2.5” of B. In other words, both excerpts A and D share a core of the same 2.5” of music (y), with A including 2.5” onto the beginning (x) and D including 2.5” onto the end (z). As a result, each 2.5” of music in the piece belongs to two different excerpts. The resulting redundancy allows for direct comparison between A and D to examine how new

![Example 4.1](image)

**Example 4.1.** Measures 37-39, spanning 140”-150” of Beethoven’s *Pathétique* sonata, second movement. Here, A indicates the first five seconds, B the second five seconds, and D the five middle seconds of the excerpt. As seen in the example, y is shared in common between A and D. A also adds x onto the beginning of y, whereas D adds z onto the end of z. This dovetailing results in a 2.5” resolution.
material influences affective ratings. Additionally, C and E also provide data for the excerpt shown, though C contains 2.5” of music before this passage and E contains 2.5” of music after this passage. The result is one rating every 2.5”, or 4 ratings for every 10 seconds of music – three complete excerpts and two half-excerpts.

In this study the effective resolution might be regarded as 2.5”. Specifically, a 2.5” excerpt is considered to be surrounded by 1.25” of musical context both before and after it. In other words, the middle 2.5” of each 5” excerpt is unique from the middle 2.5” of every other excerpt. The surrounding 1.25” can be considered as the musical context for the middle portion of each excerpt, taken from the two surrounding unique 2.5” excerpts. This approach was made explicit to participants, who were told to focus on rating just the middle of each sound excerpt.

**Procedure**

Participants were randomly assigned to one of the two offset-groups; they heard all 56 five-second excerpts comprising the entire 4’47” work from that offset-group, presented in random order. In addition, participants rated five randomly-chosen repeated excerpts, as a means of testing intra-subjective reliability. For each excerpt, the goal was to collect ratings for all fifteen affective scales attained in the initial study. However, it was deemed impractical to ask individual participants to make 15 judgments for each of the 61 musical excerpts. Listeners are apt to forget how the passage sounded by the fifteenth judgment and may become confused about the different assessment categories. Additionally, if each judgment takes 10 seconds, the 915 (15 affects X 61 excerpts)
individual judgments could take over 2.5 hours to complete. Accordingly, judgments were spread across five participants, so that each participant only judged 3 affects for each sound stimulus. In effect, a single task was divided among five participants. The 110 participants were therefore divided into 10 groups of 11 participants each: 2 offset-groups X 5 affect-groups.

At Ohio State University, the participants were tested individually in an Industrial Acoustics Corporation sound attenuation room. Participants listened to the stimuli with free-field speakers adjusted to a comfortable volume. At Westminster Choir College, the study took place in a computer lab with ten workstations. Up to ten participants were tested at a time. Participants listened to the stimuli on headphones adjusted to a comfortable volume. The experimenter read the directions aloud while the participant(s) followed along with the printed instructions.

INSTRUCTIONS:

The purpose of this study is to gather information about music and emotion. At the end of the experiment I’ll say more about our specific goals.

In this study, you will hear a series of 61 5-second excerpts from a slow piano piece. You will be asked to rate each excerpt on 3 emotional scales.

Although the excerpts are 5 second long please concentrate on only evaluating the middle portion of each sound excerpt: the fade-in and fade-out are present only to help you hear the context. So, if there is a difference between the beginning, middle, or end portions, please just rate the middle.

Use the linear scale to indicate your judgements for each of the 3 categories. Please do not evaluate your own emotional response to the music, but simply what you think the music is trying to convey.

Each emotional scale is represented by a line, with one end of the scale representing the maximum amount of that emotion and the other end representing the minimum amount of the emotion.

You can listen to an excerpt as often as you like by clicking on the PLAY button. For each excerpt, each scale will appear one at a time. You must adjust each scale before moving on to the next. Once the scale is adjusted, click NEXT EMOTION and the next
After participants read the instructions, they were given some examples of the difference between perceived and felt emotion. For example, they were told that a case of felt emotion is listening to upbeat music when you are depressed to get into better spirits, whereas a case of perceived emotion is when you listen to what you recognize as sad music without actually becoming sad. They were reminded to rate only the emotions that they perceived the music to be expressing, rather than any emotions that they felt in response to the excerpts.

Participants were also told that it would be very difficult to figure out exactly where the “middle” of each excerpt was, but that with such a short excerpt the emotional content would probably not change much. However, they were told that if they heard something at the beginning or the end of an excerpt that would change their response, they were to cut off that part of the excerpt from their rating.

After reading the instructions out loud and answering any questions the participant had, the experimenter observed while the participant attempted two practice trials. Each practice trial consisted of a five-second excerpt from the slow movement of a Beethoven piano sonata not tested in the study.

For each excerpt tested, one affect was presented at a time to the participant. An example of the interface is shown in Example 4.2. The participant used a mouse to
Example 4.2. The interface used for the main study.

adjust a graphical slider on the computer monitor. To discourage listeners from
gravitating to specific areas on the slider more than others, there were no numbers
printed on the slider. However, a series of tick marks provided a visual approximation
of distance along the slider. For every trial, the slider began by being positioned in the
center of the interface. After every evaluation of an affect for an excerpt, the listener was
presented with a new affective dimension for the same excerpt. Each participant rated
three of the fifteen affects for the duration of the study. The affect groups were chosen to
maximize the variance of the affects along the valence and arousal dimensions. The
affect groups are shown in Table 4.1.
<table>
<thead>
<tr>
<th>Affect group</th>
<th>Affective dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unsettled/Anxious, Contentment, Important/Serious</td>
</tr>
<tr>
<td>2</td>
<td>Weighty, Sincerity/Truthful, Happy/Joyful</td>
</tr>
<tr>
<td>3</td>
<td>Emotional/Moody, Carefree, Striving/Yearning</td>
</tr>
<tr>
<td>4</td>
<td>Lonely, Cheeky/Sassy, Sad/Depressed/Tragic</td>
</tr>
<tr>
<td>5</td>
<td>Suspense/Anticipation, Calm/Serene, Dark</td>
</tr>
</tbody>
</table>

Table 4.1. The five groups of three affective dimensions each. Participants were randomly assigned to one of these three affective dimension groups.

After the study, participants were each asked a number of follow-up questions. Specifically, participants were asked to describe their experience, whether they enjoyed it, how difficult they found it, what strategies they used, what they thought of the emotions they rated, what emotional labels they would suggest as being the most appropriate to this piece, how they interpreted the middle and the left of the scale, whether they knew the piece they listened to, and what they thought the study was about.

Conclusion

This study was designed in order to gather responses from participants about the perceived surface-level affective content of each of the discrete five-second excerpts from Beethoven’s *Pathétique* sonata, second movement. After data collection, reassembling the responses in the correct time-sequence will provide a mosaic analysis of the perceived surface-level affective content of the movement.
This measurement method was newly developed for this study. Whenever a new measurement method is developed, it is appropriate to investigate that method’s accuracy or reliability. The following chapter will address these concerns by conducting an investigation into the intra- and inter-subjective reliability of the data collected. In this way, data from unreliable participants will be eliminated and entire measurement scales that are unreliable will be eliminated.
Chapter 5: Reliability Measures for the Progressive Exposure Method

Introduction

In Chapter 4, the procedure for the main study was outlined. This study employed the fifteen affective dimensions attained from the initial study in a progressive exposure paradigm. Specifically, listeners rated the perceived affective content of 112 excerpts – two sets of 56 excerpts offset by 2.5” – spanning the entire duration of Beethoven’s *Pathétique* sonata, second movement. The amalgamated data provide a mosaic portrait of the entire movement for each of the fifteen dimensions, which can be likened to a surface-level affective analysis of the movement.

A common criticism of affective analysis is that listeners experience music in highly personalized ways, and thus generalization is suspect. This is properly an empirical question: To what degree are experiences of perceived affect in music shared among listeners? This chapter addresses this question by measuring intra- and inter-rater agreement.

Using a progressive exposure method with this unique set of discrete affective terms is a new approach to measuring perceived affect in music. Therefore, before carrying out any data analysis, it is appropriate to identify whether this method provides reliable data. A common strategy for identifying how well a given method
performs is to measure the consistency of ratings attained by that method. Two types of reliability are of interest: (1) the reliability of individual participants, and (2) the reliability of a particular judgment scale. It is common to set exclusion criteria by which data is eliminated if it fails to be sufficiently reliable. The decision was made *a priori* to examine individual reliability and eliminate data sets that were not reliable across an individual’s responses. With the remaining data, the reliability across participants was examined and data sets were eliminated that were unreliable across participants.

**Normalization of ratings**

For the purposes of comparison between subjects and for the later analyses of perceived affect, the data was normalized across subject-scale. There were two reasons for this. First, participants may have used the scales differently from one another. For example, one participant may have only used the middle portion of the rating scale while another participant used the entire scale. Additionally, because the slider begins each trial positioned in the center of the scale, some participants may choose to primarily adjust the slider upwards while others adjust it primarily downwards.

Another reason for using normalized data is that the range of the scale is somewhat musically arbitrary. Specifically, the variance of perceived affect across the movement is likely constricted compared to the variance of perceived affect in all of music, or at least in all of music in the style of early Romantic piano music. For example, even the moments of the highest amounts of striving/yearning for the second movement of the *Pathétique* will probably be lower than the highest moments in the
broader style. Likewise, the moments of the lowest amounts of striving/yearning for
this movement will probably be higher than the lowest moments in the broader style.
The same is likely true of all fifteen affective dimensions. Nevertheless, participants will
likely use the entire scale. For this reason, the upper and lower ends of the scales should
not be considered as the maximum and minimum values of each affect for all of music in
this style. Rather, the endpoints of each scale may tend to map onto the extremes of the
degree of affect perceived within this work. In effect, despite the isolating effect of
presenting excerpts one at a time, there is nevertheless a musical context for
experiencing these excerpts – namely, the rest of the excerpts in the study. Essentially,
the context for one excerpt is the entire set of excerpts, and so it is likely that there is a
comparison happening between excerpts.

Recall that the primary purpose of this study is to trace changes in perceived
surface-level affects across the entire movement, not to unearth some universal value of
perceived affect in isolated excerpts. It is therefore appropriate to identify how the
perceived affect of one excerpt compares to the perceived affect of other excerpts within
this movement. Accordingly, ratings were normalized scores across subject-scale. The
use of normalized scores provides a method to compare ratings from different
individuals who use the scale in different ways. For example, if one participant
primarily uses the top half of the scale, a rating of 50% should be considered a low score,
whereas the same rating of 50% should be considered a high score for someone who
primarily uses the bottom half of the scale. The result of this approach is a score that is
converted into a z-score by measuring the distance in standard deviations away from the
mean of that score. In this case, the proper mean is the average of scores obtained for
one participant along one affective dimension. Different scales may be used in different
ways by the same participant, and the same scale may be used in different ways by
different participants. These normalized z-scores are used for the following calculations
of reliability and for the affective analyses.

1. Intrasubjective reliability

Whenever one collects data from human participants, a concern is whether the
participants are skilled in performing the required task. One piece of evidence that
someone is skilled at a task is that the person is able to perform the task in a similar way
each time. A measure of skill is the consistency of responses across identical stimuli.
This is typically measured with test-retest correlation. Since data was collected for
duplicate stimuli, it is possible to measure test-retest reliability for each participant.

It is possible that a given participant might provide inconsistent responses for
one scale, but provide consistent responses for the other two scales. Normally, test-retest
reliability is used to exclude data from unreliable participants. However, since some
scales may be ambiguous or difficult to define, even the most conscientious and skilled
participant may have difficulty forming consistent judgments for a given scale.
Accordingly, data was eliminated by subject-scale, rather than by subject. Specifically,
an a priori elimination criterion of +0.4 was established for intrasubjective correlation. If
the intrasubjective correlation for a scale between the responses for the test-retest stimuli
was less than +0.4, that subject-scale was eliminated from analysis.
Each participant evaluated the affective content of five randomly-chosen excerpts two separate times. Two columns of data were generated: the first column consisted of the initial evaluations of the perceived affective content of the chosen excerpts; the second column consisted of the evaluations of the perceived affective content of the same excerpts for the later duplicate trial. Correlations were calculated for these two columns of data, providing a measure of the reliability of that affective scale for the given participant. Intrasubjective correlations for all 330 subject-scales are plotted in Figure 5.1.

After plotting the 330 subject-scales from the study, the elimination criterion of a correlation of +0.4 seemed too strict. That is, 100 subject-scales failed to meet the

![Intrasubjective correlations for all 15 affects (n=330)](#)

**Figure 5.1.** The intrasubjective correlations for all three affective dimensions for all 110 participant. The result is 330 subject-scales. The original exclusion criterion of +0.4 seemed too strict. The exclusion criterion was therefore revised downward to +0.25. The 76 subject-scales that failed to meet this criterion were eliminated from further consideration.
criterion, nearly one-third of the data collected. The decision was therefore made *a posteriori* to lower the exclusion criterion to a minimum correlation of +0.25, indicated on the graph with a vertical line. This exclusion criterion eliminated only 76 subject-scales while still retaining relatively consistent data. The 76 subject-scales that failed to meet the criterion were eliminated from further consideration for the analyses below.

Several interesting observations can be made from the distribution of Figure 5.1. First, consider that the most common subject-scale correlation is between +0.95 and +1.00. This suggests that many of the participants have some degree of skill in making the pertinent affective judgments. In addition, the prevalence of high correlations is consistent with the notion that this measurement method provides a reliable means of gathering participant self-report of perceived surface-level affective content in short excerpts.

Second, notice that there are five participants with very large negative correlations. Highly negative correlations indicate that there is a strong relationship in the opposite direction between the evaluations for the initial presentations of the five excerpts and the subsequent repeated presentations. This data could perhaps be the result of chance, given that only five repeated tests were carried out and that there were so many subject-scales. The responses from two participants resulted in correlations near -1.0: Subject-scale 209 had an intrasubjective correlation of -0.98 and subject-scale 232 had an intrasubjective correlation of -0.95. It is possible that these participants got confused about which end of the scale was which or systematically reversed their strategy at some point in the study. In post-experiment interviews, a few participants...
indicated that they changed their judgment strategy mid-way through the experiment. It could be that these participants were very conscientious, but that they altered the definition of the terms or consciously changed the musical parameters for which they were listening.

Also, note that there are a large number of subject-scale correlations that are approximately evenly distributed around zero. Correlations close to zero are obtained when there is no discernible relationship between the first set of ratings and the second. One possible explanation for why these participants had relatively low intrasubjective correlations could be that they were simply not conscientious. If participants quickly entered their responses to finish the study more quickly, the data they provided could be less reliable.

In order to investigate the possibility that participants with scores between -0.5 and +0.5 were less conscientious than participants with high correlations (whether positive or negative), a test was performed on the amount of time that participants took for the study. In total, it was the case that for 71 of the 110 participants, the intrasubjective correlation for at least one affective dimension was between -0.5 and +0.5. The subject-scale correlations were either above +0.5 or below -0.5 for all three of the affective dimensions for the remaining 38 participants. The mean amount of time spent on the study by the participants with correlations between -0.5 and +0.5 was 30'40” (standard deviation of 11’35”). In contrast, the other 38 participants took an average of 31’18” ($sd = 8’06”$). This difference was not significant at $p < 0.05$. This result
is not consistent with the hypothesis that participants who take more time in using these scales provide more accurate results.

Another possible explanation for correlations close to zero could be that some of the affective dimensions are more reliable than others. If some affective dimensions are more heavily represented in the group of subject-scales that are between -0.5 and +0.5, then they might be regarded as less reliable. Of the 71 subject-scales in this category, dark, emotional/moody, and sincerity/truthful were the most highly represented, with nine out of the 22 given subject-scales failing to meet the exclusion criterion.

In order to further investigate the difference in reliability between scales, the individual subject-scale correlations were averaged across affective dimension. This was accomplished by using Fisher’s z-transformation on the correlations so that they could be averaged (Silver & Dunlap, 1987). After averages were taken, they were retransformed back into correlations. The resulting means and standard deviations are shown in Table 5.1. As can be seen, the affective dimension weighty exhibits the highest intrasubjective reliability, with an average subject-scale correlation of +0.814, followed by contentment with an average correlation of +0.810 and striving/yearning with an average correlation of +0.732. Emotional/moody and sincerity/truthful had the lowest average correlations, with +0.448 and +0.362, respectively.

It should be noted that the average subject-scale correlation for cheeky/sassy was unable to be calculated. One subject-scale correlation for this affect was exactly 0, whereas an additional three subject-scale correlations were exactly 1. Correlations are bounded between -1 and +1, and are therefore not additive and cannot be averaged.
<table>
<thead>
<tr>
<th>Affect</th>
<th>Correlation Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighty</td>
<td>0.814 (0.64)</td>
</tr>
<tr>
<td>Contentment</td>
<td>0.810 (0.69)</td>
</tr>
<tr>
<td>Unsettled/Anxious</td>
<td>0.758 (0.67)</td>
</tr>
<tr>
<td>Striving/Yearning</td>
<td>0.732 (0.48)</td>
</tr>
<tr>
<td>Important/Serious</td>
<td>0.729 (0.60)</td>
</tr>
<tr>
<td>Suspense/Anticipation</td>
<td>0.686 (0.77)</td>
</tr>
<tr>
<td>Lonely</td>
<td>0.679 (0.65)</td>
</tr>
<tr>
<td>Dark</td>
<td>0.637 (0.72)</td>
</tr>
<tr>
<td>Carefree</td>
<td>0.632 (0.48)</td>
</tr>
<tr>
<td>Calm/Serene</td>
<td>0.591 (0.62)</td>
</tr>
<tr>
<td>Happy/Joyful</td>
<td>0.586 (0.59)</td>
</tr>
<tr>
<td>Sad/Depressed/Tragic</td>
<td>0.558 (0.76)</td>
</tr>
<tr>
<td>Emotional/Moody</td>
<td>0.448 (0.61)</td>
</tr>
<tr>
<td>Sincerity/Truthful</td>
<td>0.362 (0.50)</td>
</tr>
<tr>
<td>Cheeky/Sassy</td>
<td>cannot be calculated</td>
</tr>
</tbody>
</table>

Table 5.1. The mean intrasubjective correlations for each of the fifteen affective dimensions. Standard deviations are included in parentheses. Averages were calculated by applying Fisher’s z-transformation, averaging, and then retransforming back to correlations. As seen, all but three of the affective dimensions exhibited mean intrasubjective correlations above +0.5. The mean intrasubjective correlation could not be calculated for cheeky/sassy because several subject-scale correlations were +1.0; correlations of +1.0 map onto ∞, and so cannot be averaged.

Fisher’s z-transformation maps correlations onto the space −∞ to ∞, hence enabling correlations transformed in this way to be averaged. However, +1 maps onto ∞; as three of the 22 subject-scales had an intrasubjective correlation of +1, calculating an average was not possible using Fisher’s z-transformation.
Post-experiment interviews revealed that many participants thought that the cheeky/sassy scale was inappropriate for the movement, and therefore rated that affect with the lowest possible rating for almost every excerpt. The participant whose intrasubjective correlation for this scale was 0 rated the perceived cheeky/sassy content for each of the five duplicated excerpts with the lowest possible rating, both in the first exposure and for the repeated trial, except for one excerpt, resulting in a correlation of zero. All three participants whose intrasubjective correlations were +1 rated each excerpt with the lowest possible rating for every excerpt but one. Because all three of these participants rated both the first and second trials for this excerpt at higher than the lowest rating, the correlation was +1. Therefore, even the consistency of the other ratings for cheeky/sassy may be more an artifact of the perceived inappropriateness of the scale for the movement rather than of the reliability of the scale. Due to these considerations, cheeky/sassy was discarded from further examination for the remainder of the study.

2. Intersubjective reliability

Even though two different participants may each be highly internally consistent, they may provide data that is inconsistent with one another. It is therefore appropriate to investigate reliability across participants. A possible objection to investigating intersubjective reliability is that listening to music is an incredibly individual phenomenon. One could expect that differences between individual listeners are a non-reducible component to the phenomenological experience of what it means to engage
with listening to music. In other words, although musical creativity is typically understood in terms of the composer’s creativity or the performer’s creativity, there is also room for creative listening. In the same way that one would not expect composers or performers to behave uniformly, one should not expect that all listeners should behave in an identical fashion. If a listener’s musical experience differs dramatically from that of other listeners, this should not be regarded as evidence of abnormality or deviance in any pejorative sense.

However, it is much more difficult to study idiosyncratic listening experiences. Before addressing the range or variability of listening behaviors, it is appropriate to first attempt to understand those aspects of a listening experience that are broadly shared by a community of listeners.

Accordingly, a goal in this study is to attempt to draw some conclusions that generalize across different listeners. In attempting to form such generalizations, there are many potential confounds. In the first instance, different participants may be having different experiences while listening to the same music. For example, one participant may feel that the level of striving/yearning of a passage is increasing whereas another participant may feel that the level of striving/yearning is decreasing. This kind of divergence may occur even in situations where both participants have the same understanding of the term.

Alternatively, two participants may hold different conceptions of striving/yearning and so respond differently, even though they may have the same listening experience. For example, one participant might interpret the surface-level striving/
yearning of a passage principally according to harmonic criteria such as dissonance or
the prevalence of tendency tones, whereas a second participant may interpret this same
dimension principally according to increased loudness or dynamic levels. If low
correlations are observed between the responses of two participants, it is impossible to
know whether the differences arise due to different interpretations of the rating scale, or
because the listeners are actually experiencing the music differently. If high correlations
are observed between the responses of two participants, this could be purely
coincidental, but this convergence is also consistent with the idea that the participants
are interpreting the rating scale in similar ways and that they are having similar listening
experiences.

The key numerical tool in amalgamating data to form generalizations is the
average or mean. The calculation of an average is justified only when confident that the
values are drawn from the same population. Since not all participants may interpret the
rating scales in similar ways, it is important to establish criteria by which some values
may be excluded from calculating the average. Again, it is possible that a given
participant might provide responses that are highly consistent with other participants,
but only for one of the three scales judged by that participant. Accordingly, a minimum
cross-response correlation of +0.1 for a given subject-scale was established a priori. That
is, for any given subject-scale response set, only those sets whose average correlation
with corresponding sets was greater than +0.1 would be included.

As mentioned above, before calculating intersubjective correlations, subject-
scales with intrasubjective correlations below +0.25 were first eliminated. In the case of
Table 5.2. The intersubjective correlation matrix for the eight participants who judged sad/depressed/tragic and who had intrasubjective correlations above +0.25. Participant 5 has a much lower average intersubjective correlation than the other seven participants.

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.553</td>
<td>0.341</td>
<td>0.474</td>
<td>0.019</td>
<td>0.426</td>
<td>0.634</td>
<td>0.441</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.405</td>
<td>0.408</td>
<td>0.037</td>
<td>0.298</td>
<td>0.556</td>
<td>0.507</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.170</td>
<td>0.187</td>
<td>0.288</td>
<td>0.547</td>
<td>0.306</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.035</td>
<td>0.291</td>
<td>0.435</td>
<td>0.356</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.349</td>
<td>0.118</td>
<td>0.230</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.536</td>
<td>0.590</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.567</td>
<td>0.437</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.426</td>
<td>0.406</td>
<td>0.326</td>
<td>0.316</td>
<td>0.142</td>
<td>0.404</td>
<td>0.497</td>
<td></td>
</tr>
</tbody>
</table>

For the 0-second offset group for sad/depressed/tragic, only eight of the 11 participants in this group had intrasubjective correlations above +0.25. Table 5.2 shows the correlation matrix for these 8 participants who judged sad/depressed/tragic for the 0-second offset stimuli. Similar correlation matrices were calculated for all fifteen affective judgment scales for the listeners who heard the 0-second offset and the 2.5-second offset stimuli. As seen in the table, for each of the eight participants who exhibited intrasubjective correlations above +0.25 for this scale, the average intersubjective correlation surpassed the exclusion criterion of +0.1. However, notice that the mean for Participant 5 is much lower than the other participants. For the case of this participant, the correlations with three of the other participants are below +0.04. For this data set, it is reasonable to suppose that the data for Participant 5 represents responses from a different listening population.
Table 5.3. All of the subject-scales that exhibited a lower intersubjective correlation than +0.25. Only five exhibited lower correlations than the a priori exclusion criterion of +0.1. Thirteen additional subject-scales failed to meet the more stringent criterion of +0.2.

Table 5.3 shows the mean intersubjective correlations for the subject-scales that failed to meet the exclusion criterion. As shown in the left-hand column of Table 5.3, it was the case that the mean intersubjective correlations for only five of the remaining subject-scales fell below the exclusion criterion of +0.1: Participants 2, 6, and 7 from the 2.5-second offset group judging sincerity/truthful, Participant 2 from the 0-second offset group judging emotional/moody and Participant 9 from the 2.5-second offset group judging happy/joyful.

The purpose of investigating intersubjective correlations is to eliminate data sets that may not be from the same listening population as the rest of the data collected. The
Figure 5.2. The mean intersubjective correlations for all 254 subject-scales remaining after the elimination of the 76 subject-scales with an intrasubjective correlation below +0.25. Notice the sharp decline of means below about +0.25. The dotted line at +0.2 indicates the more stringent \textit{a posteriori} exclusion criterion.

five subject-scales described above, with very low correlations with the other corresponding subject-scales, were consequently eliminated. Upon examination of the data, the original exclusion criterion may be too low. For example, in the case above of the 0-second offset group for sad/depressed/tragic (Table 5.2), the mean intersubjective correlation for one subject-scale is clearly lower than the rest, though it remains above the \textit{a priori} cut-off criterion. Consider Figure 5.2, which plots the means of the correlation between each subject-scale and all of its corresponding sets – that is, all of the corresponding sets that were not eliminated because of low intrasubjective reliability. The graph reveals a sharp decline in the number of intersubjective correlation means below +0.2. Because the paradigm is novel, it is difficult to \textit{a priori} set an appropriate
exclusion criterion. A more conservative exclusion criterion, then, would be to eliminate subject-scales with intersubjective correlations lower than +0.2. This *a posteriori* criterion was chosen before any further data analysis was carried out.

The center column of Table 5.3 shows the subject-scales that failed to meet the more stringent criterion of a mean intersubjective correlation of +0.2. Additionally, the right-hand column shows average subject-scale correlations that were lower than +0.25. Subject-scales are identified by an abbreviated affect label, the offset, and the participant number (e.g. Joy; 2.5; 9). This identification is followed by the mean subject-scale correlation for the corresponding scales in **bold**.

Though each affective dimension originally began with 22 subject-scales (11 per offset group), many subject-scales were eliminated. After the elimination of subject-scales due to low intrasubjective reliability and the elimination of additional subject-scales due to the *a posteriori* intersubjective exclusion criterion, the number of subject-scales for each affective dimension was reduced. However, some affective dimensions were left with fewer subject-scales than others. The affective dimensions with the fewest remaining scales were: sincerity/truthful with 4, emotional/moody with 10, dark with 13, sad/depressed/tragic with 14, important/serious with 15, and happy/joyful with 15.

In the case of sincerity/truthful, only 6 subject-scales remained for each offset group after affects with low intrasubjective correlations were eliminated. Eight additional scales were eliminated due to low intersubjective reliability. This small amount of data makes calculating an average suspect, especially when considering that 82% of the data had already been eliminated. As shown in Table 5.3, an additional
subject-scale for this dimension had an intersubjective correlation below +0.2, leaving only three subject-scales, all in the same offset-group. A similar problem (though less marked) existed with emotional/moody. In this case, 12 subject-scales had been eliminated due to low intra- and inter-subjective correlations, leaving only 10, six in one offset-group and four in the other. Again, fifteen subject-scales remained for important/serious after the elimination of subject-scales with low intra- and intersubjective reliability, with eight in one offset-group and seven in the other. However, three additional subject-scales had average intersubjective correlations below +0.25, leaving only twelve, seven in one offset-group and five in the other.

The low levels of reliability evidenced in these three affective dimensions does not necessarily mean that measuring the perceived affective content of music along these dimensions is not possible nor that it is even problematic. It could be that other participants may provide reliable judgments along these dimensions. This could result if participants were more diligent in their ratings. Another possibility could be that the definitions of these terms were simply ill-defined for these participants. A repetition of this study might provide reliable and meaningful data along these affective dimensions. However, for the purposes of this study, the low levels of reliability for sincerity/truthful, emotional/moody, and important/serious were cause for caution in generalizing the data by using a mean. These affective dimensions were therefore eliminated from further consideration.

By way of summary, Table 5.4 shows the mean intersubjective correlations for all 15 affective dimensions. These means were calculated by using Fisher’s z-
<table>
<thead>
<tr>
<th>Affect</th>
<th>Mean intersubjective correlation for all data (SD) [n]</th>
<th>Mean intersubjective correlation for data not excluded (SD) [n]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calm/Serene</td>
<td>0.661 (0.41) [22]</td>
<td>0.500 (0.18) [16]</td>
</tr>
<tr>
<td>Contentment</td>
<td>0.636 (0.39) [22]</td>
<td>0.493 (0.20) [20]</td>
</tr>
<tr>
<td>Unsettled/Anxious</td>
<td>0.641 (0.39) [22]</td>
<td>0.490 (0.19) [18]</td>
</tr>
<tr>
<td>Cheeky/Sassy</td>
<td>0.636 (0.41) [22]</td>
<td>0.476 (0.19) [16]</td>
</tr>
<tr>
<td>Dark</td>
<td>0.623 (0.40) [22]</td>
<td>0.461 (0.21) [16]</td>
</tr>
<tr>
<td>Lonely</td>
<td>0.558 (0.43) [22]</td>
<td>0.426 (0.15) [13]</td>
</tr>
<tr>
<td>Suspension/Anticipation</td>
<td>0.579 (0.31) [22]</td>
<td>0.410 (0.14) [16]</td>
</tr>
<tr>
<td>Happy/Joyful</td>
<td>0.589 (0.40) [22]</td>
<td>0.405 (0.16) [15]</td>
</tr>
<tr>
<td>Sad/Depressed/Tragic</td>
<td>0.562 (0.38) [22]</td>
<td>0.399 (0.19) [14]</td>
</tr>
<tr>
<td>Carefree</td>
<td>0.532 (0.38) [22]</td>
<td>0.389 (0.17) [19]</td>
</tr>
<tr>
<td>Emotional/Moody</td>
<td>0.473 (0.37) [22]</td>
<td>0.374 (0.15) [12]</td>
</tr>
<tr>
<td>Weighty</td>
<td>0.564 (0.33) [22]</td>
<td>0.367 (0.17) [18]</td>
</tr>
<tr>
<td>Striving/Yearning</td>
<td>0.461 (0.42) [22]</td>
<td>0.328 (0.09) [17]</td>
</tr>
<tr>
<td>Important/Serious</td>
<td>0.412 (0.44) [22]</td>
<td>0.274 (0.17) [15]</td>
</tr>
<tr>
<td>Sincerity/Truthful</td>
<td>0.254 (0.38) [22]</td>
<td>0.182 (0.18) [4]</td>
</tr>
</tbody>
</table>

Table 5.4. The mean intersubjective correlations for all of the fifteen affective dimensions. Standard deviations are shown in parentheses and the number of participants in the calculation are shown in brackets. The center column shows intersubjective correlations across all subject-scales for each affective dimension. The right-most column shows the mean intersubjective correlations for all subject-scales not excluded due to low intrasubjective correlation. Due to low intersubjective correlations, and lack of data, sincerity/truthful, important/serious, and emotional/moody were eliminated from further analysis.

transformation, averaging, and then transforming the z-scores back into correlations.

For comparison, the center column shows the intersubjective reliability for the subset of subject-scales not eliminated from the intrasubjective reliability criterion. Notice that in
Figure 5.3. The histograms of all pair-wise intersubjective correlations for participants that judged the same offset for each affective dimension. Some dimensions, such as contentment, weighty, and suspense/anticipation exhibited high reliability while other dimensions, such as sincerity/truthful, important/serious, and striving/yearning exhibited low reliability.

In each case, the average correlation decreases when the data from participants with low intrasubjective correlation is removed. This was not anticipated, and is a puzzling result.
As seen in the table, each affective dimension performed to a different degree of intersubjective reliability. Calm/serene exhibited the highest levels of intersubjective reliability, with a correlation of +0.500, followed by contentment with a correlation of +0.493 and unsettled/anxious with a correlation of +0.490. Important/serious and sincerity/truthful exhibited very low intersubjective correlations of +0.274 and +0.182, respectively.

In order to better visualize the intersubjective reliability of the various affective dimensions, Figure 5.3 displays the pair-wise correlations between all participants that judged comparable offsets for each affective dimension. These pair-wise comparisons are not the subject-scale means for each participant, but rather direct comparisons between each pair of participants. As can be seen in the graph, affective dimensions such as contentment, weighty, and striving/yearning have no pair-wise intersubjective correlations below zero, whereas sincerity/truthful, important/serious, striving/yearning, emotional/moody, and others all exhibit negative pair-wise correlations.

Conclusion

In this chapter, rigorous reliability investigations were conducted on the participant responses that were obtained from using the fifteen affective dimensions from the initial study with the progressive exposure method on Beethoven’s *Pathétique* sonata, second movement. Generally, both intra- and intersubjective reliability were high. Twelve of the fifteen affective dimensions exhibited mean intrasubjective correlations above +0.5. As is typical, mean intersubjective correlations were lower than intrasubjective
correlations. Nevertheless, eight of the fifteen affective dimensions exhibited intersubjective correlations above +0.4 after subject-scales with low intrasubjective correlations were eliminated. These high levels of reliability are consistent with the idea that using the progressive exposure method with these affective dimensions on this movement produces meaningful results that are appropriate to average together with which to conduct an analysis of the affective content of this movement.

Many subject-scales were eliminated from consideration due to low reliability. Of the 330 subject-scales collected, 76 failed to meet the exclusion criterion of +0.25. These, however, were not evenly distributed among the affective dimensions. Dark, emotional/moody, and sincerity/truthful were the most highly represented among the 76 subject-scales eliminated, consistent with the idea that they are less reliable dimensions than the others. Several more subject-scales were eliminated because the mean intersubjective correlations for those scale were below +0.2. Due to these considerations, cheeky/sassy, emotional/moody, sincerity/truthful, and important/sincere were eliminated from further consideration.

While the progressive exposure method certainly was not a method that provided universally reliable data with the fifteen tested affective dimensions applied to this movement, there is still a fairly high level of reliability. The data is not consistent with the idea that every participant is having a unique experience of perceived affective content in the five-second excerpts used in this study. These considerations encourage continuing on to an analysis of the perceived affective content of the movement, by sequentially ordering the data to provide a mosaic portrait of each of the remaining
eleven affective dimensions. This affective analysis is carried out in the following chapter.
Chapter 6: Affective Analysis of the Second Movement of Beethoven’s *Pathétique* Sonata
(No. 8, Op. 13)

Introduction

In Chapter 3, a study was described that resulted in fifteen affective dimensions. These dimensions were used with the progressive exposure method to measure the perceived surface-level affective content of 112 interleaved five-second excerpts from Beethoven’s *Pathétique* sonata, second movement. In Chapter 5, the reliability of each subject-scale was investigated by examining the intrasubjective and intersubjective reliability of each scale. Of the 330 subject-scales collected, 76 were discarded from further consideration due to intrasubjective correlations below +0.25. Of the remaining subject-scales, eighteen additional scales were discarded from further consideration due to an average intersubjective correlation with all corresponding subject-scales below +0.2. Of the remaining subject-scales, there were too few for each of the affective dimensions cheeky/sassy, important/serious, emotional/moody, and sincerity/truthful to justify calculating a mean. These four affective dimensions were eliminated from further consideration, leaving eleven affective dimensions for an affective analysis of the movement. In this chapter, an affective analysis is performed on the movement based on the results from the progressive exposure method.
Results

Affective Analysis

The results of this study provide a mosaic portrait for each of the eleven affective dimensions across the entire movement. The complete results from this study, which plot all eleven affective dimensions across the entire work, are shown in Appendix A. For the purpose of illustration, consider Figure 6.1, a short excerpt spanning measures 34-44 from the movement. Here, six of the eleven affective dimensions are plotted: happy/joyful, contentment, calm/serene, lonely, dark, and unsettled/anxious. On that affective dimension graph, each data point reflects the mean participant rating for that particular five-second excerpt. The data from both offset-groups have been combined into a single graph. Since the stimulus sets were offset by 2.5 seconds, the interleaved data provides data points every 2.5 seconds. Horizontal dotted lines indicate one standard deviation away from the mean. Vertical dotted lines indicate the excerpt presented in the stimulus judged by each participant. The points between the dotted lines represent the normalized judgments for the excerpt between the dotted lines, evaluated by the 0-second offset group. The points on the dotted lines represent the normalized judgments for the excerpt centered on the line (between the two adjacent data points), judged by the 2.5-second offset group. In other words, every excerpt includes a 1.25-second overlap with the preceding and following excerpts, which were judged by the other offset-group.
Figure 6.1. A mosaic affective analysis for ms. 34-44 of Beethoven’s *Pathétique*, second movement. The dotted lines indicate the 5-sec. excerpts rated. Data points between the lines represent ratings for excerpts between the lines. Data points centered on the line represent ratings for excerpts on the lines between the other points.
A few informal observations can be made about Figure 6.1. First, consider the results for the excerpt spanning 132.5”-137.5”. Notice that the excerpt includes the suspension immediately preceding the cadence, but stops before the resolution of the suspension. At this point, unsettled/anxious ratings exhibit a localized peak. The following excerpt, which includes the resolution of the suspension, demonstrates lower unsettled/anxious ratings. The same localized peak occurs simultaneously for weighty, sad/depressed/tragic, and suspense/anticipation, not shown here. For an analysis of all eleven affective dimensions for this excerpt, consult Appendix A. These results are consistent with the notion that musical suspensions communicate an unsettled, suspenseful, weighty, or sad affect.

At 137.5”-142.5”, the music suddenly shifts into the parallel minor, accompanied by a driving and repetitive triplet figure. Here, the ratings for all of the positively-valenced affects graphed – happy/joyful, contentment, and calm/serene – decrease, while dark and lonely ratings increase slightly and unsettled/anxious ratings increase markedly. The three positively-valenced affects graphed remain below 1 standard deviation from the mean for the next several measures while the negatively-valenced affects remain relatively high. This is consistent with common interpretations of minor as negatively valenced in this musical style. One exception is noteworthy: at 142.5”-147.5” happy/joyful ratings show a slight increase while lonely ratings show a strong decrease. These alterations occur simultaneously with the entrance of another voice into the texture. The new voice is marked with staccato articulations and a lower register, serving to open up the register and texture. One possible interpretation of this
result might be that the entrance of another voice at this point contributes to a decrease in perceived loneliness; further testing of the effect of voice numerosity on loneliness ratings is warranted.

Finally, consider the final ten seconds of the excerpt in Figure 6.1. The texture thickens substantially as four-note chords low in the left hand produce a muddy sound. At the same time, the register opens up even more, with the right hand leaping up to an eventual high $e$. While the agitated triplet figure continues, the harmonies outline a series of applied dominant seventh chords, bringing about a noteworthy modulation to the major chromatic submediant key. Over these ten seconds, happy/joyful and contentment ratings rise substantially while lonely and dark ratings decrease. Interestingly, calm/serene ratings, which up to this point in the excerpt had remained highly correlated with happy/joyful ratings, nevertheless remain quite low during this modulation. Additionally, unsettled/anxious ratings remain quite high during this passage. While the end of this passage seems to express a happy/joyful affect, it is mitigated by an unsettled/anxious affect. The data suggest, then, that this excerpt could be construed as expressing happiness, but perhaps a manic, edgy sort of happiness.

Parallel Passages

Another interesting analytical opportunity lies with the potential to investigate parallel refrain passages. In many experiments, there is often a tradeoff between ecological validity and experimental control. Gabrielsson and Lindström (2010) describe several studies that examine the perceived expression of emotion by carefully manipulating
various structural factors without musical context. Specifically, experimenters have manipulated mode (Heinlein 1928; Crowder 1985), rhythm and tempo (Motte-Haber 1968), melodic properties (Gabriel 1978; Kaminska and Woolf 2000), and synthesized tone sequences (Scherer and Oshinsky 1977) to attempt to assess the influence that these musical factors have on perceived affect. While permitting a high level of experimental control, the artificially constructed stimuli used in these studies provide a listening experience that is somewhat impoverished from typical music listening situations.

By contrast, another approach utilizes excerpts from real music and asks listeners to rate the perceived affect of these excerpts. Some approaches (such as Tagg, 2006) ask listeners to write down free responses to real musical excerpts, whereas others (such as Watson, 1942) use a forced-choice paradigm. While this approach has the advantage of simulating a real listening experience much more closely, the stimuli are composed of so many musical parameters that it is difficult to discern which musical factors are contributing most to differences in ratings.

Hevner (1936) attempted to compromise between using artificially constructed music that lacks the sense of real music on the one hand, and using real musical works that deny the researcher the experimental control to vary few musical parameters on the other hand. She did this by rewriting real musical pieces, changing only one musical feature at a time. For example, she rewrote major-mode pieces as minor-mode pieces, rising melodic lines as falling melodic lines, and simple harmonies as complex harmonies. These recompositions, however, often seemed contrived, and changes in one musical feature often interacted with other musical features in complex ways.
Additionally, because Hevner was not the original composer, there may be more substantive changes introduced into the music as a reflection of her compositional style. At the same time, her hypotheses may have influenced her compositional choices, leading to increased demand characteristics.

The second movement of the *Pathétique*, on the other hand, provides a unique opportunity to investigate the effect of parallel passages. As a Rondo movement, the theme returns four times after the initial statement. In the second instance of the theme, the only two changes involve the melody being played an octave higher and a slightly thicker texture as a result of the right hand adding an extra inner voice that parallels the left hand’s inner voice. The fourth statement of the theme mirrors the first in that it uses the original register. However, the accompaniment pattern uses a sixteenth-triplet figuration instead of sixteenth notes, the only change from the first theme. Every other element (harmony, melody, register in the accompaniment, form of the melody, articulation, etc.) of the theme remains the same. The final statement of the theme bears the same relationship to the fourth as the second does to the first, except that the authentic cadence at the end resolves on the strong beat instead of after a suspension figure. The melody is in the upper octave, and the right hand adds an inner voice with the same rhythm as the inner voice in the left hand.

Additionally, the movement offers a sort of control, by utilizing an exact, literal repeat of the opening theme after the first episode. In a real listening environment, repetition is always musically significant; repeated material is heard within the context of all of the preceding material. Heraclitus might say that you cannot step into the same
theme twice. In this study, however, recall that each five-second excerpt is played in random order. The result of random ordering is that participants have little ability to tell whether they are listening to the original opening theme or the later exact repetition. Therefore, the exact restatement of the theme serves well as a control group, and the expectation is that there will no statistically significant difference between ratings for the first statement of the theme and its exact repetition.

The five refrains in the movement provide an opportunity to investigate (1) how higher register and thicker texture influence perceived affect, and (2) how quicker rhythmic values influence perceived affect. In order to investigate the changes in perceived affective content between different statements of the theme, an ANOVA was performed for each affective dimension. In other words, the five statements of the theme were treated as “treatment conditions” to see if there were any effects of “treatment.” To correct for multiple tests, the Sidak-Bonferroni correction was used. However, because the ramifications of committing a Type I error do not pose serious threats, an $\alpha$ value of 0.1 was selected. For every affective dimension that returned significant results, pairwise comparisons were made between each pair of themes.

The ANOVA summary table for weighty is shown in Table 6.1. As seen in the table, there is an effect of “treatment” between the different statements of the refrain ($p < 0.007$, the corrected $p$ value). Of the eleven tested affective dimensions, only suspense/anticipation and striving/yearning did not have a significant effect of “treatment.” All eleven ANOVA summary tables are shown in Appendix C.
Table 6.1. The ANOVA summary table for sad / depressed / tragic. The results indicate that there is an effect of treatment ($p < 0.007$).

<table>
<thead>
<tr>
<th>Sad/Depressed/Tragic</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>4</td>
<td>59.669</td>
<td>14.9172</td>
<td>26.698</td>
<td>&lt; 2.2 e-16*</td>
</tr>
<tr>
<td>Residuals</td>
<td>423</td>
<td>236.346</td>
<td>0.5587</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For the nine affective dimensions that exhibited a treatment effect, pairwise comparisons were made between the first statement of the theme and every other statement of the theme for that affective dimension. Additionally, a pairwise comparison was made between the fourth and fifth statements. Moreover, the expectation is that there should be no significant difference between any of the affective dimensions between the first and third statements of the theme. As expected, none of the affective dimensions exhibited a significant difference between the initial statement of the theme and the “control” condition. By investigating the contrast between theme statements, it can be determined if there is an effect of register and rhythmic values on the perceived affective content. The means and confidence intervals for each of the affective dimensions are plotted in Appendix B. Additionally, tables that list Tukey’s HSD for each of the pairwise comparisons are also provided in Appendix B.

As an example, consider again the sad / depressed / tragic affective dimensions. After finding significant effects of theme through the ANOVA, pairwise comparisons were made between the first statement of the theme and each consecutive statement, as well as between the fourth and fifth statements. The results are summarized in Table 6.2.
Table 6.2. Tukey’s HSD for sad/depressed/tragic, showing significant effects for both the higher register and faster rhythmic values at $p < 0.1$.

<table>
<thead>
<tr>
<th>Sad/Depressed/Tragic</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>0.01289</td>
<td>-0.15190</td>
<td>-0.50706*</td>
<td>-0.97668*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td>-0.46962*</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.2.** The plot of the means and 90% confidence intervals for sad/depressed/tragic ratings according to theme statement. There is no significant difference between the first statement of the theme and the third statement (in effect, the “control”), nor is there a significant difference between the first theme and the second theme, transposed up an octave. However, theme 4 and theme 5 both exhibit significantly lower ratings of sad/depressed/tragic than theme 1. There also appears to be an interaction between register and rhythmic value; there is a significant difference between theme 4 and theme 5.

The means and 90% confidence intervals are plotted in Figure 6.2. The numbers represent averages calculated over all of the excerpts from both offset groups. As can be seen in the figure, there appears to be an interaction between register and rhythmic
values for sad/depressed/tragic ratings. Specifically, there is no significant effect of
register between the first and second statements. However, the fourth statement (which
is cast using sixteenth-note triplets) is significantly less sad/depressed/tragic than the
first statement, as is the fifth statement, which is additionally transposed up an octave.
There is a significant difference in register between the fourth and fifth statements,
however. These results implies that higher-pitched themes would be perceived as less
sad/depressed/tragic only if they were played with faster surface rhythmic values.

The second column of Table 6.3 presents the means and standard deviations for
ratings for all eleven affective dimensions from the first statement, whereas the third
column presents the means and standard deviations for ratings from the third statement
– a literal repetition. As expected, there are no significant differences between the two
statements. The fourth column presents the means and standard deviations for the
ratings of the second statement, played up an octave. Notice that ratings for carefree
and happy/joyful are significantly elevated with the increase in register while ratings for
dark and weighty are significantly lower. This could reflect principles from the
ethological model, in which high-pitched quiet sounds are indicative of friendliness,
whereas low-pitched quiet sounds are indicative of sleepiness (Lorenz, 1982).

The fifth column of Table 6.3 presents the means and standard deviations for
ratings of the eleven affective dimensions for the fourth statement, with a triplet
accompaniment. Notice that ratings for both happy/joyful and unsettled/anxious are
significantly higher than ratings for the first statement. It is curious that ratings for
happy/joyful and unsettled/anxious – two affects that seem incompatible with each
other – are both elevated with the faster surface rhythms. One possible explanation for these results could be that higher levels of activation and arousal are signified with faster rhythms. Conversely, ratings for calm/serene, contentment, lonely, and sad/depressed tragic in the fourth thematic statement are all significantly lower than the corresponding ratings for the first statement. Again, the higher activation levels possibly signaled by the faster rhythms might account for the decrease in ratings.

Finally, the last statement of the theme is both transposed up an octave from the original statement of the theme and arranged using faster rhythms. This theme was rated as significantly higher for carefree, happy/joyful, and unsettled/anxious. Again, the faster rhythms (suggestive of higher arousal levels) might account for this effect. This final statement was also lower in ratings for calm/serene, contentment, dark, lonely, sad/depressed/tragic, and weighty. The fact that these various affective dimensions are all significantly lower for this statement of the theme might reflect both the effects of higher register and faster rhythms. The effect could also partly be influenced by the resolution of the cadence on the downbeat rather than being followed by a suspension.

In order to tease apart the effect of register and rhythm for this last excerpt, a direct comparison was also made with the fourth statement. In the table, significant differences between these two statements are marked with § in the column for theme 5. Ratings of carefree and happy/joyful were significantly higher for the final statement of the theme than for the fourth statement. These same two affective dimensions were also significantly higher in the second statement of the theme compared with the first
<table>
<thead>
<tr>
<th>Affect</th>
<th>Theme 1</th>
<th>Control (Theme 3)</th>
<th>Theme 2</th>
<th>Theme 4</th>
<th>Theme 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calm/Serene</td>
<td>0.630 (0.72)</td>
<td>0.615 (0.67)</td>
<td>0.579 (0.69)</td>
<td>0.311 (0.74)*</td>
<td>0.076 (0.83)*</td>
</tr>
<tr>
<td>Carefree</td>
<td>-0.108 (0.75)</td>
<td>-0.126 (0.78)</td>
<td>0.137 (0.85)*</td>
<td>0.086 (0.68)</td>
<td>0.607 (0.76)*§</td>
</tr>
<tr>
<td>Contentment</td>
<td>0.547 (0.72)</td>
<td>0.393 (0.78)</td>
<td>0.472 (0.63)</td>
<td>0.122 (0.82)*</td>
<td>0.225 (0.69)*</td>
</tr>
<tr>
<td>Dark</td>
<td>-0.012 (0.80)</td>
<td>0.054 (0.77)</td>
<td>-0.472 (0.71)*</td>
<td>-0.093 (0.69)</td>
<td>-0.464 (0.71)*§</td>
</tr>
<tr>
<td>Happy/Joyful</td>
<td>-0.171 (0.85)</td>
<td>-0.173 (0.75)</td>
<td>0.298 (0.82)*</td>
<td>0.159 (0.72)*</td>
<td>0.572 (0.71)*§</td>
</tr>
<tr>
<td>Lonely</td>
<td>0.568 (0.71)</td>
<td>0.618 (0.73)</td>
<td>0.390 (0.69)</td>
<td>-0.114 (0.76)*</td>
<td>-0.412 (0.76)*§</td>
</tr>
<tr>
<td>Sad/Depressed/Tragic</td>
<td>0.511 (0.69)</td>
<td>0.524 (0.71)</td>
<td>0.359 (0.66)</td>
<td>-0.004 (0.81)*</td>
<td>-0.465 (0.86)*§</td>
</tr>
<tr>
<td>Striving/Yearning</td>
<td>0.047 (0.77)</td>
<td>0.064 (0.88)</td>
<td>0.164 (0.80)</td>
<td>-0.084 (0.80)</td>
<td>-0.041 (0.90)</td>
</tr>
<tr>
<td>Suspense/Anticipation</td>
<td>-0.282 (0.93)</td>
<td>-0.224 (0.84)</td>
<td>-0.297 (0.84)</td>
<td>-0.142 (0.79)</td>
<td>-0.359 (0.88)</td>
</tr>
<tr>
<td>Unsettled/Anxious</td>
<td>-0.596 (0.72)</td>
<td>-0.536 (0.79)</td>
<td>-0.417 (0.73)</td>
<td>-0.137 (0.76)*</td>
<td>-0.071 (0.77)*</td>
</tr>
<tr>
<td>Weighty</td>
<td>0.209 (0.79)</td>
<td>0.336 (0.67)</td>
<td>-0.249 (0.77)*</td>
<td>-0.320 (0.77)</td>
<td>-0.475 (0.87)*§</td>
</tr>
</tbody>
</table>

**Table 6.3.** Group means for affective dimensions averaged over the entire movement for the five statements of the theme. ANOVA analyses are presented in full in Appendix C. Mean ratings for each statement that are significantly different from mean ratings for the same affective dimension for the initial statement of the theme are marked with *, at $p = 0.1$. In the right-most column, fifth statement mean ratings that are significantly different from fourth statement mean ratings for the same affective dimension are marked with §, at $p = 0.1$. These pairwise comparisons were calculated using Tukey’s HSD test, with corrections for multiple tests. Notice that throughout the movement, there is a gradual increase in positive and high-arousal affective dimensions and a decrease in negative and low-arousal affective dimensions.

statement, suggesting that higher ratings for these affective dimensions are associated with higher register. Ratings for dark, lonely, sad/depressed/tragic, and weighty were all significantly lower for the final statement of the theme than the fourth statement.
Taken together, the five statements of the theme suggest a gradual unfolding of the affective content of the movement from a more negatively-valenced and lower arousal affect in the beginning, towards a more positively-valenced and higher arousal affect towards the end. One possible affective narrative that could fit the data might suggest that in the midst of the *pathos* that the title suggests, there is a contrasting and growing sense of hope or lightness in the midst of suffering.

“Extreme” ratings

In many empirical studies, researchers test whether a given data set is significantly different from 0. If there is an effect of a treatment condition, then the data will differ from 0. In the case of this study, the “treatment conditions” are the 56 different five-second excerpts played for each participant. In this study, however, an average rating of 0 is not an insignificant finding. Recall that with normalized scores, 0 indicates that the participant rates a given excerpt as being average on some affective dimension. One potential reason that a score of 0 might be observed could be if no agreement existed between participants about the affective content of a given excerpt. For example, it could be that some participants rate a given excerpt as very high in calm/serene because the harmonic rhythm is slow, whereas other participants could rate the same excerpt as very low in calm/serene because there are a lot of chromatic passing tones. If there is little agreement between participants, the data will be noisy and many excerpts will have means close to 0.
However, another possible reason for finding a mean of 0 could be that a lot of participants agree that a given excerpt is actually average in its expression of a given affect. Therefore, an important consideration is not whether the mean significantly differs from 0, but how large the variance is in participant responses.

By way of example, consider again Figure 6.1. The mean rating for dark in the excerpt spanning 132.5”-137.5” is very close to average (−0.12). By contrast, the mean rating for unsettled/anxious for the same excerpt is much further from the mean (+0.48). However, the standard deviation of scores for the former of these two excerpts is only 0.58, whereas the standard deviation for the latter of these excerpts is 1.05. In other words, even though the former excerpt was evaluated as much closer to an average amount of perceived dark than the latter was to an average amount of perceived unsettled/anxious, participants agreed much more conclusively about it.

Nevertheless, a worthwhile investigation is to examine what might be termed “extreme ratings.” For the purposes of this discussion, an “extreme rating” is defined as any rating that is more than one standard deviation away from the mean for that affect, whether positive or negative. For each five-second excerpt, the number of extreme ratings can be tallied for each affect. These tallies can give a loose indication of which affects are most strongly expressed in an excerpt. Moreover, the tallies for all of the affects in an excerpt can be added together. The total number of extreme ratings across all affects for each excerpt can be loosely interpreted as a measure of the general expressive capabilities of each excerpt as compared to the other excerpts.
Leonard Meyer (1956) has hypothesized that the less predictable sections of a piece of music have the greatest potential for eliciting or conveying strong emotion in music. Similarly, David Huron (2006) has outlined the role that unrealized expectation or uncertain continuation plays in the emotional experience of a passage. If these theories are correct, one would expect to see more extreme ratings in sections of the *Pathétique* in which the latent probabilities of continuation in the music are less clear, or where the actual continuation of the music is more surprising.

In Rondo movements, it is often the case that the episodes are the most musically unpredictable and harmonically adventurous. Therefore the number of extreme ratings in differing sections of the movement were investigated *post hoc*. The movement is a five-part Rondo, and so there are three refrain sections and two episode sections. There is an additional coda after the last refrain.

The first refrain is sixteen measures, a total of 60 seconds, in which there were 576 extreme ratings. The second refrain (ms. 29-36) lasts 35 seconds, and had a total of 284 extreme ratings. The final refrain (ms. 51-66) is 57.5 second long, and had a total of 538 extreme ratings. Tallying all sections together, there were a total of 1,398 extreme ratings for 152.5 seconds of music in the refrain portions of the movement.

The first episode (ms. 17-28) is 47.5 seconds long, and had 834 extreme ratings. The second episode (ms. 37-50) is also 47.5 seconds long, with a total of 938 extreme ratings. Finally, the coda (ms. 66-72) is 45 seconds long, with a total of 525 extreme ratings. This was a total of 2,297 extreme ratings in 140 seconds of music in non-refrain portions of the movement.
The average of extreme ratings for the refrain sections was 9.17 extreme ratings per second. Compare this to the episode and coda sections, which had 16.41 extreme ratings per second. These results are consistent with the hypothesis that less musically predictable formal sections – in this case episode and coda sections in Rondo form – evoke stronger affective evaluations by listeners.

By tallying extreme ratings, one might also be able to identify the most expressive five-second excerpts in the movement. The passage with the highest number of extreme ratings per second, measures 20-29, is shown in Figure 6.3. Happy/joyful, carefree, contentment, weighty, dark, and unsettled/anxious are plotted with the music. In this span of music, lasting 35 seconds, there are 517 extreme ratings, averaging 14.77 extreme ratings per second. This passage is especially high in chromatic descending lines and is also the only passage in the movement that has a flourish of grace notes.

Perhaps most illustrative are the five-second excerpts with the highest number of extreme ratings. There are four excerpts that have more extreme ratings than any of the other excerpts, two sets of two adjacent excerpts each. Two of these excerpts are 177.5”-182.5” and 180”-185”, with 63 and 66 extreme ratings, respectively. These two contiguous excerpts, measures 48-49, are shown in Example 6.1. These two measures, immediately preceding the final refrain, serve as preparation for the retransition. Staccato articulations outline an applied fully-diminished seventh chord of the dominant in a very low register. The affective dimensions with the largest proportion of positive extreme ratings for this excerpt are dark with 92% of the 13 participant responses who each rated one of the two excerpts, unsettled/anxious with 83% of 18 responses,
Figure 6.3. Measures 20-29, the passage with the highest concentration of extreme ratings, totaling 517 extreme ratings in 35 seconds, averaging 14.77 extreme ratings per second. This passage marks the retransition to the second refrain section.
Example 6.1. Measures 48-49, the two contiguous five-second excerpts with the highest number of extreme responses with 129 in ten seconds, averaging 12.9 per second. These two measures use the fully-diminished seventh chord of the dominant in a very low register to set up the retransition.

suspense/anticipation with 78% of 18 responses, and weighty with 72% of 18 responses. The affective dimensions with the largest number of negative extreme ratings for this excerpt were calm with 100% of 16 responses, contentment with 95% of 20 responses, happy/joyful with 73% of 15, and carefree with 63% of 19 responses.

The other two excerpts with the highest number of extreme ratings are 100”-105” and 102.5”-107.5”, both with 64 extreme ratings. These two contiguous excerpts are shown in Example 6.2. Interestingly, this excerpt is also a retransition, but marks the retransition to the second refrain. Though the harmony for these two excerpts is a dominant triad, the chord itself is sounded before the beginning of the second excerpt, beginning halfway through measure 23. The only thing heard in this excerpt is a chromatic line, which first descends from the submediant to the dominant with a chromatic passing tone. The dominant is then surrounded by its chromatic neighbors. This highly chromatic passage is without harmony, slow, and fairly low in register. The affective dimensions with the largest number of positive extreme ratings for this excerpt
Example 6.2. Measures 23-24, the second highest concentration of extreme ratings, with 120 in ten seconds, averaging 12 per second. The predominant musical feature of this excerpt is the descending chromatic line in a low register, which surrounds the dominant with chromatic half-steps, implying an augmented-sixth predominant sonority.

are unsettled/anxious with 78% of 18 responses, dark with 77% of 13 responses, suspense/anticipation with 72% of 18 responses, weighty with 72% of 18 responses, and sad/depressed/tragic with 64% of 14 responses. The affective dimensions with the largest number of negative extreme ratings for this excerpt are contentment with 80% of 20 responses, carefree with 74% of 19 responses, happy/joyful with 67% of 15 responses, and calm with 63% of 16 responses.

Excerpts with maximum and minimum ratings for each affect

Another approach to gaining insight into the relationship between the musical excerpts used in this study and the eleven tested affective dimensions is to examine the excerpts that exhibited the highest and lowest mean ratings for each affective dimension. First consider Example 6.3, the final measure of the movement. The measure comprises the last excerpt for each offset-group, and contains nothing but tonic triad and rests. This measure contains the two excerpts that were rated lowest for unsettled/anxious,
Example 6.3. The final measure of the movement, rated with the lowest ratings for unsettled/anxious, suspense/anticipation, and striving/yearning and the highest ratings for calm/serene and contentment. These ratings are all consistent with the typical understanding of the end of movements as the point of strongest repose. At the same time, this measure contains the two excerpts that were rated highest for calm and contentment. These results are consistent with the notion that cadences – especially final cadences – are points of strong repose, where the motion of the piece or movement slows down and ultimately ceases. Additionally, the two measures prior to the last measure (not shown), also contain tonic chords on the downbeats, preceded by a strong dominant-tonic perfect authentic cadence. In similar fashion, these two measures are the excerpts with the lowest ratings for weighty and dark, respectively.

In contrast, consider again Example 6.1, comprising measures 48-49. These two measures contain the adjacent excerpts 117.5″-122.5″ and 120″-125″. Both measures contain exactly the same music – a fully-diminished seventh chord of the dominant simultaneously sounded in agitated triplets and arpeggiated in a very low register. These two excerpts account for the highest ratings of dark and suspense/anticipation and the lowest ratings of calm and contentment.
Next, consider Example 6.4, consisting of measures 17b-19a. These two excerpts are marked by a single, lyrical, melody in the highest register used in the movement accompanied by a simple chordal dominant-tonic motion in $f$ minor in the middle register of the left hand. These measures elicited the highest ratings for sad/depressed/tragic, lonely, and striving/yearning for the whole movement.

Measures 27-28, already discussed in the previous section, are noteworthy for being one of the excerpts responsible for generating the most “extreme” ratings. This passage, a single, relatively low-pitched chromatic line spanning a diminished third prompted the lowest ratings for carefree, and happy/joyful.

Finally, consider Example 6.5, spanning measures 43-44 and two excerpts. The first excerpt, including the strong cadence modulating to the major chromatic submediant on the downbeat of measure 44, prompted the lowest ratings for lonely. The second excerpt, measure 44 firmly in the new key, elicited the lowest ratings for sad/depressed/tragic. Perhaps also contributing to these low ratings are the high melody, the active, dense chords in the left hand, and the loud dynamic. Interestingly, the
Example 6.5. Measures 43-44, exhibiting the lowest ratings for lonely and sad/depressed/tragic. These two measures utilize only major (or major-minor seventh) chords, and modulates to the key of the major lowered submediant. Additionally, the melody, doubled in octaves, is very high with a lot of activity throughout in the left hand, with repeated triplet figures and dense chords.

unstable music immediately preceding the move to the major submediant key, measures 40-41 (not shown here), elicited the highest ratings for unsettled/anxious.

The complete list of excerpts prompting the highest and lowest mean ratings of perceived affective content for each of the affective dimensions is provided in Table 6.4. These excerpts are consistent with intuitive notions of which excerpts might be considered as the most extreme examples (both positive and negative) for each of the eleven affective dimensions tested in the study.

Maximum and minimum affective dimensions

Traditional musical analysis has historically not focused on emotional content, but rather on the structures, syntax, and form of music. When music scholars have commented on emotional content, there has been a tendency to describe the mood or emotion of an entire movement or work using just one or a few words. In effect, this amounts to a synchronic affective characterization.
Table 6.4. For each affective dimension, the measure(s) representing the excerpts with the highest and lowest ratings, respectively.

<table>
<thead>
<tr>
<th>Affective dimension</th>
<th>Excerpt with highest mean rating</th>
<th>Excerpt with lowest mean rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calm</td>
<td>ms. 73</td>
<td>ms. 49</td>
</tr>
<tr>
<td>Carefree</td>
<td>ms. 22b-23</td>
<td>ms. 27b-28a</td>
</tr>
<tr>
<td>Contentment</td>
<td>ms. 73</td>
<td>ms. 49</td>
</tr>
<tr>
<td>Dark</td>
<td>ms. 48</td>
<td>ms. 71b-72a</td>
</tr>
<tr>
<td>Happy/Joyful</td>
<td>ms. 43b-44</td>
<td>ms. 27b-28a</td>
</tr>
<tr>
<td>Lonely</td>
<td>ms. 18-19a</td>
<td>ms. 43-44a</td>
</tr>
<tr>
<td>Sad/Depressed/Tragic</td>
<td>ms. 18-19a</td>
<td>ms. 43b-44</td>
</tr>
<tr>
<td>Striving/Yearning</td>
<td>ms. 17b-18a</td>
<td>ms. 73</td>
</tr>
<tr>
<td>Suspense/Anticipation</td>
<td>ms. 49</td>
<td>ms. 73</td>
</tr>
<tr>
<td>Unsettled/Anxious</td>
<td>ms. 40-41a</td>
<td>ms. 73</td>
</tr>
<tr>
<td>Weighty</td>
<td>ms. 24b-25a</td>
<td>ms. 70b-71a</td>
</tr>
</tbody>
</table>

There are many historical sources from educators, performers, and theorists in which analyses of this movement have been performed. Again, while most of these analyses focus on the musical structures of this movement, a few scholars have discussed the emotional content of the work. For example, A. Forbes Milne (1928) describes the first theme of the second movement as “a quiet, pensive melody,” the first episode as “serene and graceful,” and the move to the parallel minor as “disturbing the atmosphere of serenity with its fretful petulance.” Robert Taub (2002), in a treatise for piano players, describes the theme as “unabashedly lyrical” with “long, singing phrases,” the first episode as having a “questioning character,” and the second episode
as having “introspective conversant qualities.” Elaine Sisman (1994) describes the movement as “transcendentally lovely... – often described as ‘consoling’ or ‘healing’ – set into relief by increasingly agitated episodes.” Bengt Edlund (1997) says that the opening theme is “in perfect repose,” but that the rising gesture in measure 3 “betrays a striving for completion.”

Many of these poetic descriptions bear strong resemblances to some of the terms used in this study. Repose and serene are synonymous with the terms “calm/serene,” agitated and fretful seem similar to “unsettled/anxious,” and striving parallels “striving/yearning.” While the results from the progressive exposure method portray a mosaic portrait of how each of the eleven affective dimensions used in the study change throughout the movement, it is nevertheless possible to determine which affective dimensions have the highest mean rating over the entire movement. The affective dimension that exhibits the highest overall value might be presumed to be closest to the affective synchronic character of the entire movement.

Normalized data will always have a normal distribution centered on 0, with a standard deviation of 1. The mean of the normalized data for each affective dimension is therefore 0. Because the data for each affective dimension were normalized, each affective dimensions has exactly the same mean of 0. For this reason, it was necessary for the purpose of this analysis to revert to using the raw data.

The mean surface-level ratings across the entire movement for each of the eleven affective dimensions not eliminated due to low reliability are shown in Table 6.5. As indicated in the table, the affective dimensions with the highest mean surface-level
ratings were 1) calm, 2) contentment, 3) striving/yearning, and 4) suspense/anticipation. These four affects were followed closely by 5) weighty and 6) sad/depressed/tragic.

Can this movement be both calm and full of striving? While these two terms seem antithetical, the short survey of the analytical literature discussed above certainly reveals the use of both of these terms to describe the affective content of the movement. It may be that there is an alternation between these two affects that are paralleled in the alternation of theme and episode sections. However, part of the pathos of the movement

<table>
<thead>
<tr>
<th>Affect</th>
<th>Mean ratings not normalized (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calm</td>
<td>4.498 (1.59)</td>
</tr>
<tr>
<td>Contentment</td>
<td>4.445 (1.65)</td>
</tr>
<tr>
<td>Striving/Yearning</td>
<td>4.349 (1.51)</td>
</tr>
<tr>
<td>Suspense/Anticipation</td>
<td>4.111 (1.67)</td>
</tr>
<tr>
<td>Weighty</td>
<td>4.092 (1.58)</td>
</tr>
<tr>
<td>Sad/Depressed/Tragic</td>
<td>4.012 (1.57)</td>
</tr>
<tr>
<td>Lonely</td>
<td>3.929 (1.67)</td>
</tr>
<tr>
<td>Happy/Joyful</td>
<td>3.985 (1.31)</td>
</tr>
<tr>
<td>Carefree</td>
<td>3.697 (1.50)</td>
</tr>
<tr>
<td>Unsettled/Anxious</td>
<td>3.603 (1.84)</td>
</tr>
<tr>
<td>Dark</td>
<td>3.287 (1.59)</td>
</tr>
</tbody>
</table>

Table 6.5. The mean rating for each affective dimension averaged over the entire movement. Because normalized scores always have a mean of 0, these mean scores represent the raw data. Notice that calm and contentment are the strongest affective dimensions over the movement, but that striving/yearning and suspense/anticipation are the next highest.
might be achieved through the simultaneous juxtaposition of such contrasting affects. It may be that the combination of positively- and negatively-valenced affects and low- and high-arousal affects is part of what creates “bittersweet” emotional experiences.

Correlations between affective dimensions

A common question about measurement methods for perceived affective content in music is the amount of redundancy in the rating scales used. In an empirical study designed to test different models of induced emotion in participants as they listened to music, Eerola and Vuoskoski (2011) used a principle components analysis to discover redundancies in the rating scales. They tested GEMS, a basic emotion model, and a two-dimensional arousal-valence model, and found that all of these scales could be reduced to two principal dimensions and retain 89.9% of their variance. The two dimensions approximately mapped onto valence and arousal; the valence dimension accounted for much more of the variance (55.4%).

In order to investigate the amount of redundancy in the measurement scales used in this method, a correlation matrix was calculated between each of the affective dimensions used. For each of the 112 excerpts, means were calculated across all participants who judged each affective dimension for that excerpt. The 112 excerpt-means for each affective dimension were then correlated with the 112 excerpt-means for each other affective dimension. The results are reported in Table 6.6. Strong correlations (above +0.75 or below -0.75) are highlighted and in **bold**.
Several observations can be made from the data in Table 6.6. The strong positive correlations are consistent with intuitive notions of how the affective dimensions tested relate to one another. For example, sad/depressed/tragic and lonely are highly positively correlated, approaching tautology with a correlation of +0.926; suspense/anticipation and unsettled/anxious, happy/joyful and carefree, and calm/serene and contentment are also highly positively correlated with one another.

The strong negative correlations are also consistent with intuitive notions of the relationships between the tested affective dimensions. Unsettled/anxious is almost

<table>
<thead>
<tr>
<th></th>
<th>Anx</th>
<th>Calm</th>
<th>Cfree</th>
<th>Content</th>
<th>Dark</th>
<th>Joy</th>
<th>Lone</th>
<th>Sad</th>
<th>Susp</th>
<th>Weight</th>
<th>Yearn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anx</td>
<td>-</td>
<td>-0.849</td>
<td>-0.567</td>
<td>-0.941</td>
<td>0.661</td>
<td>-0.489</td>
<td>0.053</td>
<td>0.134</td>
<td>0.847</td>
<td>0.237</td>
<td>0.512</td>
</tr>
<tr>
<td>Calm</td>
<td>-</td>
<td>0.377</td>
<td>0.795</td>
<td>-0.510</td>
<td>0.270</td>
<td>0.262</td>
<td>0.169</td>
<td>-0.742</td>
<td>-0.196</td>
<td>-0.339</td>
<td></td>
</tr>
<tr>
<td>Cfree</td>
<td>-</td>
<td>0.661</td>
<td>-0.814</td>
<td>0.815</td>
<td>-0.590</td>
<td>-0.654</td>
<td>-0.581</td>
<td>-0.657</td>
<td>-0.640</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>-</td>
<td>-0.773</td>
<td>0.610</td>
<td>-0.185</td>
<td>-0.259</td>
<td>-0.836</td>
<td>-0.288</td>
<td>-0.511</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dark</td>
<td>-</td>
<td>-0.832</td>
<td>0.456</td>
<td>0.513</td>
<td>0.666</td>
<td>0.533</td>
<td>0.553</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joy</td>
<td>-</td>
<td>-0.595</td>
<td>-0.635</td>
<td>-0.500</td>
<td>-0.508</td>
<td>-0.420</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lone</td>
<td>-</td>
<td>0.926</td>
<td>0.195</td>
<td>0.256</td>
<td>0.419</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sad</td>
<td>-</td>
<td>0.279</td>
<td>0.323</td>
<td>0.498</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Susp</td>
<td>-</td>
<td>0.274</td>
<td>0.590</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>-</td>
<td>0.511</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yearn</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 6.6.** The correlation matrix between all pairs of affective dimensions. The strongest relationships are highlighted and shown in **bold.** Many relationships are not surprising, such as the strong negative correlations between unsettled/anxious and both calm/serene and contentment; between dark and each of happy/joyful, carefree, and contentment; and the strong positive correlations between unsettled/anxious and suspense/anticipation; and lonely and sad/depressed/tragic.
diametrically opposed to both contentment and calm/serene. Dark, in a similar way, is strongly contrasted against carefree, contentment, and happy/joyful.

Additionally, there are also some relationships between affective dimensions that are much weaker than one might expect. For example, notice that the correlation between sad/depressed/tragic and happy/joyful is only -0.635. While this is a strong negative correlation, it is much weaker than many of the other correlations. In post-experiment interviews, some participants mentioned that some of their ratings were counter-intuitive. Specifically, some participants said that they rated highly both of what seemed to be contradictory affective dimensions. The weaker-than-expected negative correlation between these two affective dimensions recalls the discussion of the highest rated synchronic affective dimensions. This observation is consistent with the movement simultaneously expressing negatively- and positively-valenced affects.

Consider again the ethological model, in which generalizations are made about the communication of affect in the animal kingdom. If music portrays affect in a similar way that the animal kingdom does, one prediction of this model would be that low-pitched, quiet sounds in music would be evocative of both sleepiness (or low arousal states) and also negatively-valenced, low-arousal affects like sadness and loneliness (Lorenz, 1982). If this were the case, one would expect listeners to more easily confuse the low arousal calm/serene with sad/depressed/tragic and lonely. This in turn would result in relatively high correlations between these affective dimensions. However, contrary to these predictions, calm/serene ratings are only correlated +0.169 with sad/depressed/tragic ratings and +0.262 with lonely ratings. While these correlations are
skewed in the expected directions, calm/serene ratings only account for 2.9% of the variance of sad/depressed/tragic ratings and only 6.9% of lonely ratings. These unexpected results could be an artifact of positively-valenced connotations of calm/serene contrasting with the negatively-valenced connotations of sad/depressed/tragic and lonely.

Discussion

The progressive exposure method was developed to provide a sort of compromise between the purposes and methods of both the retrospective response to excerpts paradigm and the continuous data collection paradigm. The data collected from participants using the progressive exposure method shares the strengths of data collected in experiments that test discrete excerpts; namely, the data exhibits a high degree of reliability and allows direct comparisons between the affective ratings and measurable elements of the music. For example, various portions of the score were analyzed in terms of the affective dimensions tested, portions of the movement that exhibited a high degree of strong affective response were identified, and the different statements of the theme were directly contrasted.

On the other hand, by amalgamating the data in the correct order of the movement, a mosaic portrait of each affective dimension was created. This was portrayed as a line graph tracing the normalized scores. While not truly continuous data, the data provided by the progressive exposure method simulates continuous data,
in that the way each affective dimension changes over the movement can be tracked along with the development of the music.

The strengths of the method notwithstanding, there are also drawbacks to the progressive exposure method. Despite the 1.25" of “context” around each data point representing 2.5" of music, the experience of listening to five-second excerpts is greatly impoverished compared to listening to the complete movement in one hearing. As mentioned in Chapter 2, the presentation of short excerpts necessarily mitigates the tendency to use top-down ways of processing the perceived affective content of music. Presenting the passage in short excerpts also tends to increase the effect of short-term, more bottom-up musical features on the perceived affective content. This type of bottom-up processing may increase the effect of low-level musical elements like register, texture, density, articulations, and dissonance on affective ratings and decrease the effect of implications and subsequent realizations, large-scale harmonic structure, Schenkerian lines, modulation, or contrasts in formal sections of the movement. A possible criticism of presenting a movement in small chunks in this manner might argue that the entire experience is less musical and more ephemeral.

Nevertheless, despite the discrete nature of the presentation of each excerpt, the experience of the entire study does present a sort of “context” for each excerpt. Specifically, the entire collection of excerpts creates a backdrop against which each excerpt can be contrasted. This creates its own unique set of strengths and weaknesses. In post-experiment interviews, participants who were familiar with the movement sometimes mentioned that they had trouble restricting their rating to just what they
heard and not the bigger context surrounding the excerpt. Additionally, two excerpts that would not normally be juxtaposed in listening to the movement might be randomly selected to be played consecutively. The affective ratings for one excerpt may be impacted if the excerpt heard contrasted strongly in some way with the excerpt immediately preceding it. However, the effect of randomizing the order of presentation should reduce the effect when all of the ratings are averaged together.

It should also be noted that this study was conducted on a piece of music that is very well known. Evidence for this includes the fact that the recording used belonged to a compilation intended for a general audience (*Relaxing Piano for Lovers*). Additionally, this movement is part of the sophomore-level music theory and music history sequence at Ohio State in both the Romantic harmony quarter and the Form and Analysis quarter. As a result it is very likely that the members of the participant pool at Ohio State, composed of sophomore music majors, are familiar with the movement. Moreover, studies have shown that music may become associated with autobiographically salient events and may therefore come to have deep emotional significance that may not be related to the musical content (Janata, *et al.* 2007; Paul 2010; Barrett, *et al.* 2010). It could be that the perceived affective content of this movement may be different for someone who is deeply familiar with it than for someone who is not. While conducting an affective analysis on an frequently analyzed movement may nevertheless be useful for discovering new insights into the movement, the dangers of a familiarity effect must be kept in mind when interpreting the results.
Another potential issue with this paradigm is the actual nature of the interface used in the study. Participants used a continuous scale in which the slider began in the middle position. Participants were encouraged to make use of the entire scale; if the slider had started at one end, it seems likely that the scores would have clustered toward that end. However, the decision to initiate the slider in the center position seems to have created some problems with the way participants conceptualized the scale. In post-experiment interviews, participants were asked how they interpreted the center (and by extension, the left end) of the scale. Though the ends of the scale were labeled “very [affect]” and “not at all [affect],” some participants conceptualized the center of the scale as a sort of “0” on a number line. For these participants, the center meant that the excerpt was neutral for that affective dimension or that the affective dimension did not apply to the excerpt. These participants tended to think of the extreme left side of the scale as a strong instance of the opposite affect.

Another group of participants tended to think of the middle of the scale as an average amount for that affective dimension. For these participants, the left end of the scale represented a complete lack of that affect, or that the affective dimension did not apply to the excerpt. An example of the different conceptualizations of the scale could be likened to the difference between an affect that is unhappy, and one that is not happy. While the disagreement between these two conceptualizations of the scale may seem like a major problem, it is not clear how these two different conceptualizations of the scale actually affected participants’ ratings. The line between absolutely none of an affect and the opposite of an affect may not be well-defined and may amount to the same treatment
of the scale, at least in terms of how participants rated the excerpts. Moreover, recall that the values were all normalized so that effects of unique treatments of the scale by different participants was reduced.

A more serious issue regarding the slider concerns the tendency for the data to be bimodally distributed. As an example, consider the left plot for Figure 6.4. This is a histogram of all of the raw scores for happy/joyful, divided into 120 bins. It seems that there are three different kinds of responses for this affective dimension. There are those

![Histogram of Happy/Joyful](image)

**Figure 6.4.** The distributions of ratings for happy/joyful. The left graph shows the raw scores and the right graph shows the normalized scores. Notice that the most common rating in the left graph is exactly the middle, indicative of participants not touching the slider. Ratings that were greater than or less than the center this have their own localized means about 1 point from the center. Almost no scores are immediately next to the center. This artifact can still be noticed with the normalized scores, but the effect is reduced. The full set of score distributions are shown in Appendix C.
responses that, for whatever reason, leave the slider at the center of the scale, observable by a big spike in the center. There are responses that move the slider upward from the center. Finally, there are responses that move the slider downward. Notice that, aside from those responses that left the slider where it began, there are very few responses toward the center of the scale. Each of the other two distributions within the plot seem to exhibit a localized mode near about 1/3 of the distance toward the extremity of the slider. These results can be seen for every affective dimension, though the effect is more pronounced for happy/joyful (and balanced between the positive and negative distributions) than many of the other affective dimensions. The effect is mitigated substantially when the normalized scores are plotted instead (the left plot of Figure 6.2). The full results displaying the histograms for all affective dimensions are provided in Appendix D. One possible explanation for this pattern in the data could be that participants felt obliged to move the slider one direction or another, and so ratings tended to cluster away from the middle of the scale.

As mentioned previously, some of the affective dimensions used were compound terms (e.g. happy/joyful, striving/yearning, sad/depressed/tragic). While these compound terms were directly obtained from a content analysis of the free responses given in the initial study, some participants reported confusion about some of them. Specifically, emotional/moody, sad/depressed/tragic, and striving/yearning were all mentioned as terms whose components were sometimes deemed as not completely synonymous. While the compound terms thus provided a means of capturing more fully the nuances of the free response comments from the initial study, they may also
have served to increase noise in the data. It may be for this reason that emotional/moody, sincerity/truthful, or important/serious exhibited reduced reliability and so were discarded. Future studies may benefit from choosing simple affective dimensions that are more easily defined.

Conclusion

In this study, the PEM data was used to conduct a listener-generated analysis of the perceived surface-level affective content of Beethoven’s *Pathétique* sonata, second movement. Specifically, participants rated the perceived surface-level affective content of five-second excerpts along fifteen affective dimensions obtained from the initial study. Participants were able to introspect as long as desired for each excerpt, and the ratings for the perceived affective content collected in this fashion produced data that exhibited a high degree of reliability for eleven of the fifteen affective dimensions.

In order to conduct the mosaic affective analysis of the movement, the 112 participant-rated excerpts were assembled in order, producing a line-graph of each of the remaining eleven affective dimensions over the course of the entire movement. The resulting affective analysis of the movement was compared to the musical surface of the movement. The result was an empirically-informed analysis of how the musical surface conveys affect.

The movement also offers a unique opportunity to compare differences between parallel passages. Specifically, the only change between the first statement of the theme
and the second is an octave displacement and a thickening of the texture. The third statement is a literal restatement of the opening theme. The final two statements are modified from the original statement by utilizing the faster rhythms of sixteenth-triplet notes and (in the case of the final statement) the higher register and strong metric close. An analysis of the five thematic statements reveals a gradual change in affect that correspond to an increase in valence and arousal throughout the movement.

Additionally, it was the case that the episode and coda sections of the movement exhibited greater numbers of “extreme” ratings above 1 standard deviation from the mean. This result is consistent with the notion, initially presented by Meyer (1956), that musical passages that progress less predictably (represented in Rondo form by non-refrain passages) tend to have the potential to evoke or portray stronger emotions.

It was also found that the affective dimensions with the highest mean ratings for the movement were calm, contentment, striving/yearning, and suspense/anticipation. While simultaneously high levels for these four affective dimensions seems counterintuitive, this finding is consistent with historical analyses of this movement that describe it as portraying a complex mix of positive and negative affects. This finding is further corroborated by the relatively low negative correlation between happy/joyful and sad/depressed/tragic ratings. Finally, correlations between the eleven tested affective dimensions were carried out to investigate possible redundancies in the scales used.

These considerations raise important questions about the capacity of music to portray affects, and how strong of a role the music itself plays in listener-generated
ratings of perceived affective content. An additional investigation would be appropriate that examines which measurable musical features (such as mode, presence of dissonance, pitch height, surface rhythmic speed, etc.) are correlated with ratings along different affective dimensions. In order to investigate these questions, the following chapter describes a follow-up study that was conducted to build and test a mathematical model of perceived surface-level affective content for not only this movement, but generalized to all of Beethoven’s piano sonatas.
Chapter 7: Building and Testing a Model of Perceived Surface Affective Content for the
Entire Corpus of Beethoven Piano Sonatas

Introduction

In the main study, described in Chapters 4-6, an affective analysis was carried out on one
movement of a Beethoven piano sonata using the progressive exposure method. In that
study, 110 participants listened to the entire second movement of Beethoven’s *Pathétique*
sonata in randomly-ordered discrete five-second excerpts and rated the perceived
surface-level affective content of each excerpt on fifteen affective dimensions. The result
was a mosaic portrait of each affective dimension as it changes throughout the work,
provided in Appendix A. Moreover, direct comparisons were made between the mean
ratings across participants for each excerpt and the musical content of each excerpt. This
data allows an affective narrative to be constructed as a type of analysis that relates what
is happening musically with ratings of perceived affect. Additionally, direct
comparisons were made between the different statements of the theme, and there was a
significant difference in many of the affective dimensions due to theme. Because the
only thing that changed between different statements of the theme was 1) the rhythmic
values and 2) octave displacement of the melody, there appeared to be a direct effect of
rhythmic values and register on affective content.
These results are suggestive of two deeper questions. The first question is to what extent are some of the other measurable musical parameters in each excerpt (e.g., mode, harmonic rhythm, pitch height, presence of dissonance, presence of cadence, dynamics etc.) contributing to or influencing ratings of perceived affective content in this movement? The second question is how generalizable to other works is the relationship between the perceived affective ratings of the excerpts in the *Pathétique* and the musical content of those excerpts that led to an affective narrative? Both of these questions can be summarized by asking how stable the relationship is between the musical surface of excerpts and the perceived affective content of those excerpts both within the movement tested and across other pieces of music.

As noted in Chapter 6, the progressive exposure method necessarily limits the influence of more top-down types of processing in determining the perceived affective content of a short excerpt of music. As a result, the effect on affective ratings for deeper long-term processes like implications and realizations, large-scale harmonic structure, Schenkerian lines, modulation and key areas, or contrasts in sections, would be consequently mitigated. It seems likely that, instead of these more top-down types of features, listeners would evaluate the perceived affective content of short excerpts based on more bottom-up types of musical features, such as register, texture, density, articulations, and dissonance, for example. The data from the dependent sample test between different statements of the theme, discussed in Chapter 6, are consistent with the idea that low-level features such as rhythmic activity or registral displacement have an effect on the perceived affective content of short musical segments.
It therefore seems appropriate, in building a model of perceived affective content in short excerpts of music, to take into account the types of bottom-up features that a listener may use to make judgments of the perceived affective content of the excerpts. For this study, several low-level analytical observations will be made for each of the excerpts used in the initial study. These observations will be treated as predictor variables in multiple regression analyses for each of the affective dimensions not eliminated due to low reliability. The result is a linear model for each affective dimension that predicts listener ratings of perceived affective content in five-second excerpts in Beethoven’s *Pathétique* sonata, second movement.

After the construction of the linear model, a formal study will be described in which the purpose is to investigate the extent to which the model is generalizable to other works. In this follow-up study, fifteen excerpts sampled from the corpus of Beethoven piano sonata movements are analyzed according to the same low-level musical parameters. These parameters are fed into the linear models in order to make predictions of listener ratings of perceived affective content for each affective dimension.

Constructing a model of perceived affective content in the *Pathétique*

**Predictor variables**

In constructing any linear model, the goal is to be able to account for as much variability as possible in the independent variable by using a set of predictor variables. In building a model of the data from the main study, the independent data to be predicted are the responses given by the participants that rated the perceived surface-level affective
content of each excerpt. The linear model constructed will attempt to explain as much of
this variance as possible using low-level musical features as predictors.

When dealing with human participants, accounting for the variability in
responses is a difficult task. The amount of variability in responses is likely to be further
increased because the main study asks participants to perform an inherently aesthetic
task, judging the affective content of music. There are many potential reasons why
participants could offer differing judgments of the affective content of the same excerpt.
The mood that participants are in when they participate in the study, their general
disposition, or differences in their personalities might affect the way they perceive
affective content in music. Studies have shown that people who score high on the
neuroticism trait are more likely to rate music as sad, for example (such as Schellenberg
& Ladinig, 2011). A participant’s past experiences may also affect their ratings of the
perceived affective content of musical excerpts. Participants that spend more time
listening to classical music might respond differently than participants who do not.
Even if someone listens to classical music, responses may differ depending on a
participant’s level of familiarity with early Romantic piano music. Furthermore,
someone who is familiar with this particular movement may respond differently than
someone who is unfamiliar with it, especially if there are autobiographically salient
memories associated with it. A participant who is familiar with the work may allow
knowledge of the surrounding context of the excerpt influence their ratings.
Additionally, not all participants may be paying the same degree of attention to the
music, or their preferences (or dislikes) for the music they listen to may impact their
ratings. Again, participants may enjoy being in studies or dislike the experience and want to penalize the study because they were forced to participate. The level of conscientiousness of a participant generally, what time of day they participate, when in the term they participate, and how much sleep they have had all may effect their responses. Piano players may engage with piano music differently than performers of other instruments, and participants of differing ages or genders may respond differently. As mentioned previously, people may define the tested affective terms in different ways or may listen for different components of the sound (e.g., harmonic content vs. dynamic content) to make their judgments. There are many more sources of variability in responses that could be considered, as well. In fact, when considering all of the possible sources of variance, it is a wonder that any patterns can be detected in participant responses at all! Though these possible sources of variance could be considered and measured as predictor variables under the right circumstances, they will all be considered as part of the error term for the purposes of this study.

Nevertheless, in spite of all of these possible sources of variance, another reason participant responses may vary could be due to differences in the musical excerpts presented to them. Some suggestion of this was seen in the previous chapter by the systematic way in which responses varied due to different statements of the theme arranged in different ways. For this study, only the differences in low-level musical features were considered between excerpts as explanatory variables to predict participant responses. All of the other sources of variance, though certainly worthy of
There are a number of musical parameters that might influence a listener’s rating of the perceived affective content of an excerpt of music. For this study, sixteen different features of an excerpt were identified that may effect listener ratings, shown in Table 7.1. Most of these musical parameters were calculated by listening to each excerpt, identifying the corresponding boundaries of the excerpt on the score, and visually inspecting the notated segment of the music. This method was sufficient, for example, for counting the number of concurrently attacked notes, determining the direction of the melodic line, or determining whether there were more major or minor harmonies. However, for the two musical parameters that describe how loud the music sounds, the decision was made to analyze the recording of the excerpt rather than the score. This decision was motivated by the common practice of interpreting the relatively sparse dynamic information that Beethoven provides with additional dynamic nuance. Listeners did not have access to the score, and so it seemed more relevant to make this judgment from the recording than from the score. Likewise, performers often regularly introduce rubato into their performance, and so the recording provides a more relevant source from which to make judgments about the tempo.

*Regression Diagnostics*

Before performing the linear regression, it is appropriate to examine the data to determine if there are any violations of the assumptions required to perform multiple
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Operationalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>dyn</td>
<td>1-8, mapping onto ppp-fff; judged from the recording</td>
</tr>
<tr>
<td>cres</td>
<td>dramatic crescendo (2), cresc. (1), no change (0) diminuendo (-1), dramatic dim. (-2); judged from the recording</td>
</tr>
<tr>
<td>dense</td>
<td>maximum number of concurrently attacked notes within the segment</td>
</tr>
<tr>
<td>speed</td>
<td>1-4, number of unique onset moments: 1 = 0-5 onsets, 2 = 6-10, 3 = 11-15, 4 &gt; 15</td>
</tr>
<tr>
<td>tempo</td>
<td>measured from 1-5, with 3 = “moderato,” 1 = “largo,” and 5 = “allegro molto”; judged from the recording</td>
</tr>
<tr>
<td>artic</td>
<td>1 = legato, 3 = staccato, and 2 = neutral. For mixtures, code the most prominent</td>
</tr>
<tr>
<td>direct</td>
<td>1 = predominantly ascending melodically, -1 = descending, 0 = essentially static</td>
</tr>
<tr>
<td>surp</td>
<td>-1 = only diatonic notes, and all intervals either step-wise or arpeggiated; 1 = non-arpeggiated leap or non-diatonic leap; 0 = diatonic arpeggations, if chromatic tones, they are step-wise are resolve to the note they tonicize</td>
</tr>
<tr>
<td>tend</td>
<td>1 = presence of non-incidental tendency tone, 0 = no non-incidental tendency</td>
</tr>
<tr>
<td>mode</td>
<td>1 = predominantly major harmonies, -1 = minor harm. or suggestive of minor</td>
</tr>
<tr>
<td>dis</td>
<td>0 = only major or minor triadic harmonies; 1 = dominant seventh sonority; 2 = any other, more dissonant chord</td>
</tr>
<tr>
<td>harmS</td>
<td>0 = all harmonic progressions using circle of fifths or thirds, or IV-V; -1 = other</td>
</tr>
<tr>
<td>Htemp</td>
<td>number of harmonies within the segment</td>
</tr>
<tr>
<td>hpi</td>
<td>pitch height of highest pitch in semitones away from 0 as middle c</td>
</tr>
<tr>
<td>lpi</td>
<td>pitch height of lowest pitch in semitones away from 0 as middle c</td>
</tr>
<tr>
<td>close</td>
<td>presence of cadence; 2 = authentic cadence, 1 = other cadence, 0 = no cadence</td>
</tr>
</tbody>
</table>

Table 7.1. The musical features used as predictors for the regression analyses. Most of the variables were categorical, and for the purpose of generalizability to other works, even the two continuous variables (speed and Htemp) were discretized into categories. None of the variance inflation factors was found to be above 2.5, suggesting that regression. First, a test was carried out to check for multicollinearity within the 16 predictors used in the study. Each of the 16 predictor variables was tested against the multicollinearity is not a problem for this analysis. Plotting the residuals of each of these tests revealed no nonlinearity or heteroscedasticity among the predictor variables,
suggesting that the modeling errors are uncorrelated. Unfortunately, the presence of so
many categorical variables restricted the value of the visual inspection of the residuals.

The next step was to test for normality in the distributions of ratings of perceived
surface-level affective content for the eleven affective dimensions tested in the study.
Histograms of both the raw data and data normalized by subject-scale for each of the
eleven affective dimensions are provided in Appendix D. As mentioned in Chapter 6,
each affective dimension exhibited non-normality as an artifact of the user interface.
Specifically, the raw data for each affective dimension exhibited a spike in raw ratings at
the center of the scale and at each of the extreme ends of the scale. Even when the data
was transformed by normalizing ratings by subject-scale, the bimodality of each
distribution was apparent.

In order to check whether distributions of the normalized data exhibited enough
of normal distribution to perform the regression, a Kolmogorov-Smirnov test was
conducted for each affective dimension. A comparison was made between each
normalized distribution and five different randomly-generated distributions of 1000
values with a mean of 0 and a standard deviation of 1. In this test, the null hypothesis is
that the two distributions are sampled from the same population. It was the case that for
all of the eleven affective dimensions except for happy/joyful, the null hypothesis was
rejected for each of the five comparisons. All five tests for happy/joyful failed to reject
the null hypothesis. This suggests that the data are not normally distributed for any of
the affective dimensions except for happy/joyful.
Following this test, QQ plots were generated for the normalized distributions for each of the eleven affective dimensions. Comparing the QQ plot to an idealized normal distribution (represented by the diagonal line) provides an opportunity to identify areas of the distribution that deviate from an expected normal distribution. As expected, the greatest deviation from the idealized normal distribution occurs at the ends of the scale. This deviation is typically seen when the tails of a distribution are wider than the normal distribution. Nevertheless, the majority of the plot for each affective dimension followed the line closely. This is not surprising, considering that each distribution exhibited very large numbers of ratings at the extremes, leading to thicker tails. Nevertheless, the majority of the distribution between the modes on the extremes exhibits normalcy.

One common reason for large stacks at the extremes of a scale is that the scale exhibits a ceiling and/or floor effect. Due to the scale being bounded, any measurement that was deemed more extreme than the ends of the bounded scale would simply be rated with the end of the scale. One could imagine that if the scale had been infinite, the same kind of stacking at the ends of the scales would not be seen. It is possible that what might be long tails at the ends of a distribution approximating normality could bunch up at the extremes of a bounded scale.

The apparent lack of normalcy is a violation of a major assumption in multiple regression analysis. However, even though most of the normalized scales failed normality tests, the data appears to approximate the normal distribution for most of the distribution. Moreover, the effects of the central limit theorem mitigate the negative
effects of non-normality in large sample sizes. In this case, each affective model was built from a large sample size of over 1,000 observations per scale. On balance, the large sample size and the relatively normal appearance of the distributions suggests that it may not be inappropriate to carry out a multiple regression analysis, despite the formal failures of the normalcy test. However, the results of such an analysis must be interpreted with caution.

Regression Analysis

There are many methods for choosing predictor variables for a regression model, and there is little agreement in the field for the superiority of any one method over another. However, it is always a good idea to choose predictor variables thoughtfully, with careful consideration of collinearity, interaction among predictors, and theory.

The first step employed in choosing a model for each affective dimension was to perform all-subsets regression. This strategy compares every possible combination of predictor variables for every possible number of predictors with every other combination, and selects those models for each number of predictors that account for the greatest percentage of the variance in participant responses. In the case of this study, there were sixteen total predictor variables (Table 7.1). The five best-fitting models were chosen for each number of predictors and two plots were constructed for each of the models: each model was plotted against its adjusted $R^2$ in one graph and Mallow’s $C_p$ in a second graph.
Generally, the greater the number of predictors included in a model, the more variance the model will be able to account for. However, even though most models are chosen to maximize the amount of variance accounted for, there is some effect of diminishing returns; the amount of variance accounted for increases marginally for every added predictor past a certain point. It is therefore important to weigh the advantages of increased variance accounted for against the disadvantages of a model that uses too many predictors and therefore over-fits the data. The visual analyses of the graphs resulted in choosing 6 predictor variables for happy/joyful, lonely, and striving/yearning, 7 predictors for contentment and sad/depressed/tragic, 8 predictors for suspense/anticipation and weighty, and 9 predictors for calm/serene, carefree, dark, and unsettled/anxious.

After selecting the number of variables for each regression, a plot was inspected for the best five models for each number of variables up to the number selected. An example of one of these plots is provided in Figure 7.1. In choosing which predictor variables to include in each model, the most important criterion was relative importance. The relative importance of each variable was estimated based on how frequently it was chosen to be part of these various models. Typically, four or five variables were more commonly chosen than the rest of the predictors. These were chosen to be part of the regression. Finally, the best five models for the chosen number of predictor variables were compared, and one model was chosen according to the author’s intuitions. The predictor variables for each affective dimension, along the amount of variance accounted for (Adjusted $R^2$), are shown in Table 7.2.
Figure 7.1. The plot for the five regression models that account for the greatest for each of up to six predictor variables for lonely. As seen in the mostly filled-in columns, mode, close, dense, speed, surp, and hpit seem like important predictors for this affect.

Table 7.2 shows some interesting results. One interesting observation is the relatively high values for Adjusted $R^2$, signifying the amount of variance in participant responses that can be accounted for by measuring only those musical parameters shown in the center column. Given all of the possible sources of error in this study, discussed above, the high numbers here are encouraging.

In many ways the predictor variables selected for each affective dimension, and their effect on perceived affective rating, are not surprising and confirm basic intuitions about the perception of affective content in music. Generally, positive affective dimensions are associated with the major mode and negative affective dimensions are associated with the minor mode. Higher arousal affective dimensions, such as unsettled/anxious, happy/joyful, and carefree are associated with higher pitches,
<table>
<thead>
<tr>
<th>Affective Dimension</th>
<th>Regression Parameters</th>
<th>Adjusted $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calm/Serene</td>
<td>Legato articulations; Major mode; Faster harmonic tempo; Lower high pitch; Diminuendo; Presence of cadence; Common melodic succession; Less harmonic dissonance; Expected melodic progression; No tendency tones</td>
<td>0.362</td>
</tr>
<tr>
<td>Carefree</td>
<td>Major; Higher highest pitch; Diminuendo; Faster surface rhythms; Cadence; Legato artic.; Less harm. dissonance; Higher lowest pitch; Faster harm. tempo</td>
<td>0.242</td>
</tr>
<tr>
<td>Contentment</td>
<td>Major mode; Legato articulations; Faster harmonic tempo; Cadence; Less dissonant harmonies; No tendency tones; Slower surface rhythms</td>
<td>0.338</td>
</tr>
<tr>
<td>Dark</td>
<td>Minor; Lower high pitch; Slower harmonic tempo; Staccato; No cadence; Uncommon harm. progression; More dissonance; Tend. tones; Fast tempo;</td>
<td>0.319</td>
</tr>
<tr>
<td>Happy/Joyful</td>
<td>Major mode; Higher highest pitch; Legato articulations; Presence of cadences; Faster surface rhythm; Decrescendo</td>
<td>0.249</td>
</tr>
<tr>
<td>Lonely</td>
<td>Slower surface rhythms; Minor mode; No cadence; Less dense harmonies; Lower highest pitch; Faster tempo</td>
<td>0.313</td>
</tr>
<tr>
<td>Sad/Depressed/Tragic</td>
<td>Minor mode; No cadence; More common melodic progression; Tendency tones; Lower highest pitch; Faster tempo; Less dense harmonies</td>
<td>0.281</td>
</tr>
<tr>
<td>Striving/Yearning</td>
<td>Minor mode; No cadence; Crescendo; Louder dynamic; Tendency tones; Lower lowest pitch</td>
<td>0.200</td>
</tr>
<tr>
<td>Suspense/Anticipation</td>
<td>Minor mode; Slower harm. tempo; Faster tempo; Tendency tones; No cadence; Staccato articulation; More dissonant harmonies; Denser harmonies</td>
<td>0.283</td>
</tr>
<tr>
<td>Unsettled/Anxious</td>
<td>Higher highest pitch; Crescendo; Minor mode; Tendency tones; No cadence; Staccato articulations; Uncommon melodic succession; More dissonant harmonies; Slower harmonic tempo;</td>
<td>0.353</td>
</tr>
<tr>
<td>Weighty</td>
<td>Lower highest pitch; Lower lowest pitch; Crescendo; No cadence; Tendency tones; Louder dynamic; Minor mode; More dense harmonies</td>
<td>0.256</td>
</tr>
</tbody>
</table>

Table 7.2. The results from the regression for each affective dimension. The rightmost column shows the adjusted $R^2$ for each affective dimension, or percentage of variance accounted for by the model. In the center column, the complete list of musical features that increase ratings of each affective dimension are given. For example, higher lonely ratings are correlated with slower rhythms, the minor mode, no cadence, fewer concurrently attacked notes, a lower highest pitch in the excerpt and a faster tempo.

whereas lower arousal affective dimensions, such as dark, calm/serene, and weighty are associated with lower pitches. More harmonic dissonance, tendency tones, and less strong or no cadential motion are generally associated with negatively valenced affective
dimensions that are typically more intense, such as unsettled/anxious, striving/yearning, and suspense/anticipation. These observations accord with the ethological model (Lorenz, 1982).

Follow-up study: Testing the model

In order to investigate the extent to which the linear model generated from data in the main study is generalizable to other works, a follow-up study was conducted to test the predictions made by the models generated from the data from the first study. By measuring the same musical parameters used in the development of the regression models for novel excerpts, predictions can be made about new stimuli. These predictions can further be tested by asking a new group of participants to rate the perceived affective content of these new excerpts and comparing the predicted results to the participant-generated responses.

Stimuli

The selection of new excerpts to test is an important issue. While it has been suggested above that presenting excerpts in short segments may result in participants focusing on more low-level aspects of the sound, some excerpts may nevertheless be more appropriate to test than others. To reduce the influence of confounding variables, excerpts of solo piano music were chosen. It seems likely that the effect of different instruments or groups of instruments would influence ratings of the perceived affective content of excerpts. For example, instruments with lower timbres such as tubas or
‘cellos may elicit higher ratings for dark or weighty than instruments with higher
timbres such as piccolos or trumpets.

An additional variable that may confound the results could be an effect of
composer or year of composition. One strategy would be to include excerpts from many
composers over many years to test how well the constructed models perform in many
different styles. However, it is likely that different composers or different musical styles
may utilize different musical features in the expression of the same affect. For example,
what musical features are recognized as happy/joyful or lonely for J. S. Bach may be
quite different from the musical features indicative of those same affects for Debussy.
Therefore, it was therefore to only include music from Beethoven’s corpus. Although
the low-level musical features that are associated with various affects may also apply to
other styles and composers, it seems that an important first step is to confirm that they
apply in the same way across different compositions by the same composer.

The final constraint on excerpt selection was that all excerpts should be taken
from the same genre of composition. It may be the case that, even while restricting
musical excerpts to those composed by Beethoven, there could be differences in how
affects are perceived between sonata-form movements and other works, such as
character pieces or dance works. Therefore, only excerpts from Beethoven’s sonatas
were considered. In total, Beethoven wrote 32 piano sonatas, from two to four
movements each, totaling 102 movements.

Another possible confounding variable could be excerpt length. Recall that five-
second excerpts were used specifically for the *Pathétique* sonata, second movement,
because that length seemed appropriate to the tempo and “affective flux” of the movement. While different lengths of excerpts may be more appropriate for other movements, using differing lengths of excerpts may have an effect on the perceived affective content of those excerpts. Therefore, all excerpts used in this study were also five seconds in length.

One approach to selecting excerpts would be to randomly sample five-second portions of Beethoven sonata movements. The low-level musical features that participate in the regression models would be measured, resulting in a prediction of the perceived affective content of the selected excerpts. These predictions could then be tested by comparing the predictions to participant responses. Selecting excerpts randomly, however, could result in the testing of excerpts that are not particularly evocative of strong affect. A more compelling approach would be to solicit excerpts from a panel of experts that are considered particularly powerful expressions of varying affects. The predictions made by the model could then be compared not only to the participant responses, but also to the explicit descriptions of the perceived affective content of the excerpts provided by the experts.

In order to find excerpts suitable for the follow-up study, professional music theorists on the Society for Music Theory ListServ and professional piano performers and educators on the Piano Street forum were polled to provide excerpts from Beethoven’s piano sonatas considered particularly evocative of emotion. The following request was sent out by email to the Society for Music Theory “talk” ListServ and was posted on the “Repertoire” section of the Piano Street web site forum:
I’m interested in some of the most expressive moments in Beethoven’s piano sonatas. What are some of your favorite excerpts that you feel express the most powerful emotions? Any emotion is fine (joy, sorrow, agitation, passion, jealousy, etc.), but ideally it would be an exemplar of that emotion. If you would like to share your favorite moments, please pass on the Sonata number (or Op. number), which movement it is, in what measures it happens, and what emotions you feel that it represents.

For example, I find the opening of the sonata No. 26 (Op. 81a, “Les adieux”) to be very moving. The C minor triad seems to exude a mournful, resigned spirit after the Eb-G of the first sonority and the Bb-F open fifth almost establish Eb major. Or, I’ve always found the “E” major modulation in ms. 42-44 in the second movement of the Pathétique to exhibit an almost paradoxical joyful, resolved feeling that is yet colored by the agitation and unease expressed by the dense, insistent triplet chordal figures underneath the melody.

In total, there were seven responses from the Society for Music Theory and eight responses from Piano Street. All responses were accepted for the study, providing that they were limited to the piano sonatas and that there was some indication as to where in the movement the described emotion happened. Several responders on the Piano Street forum provided suggestions for Beethoven’s piano concertos and some responders simply provided an entire movement or work but did not specify a section. Several responders provided multiple suggestions for excerpts, and these were all accepted.

The recordings were selected from the collection of recordings on Naxos.com. The database was searched by entering the composer’s name, opus or sonata numbers, or the descriptive terms (such as Waldstein) associated with the work. The recording corresponding to the first search result that matched the correct work was chosen as the source for the excerpt. Some of the suggested excerpts spanned much more than five seconds of time for the recording selected. In these cases, a portion of the suggested excerpt was chosen. For suggested excerpts that mentioned specific themes (e.g., “Moonlight sonata finale, opening theme”), the beginning of the theme was chosen as the
<table>
<thead>
<tr>
<th>Excerpt</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Tempest sonata, finale, mm. 169-173</td>
<td>“Determination”</td>
</tr>
<tr>
<td>2. Moonlight sonata, finale, opening theme, mm. 1-4</td>
<td>“Fury”</td>
</tr>
<tr>
<td>3. Hammerklavier sonata, I, mm. 22</td>
<td>“Hope in a dark world, in context”</td>
</tr>
<tr>
<td>4. Hammerklavier sonata, II, mm. 1-6</td>
<td></td>
</tr>
<tr>
<td>5. Waldstein sonata, I, opening of 2nd theme, mm. 35-37</td>
<td>“Contentment, nostalgia”</td>
</tr>
<tr>
<td>6. Waldstein sonata, III, mm. 240-245</td>
<td>“Heavenly”</td>
</tr>
<tr>
<td>7. Appassionata, I, beginning of 2nd theme, mm. 34-36</td>
<td>“Gaining in confidence”</td>
</tr>
<tr>
<td>8. Op. 109, finale, Var. VI, mm. 11b-12a</td>
<td>“Transcendental”</td>
</tr>
<tr>
<td>11. Op. 26, I, Variation 3, mm. 7-8</td>
<td></td>
</tr>
<tr>
<td>13. Les Adieux, finale, mm. 182-183</td>
<td>“Sobs and tears of joy, but somehow suggestive of the inscrutable quasi-masochistic fallibility of one who has made an accommodation with pain”</td>
</tr>
<tr>
<td>14. op. 111, I, m. 3a</td>
<td>“Overcoming the struggles of mortal life and transcending against all odds”</td>
</tr>
<tr>
<td>15. Op. 2, No. 1, finale, mm. 34-37</td>
<td>“Like a storm came and then left leaving a smooth, yet still unsettling, reflection upon what just happened”</td>
</tr>
</tbody>
</table>

Table 7.3. The fifteen suggested excerpts and the free-response descriptions of affective content provided by the responders to the email/forum post.

beginning of the excerpt and the excerpt was terminated at five seconds. For suggested excerpts that gave specific measures (e.g., “Op. 10, No. 3 Largo e mesto, ms. 21-25”) but that were longer than five seconds, the author listened to the excerpt and tried to select the portion that he felt was most evocative of the affective description given by the
recommender. The result was fifteen excerpts, each from a different Beethoven sonata movement, that theorists and pianists had identified as especially evocative of particular powerful emotions. Not all of the excerpts identified were accompanied by a specific description of the emotional content. For example, a respondent might simply exclaim how much they loved a moment, but not offer a specific emotional characterization. The fifteen excerpts are listed in Table 7.3, including the movement from which they were taken, the measure numbers of the excerpt, and any descriptions of the emotional content of the excerpts provided. As in the main study, each excerpt was edited with a 500-ms fade-in and fade-out to avoid abrupt onsets and offsets.

Participants

19 participants were recruited from the Ohio State University subject pool. This was one of several experiments that could be selected by participants in order for them to receive course credit for sophomore-level Aural Skills. The mean age for the participants was 20.5 years (standard deviation = 2.2), and the mean number of years of musical training was 12.5 (sd = 4.0). Eight of the participants were female and eleven were male.

Procedure

Unlike the previous study, all participants were able to listen to and rate all of the excerpts for all of the affective dimensions. This was possible because there were only fifteen excerpts and only eleven affective dimensions. Participants listened to the excerpts in blocks of 16 trials each – they were given one affective dimension and then
rated each excerpt for that affective dimension. In addition to the fifteen excerpts, one randomly-selected excerpt was presented at the end of each block, in order to test for intrasubjective reliability. However, due to a randomization glitch, some affective dimensions used more than one repeated trial at the expense of presenting every excerpt to the participant. The order of presentation was randomized for both the affective dimensions and the excerpts within each trial block.

The participants were tested individually in an Industrial Acoustics Corporation sound attenuation room. Participants listened to the stimuli with free-field speakers, adjusted to a comfortable volume. The experimenter read the directions aloud while the participant read along with the printed instructions.

“What can music express?”

The purpose of this study is to gather information about music and emotion. At the end of the experiment, I’ll say more about our specific goals.

In this study, you will hear short excerpts from early Romantic piano music and rate the emotional expression of those excerpts. Please do not evaluate your own emotional response to the music, but simply what emotions you think the music is trying to express or convey.

You can listen to each excerpt as many times as you like by clicking on the PLAY button.

Each emotional scale is represented by a line, with the right end of the scale representing the maximum amount of that emotion and the left end representing the minimum amount of the emotion.

The excerpts will be presented in blocks of 16 trials each. You will rate a different emotion for each block of trials.

For each emotion, each excerpt and scale will appear one at a time. You must adjust each scale before moving on to the next excerpt. Once the scale is adjusted, click NEXT EXCERPT and the next rating scale will appear.

For the last rating excerpt of the emotion block, you must click NEXT EXCERPT and then NEXT EMOTION to move on to the next sound excerpt.

The experiment will take roughly 45 minutes to complete.

Do you have any questions?
After participants read the instructions, they were given some examples of the difference between perceived and felt emotion. They were reminded to rate only the emotions that they perceived the music to be expressing, rather than any emotions that they felt in response to the excerpts. After answering any questions the participant had, the experimenter observed while the participant attempted four practice trials, two each for two different affective dimensions. The interface for the study was similar to the interface for the main study (Example 4.2).

After the portion of the study conducted on the computer participants were given a chart to fill out, shown in Example 7.1. The purpose of the chart was to translate the descriptions of the emotional expression of the excerpts given by responders to the email/forum post into the terms that were tested in the study. Specifically, participants were instructed to put a “+” in any box in which the affective term on the top (from this study) was similar to or close in meaning to the description given by the responders on the left, to put a “−” in any box in which the term on the top meant the opposite of the description on the left, and to put a “0” in any box in which the term on the top seemed unrelated to the description on the left.

Finally, participants were asked a number of follow-up questions. Specifically, participants were asked to describe their experience, whether they enjoyed it, how difficult they found it, what strategies they used, whether they knew any of the excerpts they heard, how they interpreted the middle and left of the scale, and what they thought the study was about.
Example 7.1. The chart filled out by participants. Affective descriptions by responders are shown on the left and the eleven affective terms used by the models are on the top.

Results – Provided descriptions

Sixteen musical features were measured in each of the excerpts from the *Pathétique* sonata used in the main study. These features were used as predictor variables to create a regression model for each of the eleven tested affective dimensions. The same sixteen musical features were measured for each of the fifteen excerpts used in the follow-up study, and these measurements were used in the regression equations to predict the perceived affective content of those excerpts. The predictions are shown in Table 7.4.
Recall that the variables used to build the model were first normalized by subject-scale, and so were measured in standard deviations away from the mean for that affective dimension for that participant. The values in Table 7.4 should therefore be interpreted for each affective dimension as standard deviations away from the mean of excerpts in the Pathétique sonata, second movement. While the actual numbers presented do not represent some sort of universal value of the amount of each affective dimension per excerpt, the rank-order of the values for the excerpts are nevertheless have the highest and lowest scores are in **bold** and highlighted.

The ability of a regression equation to make predictions from predictor variables is neither surprising nor impressive – that is how the equation is defined. It is more noteworthy when a model built from a data set is used to accurately predict data not included in the building of the model. When the predictions of a model are consistent with further empirical observations, this is evidence that the model may be generalizable to a broader population.

In the case of the predictions generated from the fifteen excerpts provided by revealing. For each affective dimension in Table 7.4, the excerpt that was predicted to responders, there are two ways in which the predictions can be tested. In the first case, the predictions can be compared with responses provided by participants who rate each excerpt on the same affective dimensions after listening to those excerpts. Those results are described in more detail below. In the second case, the predictions offered by the model can be compared with free-response descriptions of the perceived affective content of the excerpts provided by responders. However, it is difficult to determine
Table 7.4. The predictions made by the models for each suggested excerpt, measured in standard deviations away from the mean of the *Pathétique*. The excerpts with the highest and lowest predictions for each affective dimension are in **bold** and highlighted.

| Table 7.5. The mean ratings given by participants in “translating” the eleven affective descriptions given by responders into the terms used by the models. | 138 |
how free-response descriptions might relate to the affective dimensions that the model predicts. For example, the descriptions “anguished grappling with inevitable and depressing conclusion,” used to describe the Op. 10, No. 3, m. 21 excerpt, and “like a storm came and then left leaving a smooth, yet still unsettling reflection upon what just happened” used to describe Op. 2, No. 1 are not immediately translatable into the affective dimensions used in this study.

In order to translate the descriptions provided by the responders, the data was analyzed from the charts given to participants in the follow-up study. Recall that there were three categories used to describe the relationship between the terms used in this study and the descriptions of the excerpts: “+,” which was used when the terms were deemed similar; “−,” used when the terms were deemed opposite in meaning; and “0,” used when the terms were deemed neither similar nor opposite. By considering “+” to be 1 and “−” to be -1, the relationship between the descriptions provided and the affective dimensions used in this study can be assessed numerically.

An average was calculated for every box at the intersection of a description and an affective dimension, shown in Table 7.5. The descriptions are numbered according to the same numbering used in Table 7.5 for easy comparison between the excerpts and the descriptions. These averages are bounded between -1 and 1, and the numerical space can be divided into three equally-sized portions. If the average shown in Table 7.5 is between -1 and -½, the expectation is that the model’s prediction should be negative (operationalized here as less than -0.5 standard deviations from the mean). If the average is between +½ and +1, the model’s prediction should be positive
(operationalized as greater than +0.5 standard deviations from the mean). Finally, if the average is between $-\frac{1}{2}$ and $+\frac{1}{2}$, the model’s prediction should be close to 0 (between -0.5 and +0.5 standard deviations from the mean).

Only eleven of the fifteen excerpts were provided with an affective description by the responders, and so only eleven of the descriptions can be tested against the model’s predictions in this way. In total, there were seventeen participants who filled out the chart and eleven descriptions provided by responders, leading to a total of 187 ratings per affective dimension. Of these 187 ratings, only 79 of the model’s predictions for calm/serene matched the descriptions of the excerpts translated into the eleven affective dimensions, 77 matched for carefree, 75 matched for weighty, 73 matched for happy/joyful, 66 matched for unsettled/anxious, 62 matched for contentment, 59 matched for dark, 55 matched for striving/yearning, 53 matched for lonely, 52 matched for sad/depressed/tragic, and 49 matched for suspense/anticipation.

Following this descriptive investigation, a statistical test was performed on each affective dimension using Kendall’s tau, which uses a rank-order correlation to determine if two data sets are significantly correlated above 0. For this test, $\alpha$ was set to 0.1. Due to the danger of multiple tests, $\alpha$ was corrected using the Sidak-Bonferroni correction to 0.0048. Of the eleven scales, only the model’s predictions for calm/serene ($p = 0.00033$) and weighty ($p = 0.0022$) were significantly correlated with the descriptions given by the responders across all of the excerpts.

There are many reasons why the descriptions given by responders could not be accurately predicted by the model. Of course, the most obvious reason is that the
movement-specific model may not be a good predictor of the perceived affective content in other works. Nevertheless, many of the descriptions are poetically worded, and so may be confusing to understand for the participants who translated them into the terms used in the study. For example, in post-experiment interviews, many participants confessed to not knowing the meaning of “transcendental.” Likewise, some descriptions, such as “hope in a dark world, in context” or “sobs and tears of joy, but suggestive of the quasi-masochistic fallacy of one who has made an accommodation with pain” have subtle mixtures of emotions and do not permit well of a binary decision for a given emotion. In these cases, some participants may have focused on “hope” or “joy” in these descriptions while others may have focused on “dark” or “sobs...pain.”

Recall that the descriptions were provided by professional musicians or academics. As such, the descriptions may have used language that was less familiar or more difficult for undergraduate students to understand. The disagreement in the responses may therefore be more of a product of different strategies used in interpreting complex mixtures of emotion rather than the specific limitations of the model.

Another possible reason that the descriptions may not have been accurately predicted by the model could be that the descriptions might describe bigger sections of the music than the five-second windows that the study was constrained to use. In many cases, excerpts were suggested that were explicitly longer than five seconds, and editing had to be performed. In addition, the excerpts were suggested by theorists and pianists who presumably know the works well. As such, the descriptions may be addressing the perceived affective content of a moment in contrast to a much larger portion of the work.
preceding that specific excerpt. This sort of context-related affect was explicitly
mentioned by the responders in two of the eleven cases (the Waldstein second theme in
the first movement and the slow, major-mode theme at the end of the Les Adieux sonata).
As mentioned previously, using short excerpts will restrict the types of processing used
to determine affective content to favor bottom-up processes. As a result, any effect of
deep structure or longer context is necessarily mitigated.

In order to test the possibility that the failure to find significant correlations
between predictions and descriptions was due to excerpt, the same Kendall’s tau test
was performed on all of the predictions averaged across excerpt and not across affective
dimension. There were eleven excerpt descriptions and seventeen participants translated
the descriptions into the terms used by the model. Of the eleven excerpts, the model’s
predictions were significantly positively correlated with the translations of the
descriptions given for the Moonlight sonata ($p < .00001$), for the first movement of the
Waldstein sonata ($p < .00001$), for the Appassionata sonata ($p = 0.00035$), for the Op. 109 ($p
= 0.00037$), for the Les Adieux sonata ($p = 0.00089$), for the Op. 111 ($p < .00001$), and for the
Op. 2, No. 1 ($p = 0.00021$). Additionally, the Op. 10, No. 3 (ms. 21) had a $p$-value of 0.008,
just above the corrected family-wise $\alpha$-level, and the Hammerklavier, first movement, had
a $p$-value of 0.04 – skewed in the predicted direction. Taken together, seven of the eleven
predictions of the model were significantly correlated with the translations of the
descriptions. Two more predictions had $p$-values below the family-wise critical value,
but failed to meet the pair-wise criterion.
These results are consistent with the hypothesis that the predictions made by the model can be generalized to other works. Moreover, the results are consistent with the idea that the regression model built from the observations of one movement can provide predictions in line with professional musicians’ open-ended descriptions of the affective content of excerpts from other Beethoven sonata movements. It seems likely that the low correlations between the model’s predictions and the translations for the Tempest sonata ($p = 0.764$) and the Waldstein sonata, third movement ($p = 0.444$) were enough to mitigate an affective dimension effect across all excerpts. It may be that the affective descriptions provided for these two excerpts (“determination” and “heavenly,” respectively) did not have clear corollaries within the eleven affective dimensions tested in this study, and therefore were difficult to translate into the terms used in the model.

**Results – Participant Ratings**

While the regression models were clearly unable to predict professional musicians’ descriptions of the perceived affective content for two of the fifteen excerpts offered by responders to the email/forum post, it provided predictions in line with the descriptions for seven of those excerpts. Additionally, it is difficult to determine if the reason why the two excerpts that failed to exhibit significant results was because the model was ill-suited to making accurate predictions for those excerpts or because the descriptions were difficult to translate into the terms used in this study.

The data provided by participants who rated the perceived affective content of the fifteen excerpts used in the study, however, offer an additional opportunity to
investigate the relationship between the models’ predictions and the perceived affective content of the excerpts. It may be that the descriptions of the perceived affective content of the suggested excerpts provided by the responders are outliers in the population of musicians who would listen to these excerpts. It also may be that there is some loss of meaning in the translation from the descriptions offered by the responders into the language of the affective dimensions used by the models. By allowing participants to rate the perceived affective content of the excerpts directly, a comparison can be made between the predicted ratings and the actual results from the study.

Before analyzing the data, however, it is appropriate to investigate intrasubjective reliability and eliminate any data sets that are unreliable. Each participant listened to at least one repeated excerpt for each affective dimension, averaging 36.7 repeated excerpts per participant. As in the main study, the exclusion criterion was *a priori* set to +0.25. For the nineteen participants, the mean intrasubjective correlation between the ratings for the first trials and the ratings for the repeated excerpts was +0.853 (standard deviation = 0.089). The highest intrasubjective correlation was +0.966, and the lowest was +0.605. In light of the strong within-participant correlations, none of the data were eliminated.

It is noteworthy that the intrasubjective correlations are much higher than in the first study. One possible reason for this could be that the excerpts in this study were specifically selected for the expression of extreme affects. The excerpts being selected from different movements increases the variability in the excerpts. Differences between
the perceived affective content in the excerpts are therefore probably stronger, and the starker contrast between excerpts likely leads to higher correlations.

In the first study, which was exploratory in nature, the intersubjective reliability was investigated to determine which ratings scales were performing well as measures of perceived affective content in music. In that study, four affective dimensions were eliminated because the data measured on those scales were unreliable. In this study, the assumption was made that the scales were reliable means of assessing the perceived affective content of Beethoven sonata movements, and therefore it was deemed unnecessary to investigate intersubjective reliability in this study.

The raw scores were converted into normalized scores by averaging across subject-scale. This important step was necessary for the comparison of the mean participant ratings in this follow-up study with the model’s predictions, measured in normalized z-scores. Despite this corrective step, there is still not a one-to-one correspondence between the meaning of the model’s predicted ratings and the normalized scores from the participants in the follow-up study. This is because the predictions are measured in standard deviations away from the mean of ratings of excerpts within the Pathétique sonata, whereas the ratings in the follow-up study are measured in standard deviations away from the mean of ratings of excerpts used in the follow-up study. Since the models are based on the Pathétique, one should expect their predictions to not be especially accurate when applied to passages from other works – the units of measure are different. However, if the models based on the Pathétique are reasonably predictive of the affective content of other passages, then this suggests either
that the excerpts tested are similar to the excerpts in the *Pathétique* or that the models capture general aspects of affective musical features.

**Excerpt effects**

It is possible to distinguish a strong and weak version of the research hypothesis. The *strong version* takes the predictions as point-estimates of the means of ratings for the tested excerpts; the hypothesis is that these means are statistically indistinguishable from the population means at an \( \alpha \)-level of .95. A statistical test can be performed to determine whether the prediction falls within a 95% confidence interval of the mean ratings in the follow-up study. Despite the fact that the ratings are based on different means and standard deviations, this finding might be seen if participants somehow adjusted their ratings to the context of the non-*Pathétique* excerpts. This finding might suggest that the mean ratings for excerpts in the follow-up study maps onto the mean ratings for excerpts in the *Pathétique* sonata. However, if the predictions fail to fall within the 95% confidence interval around the data, a *weak version* of the hypothesis can be tested; this hypothesis holds that the predictions, though not exactly the population means, are scaled versions of the population means. This can be tested by using a Pearson’s correlation test.

1. *Tempest sonata*

This excerpt (Example 7.2), taken from the music immediately preceding the retransition, uses arpeggios to outline first a diminished-seventh chord and then a
German sixth chord, two harmonies that strongly imply resolution to the dominant. The passage is played loudly and is marked with a crescendo. The affective dimensions that exhibit mean ratings significantly different from zero are unsettled/anxious, suspense/anticipation, and striving/yearning with high ratings and calm/serene with low ratings.

Only one of the models’ predictions for this excerpt was actually inside the 95% confidence interval (calm/serene). Two of the predictions seemed to be very far from the actual participant responses. Specifically, lonely and sad/depressed/tragic predictions were very high while the participant ratings were close to 0. The reason the predictions were so high might be because of the prevalence of harmonic dissonance and tendency tones in this excerpt and the fact that it is cast in the minor mode. While these

Example 7.2. The excerpt from the *Tempest* sonata finale, mm. 169-173, the music right before the recapitulation. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.
characteristics have been associated with high ratings in the excerpts in the *Pathétique* from which the models are built, the quickness of the rhythms coupled with the strong pedal point in the bass may have subverted the ratings of perceived sad/depressed/tragic and lonely.

Nevertheless, the predictions were generally skewed in the direction of the responses given by participants. With the exception of lonely and weighty, the signs of the predictions were consistent with the participant responses. The affective models predicted high ratings for sad/depressed/tragic, lonely, dark, suspense/anticipation, unsettled/anxious, and weighty, and low participant responses for calm/serene.

From a visual inspection of the plot in Example 7.2, it seems that the predictions of the model overinflated the participant ratings of each of the affective dimensions for this excerpt. This result may occur if the models scaled the predictions differently due to a different mean and standard deviation. In order to test whether the weak hypothesis holds for this excerpt, a test was performed to determine if the predictions were correlated with the results. For this excerpt, the ratings given by participants were correlated +0.66 with the predictions given by the models for this excerpt, but this was not significant, adjusted for multiple tests ($p = 0.013$). Nevertheless, the $p$-value was much lower than the family-wise $p$-value, set at 0.1.

2. Moonlight sonata

The beginning of the finale of the *Moonlight* sonata (Example 7.3) is played very quickly and dramatically. Though the harmonic tempo is very slow (only two chords), many
note events (56!) happen in this short time span. Every two measures, there is also a very sudden fortissimo. The affective dimensions with the highest mean ratings for this excerpt are dark, unsettled/anxious, and suspense/anticipation, while participants rated the excerpt as low in calm/serene and happy/joyful.

Two of the models’ predictions were inside of the 95% confidence interval of the participant ratings for this excerpt: happy/joyful and weighty. However, the signs of the predictions are consistent with the participant ratings for every affective dimension. The highest predicted ratings were for suspense/anticipation, unsettled/anxious, dark, sad/depressed/tragic, and striving/yearning. The lowest predicted ratings for this excerpt were for contentment, calm/serene, and carefree.

Example 7.3. The excerpt from the *Moonlight* sonata finale, opening theme. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.
As in the results for the *Tempest* sonata, a visual inspection seems to indicate that the models overinflated the predictions away from the mean. With the exception of the two predictions that fell within the 95% confidence intervals, every prediction was further from the mean than the participant ratings. A Pearson’s correlation test indicated that the predictions and the results were significantly positively correlated +0.90 ($p < 0.0001$). This is consistent with the weak hypothesis that the models’ predictions are consistent with participant ratings, but to a different scale.

3. Hammerklavier sonata (I)

The excerpt taken from the first movement of the *Hammerklavier* sonata (Example 7.4) is very slow and very quiet. The harmonic rhythm is also very slow, with only a simple G major harmony, played first in root position and then in first inversion. The melody is marked by large leaps – first a major 10th from the grace note to the high $b$, followed by the minor 7th leap from $a$ up to $g$, which doubles the lower $g$ in octaves. Participant ratings were the highest for calm/serene, contentment, and lonely in this excerpt, while they were the lowest for unsettled/anxious, suspense/anticipation, dark, and weighty.

The predictions for this excerpt were more accurate than the previous two, with five of the eleven predictions within the 95% confidence interval. Two of the predictions, however, seemed quite different than the actual participant ratings: unsettled/anxious was predicted to be rated higher than 0 and calm/serene was predicted to be rated
Example 7.4. The excerpt from the first movement of the Hammerklavier sonata, mm. 22. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.

lower than 0. These two discrepancies, which seem related, may have to do with some of the surface features in the excerpt. For the model, high pitches and crescendos were associated with higher unsettled/anxious and lower calm/serene ratings. The presence of both high pitches and crescendos in this excerpt may have counterbalanced the slow rhythmic values and simple major triads to push the predictions closer to 0. Despite the high number of predictions that fell within the 95% confidence interval, the correlation between the predictions and the participant ratings was only +0.41. These two discrepancies may have been enough to lead to an insignificant result for the Pearson’s correlation test (p = 0.108).
4. Hammerklavier sonata (II)

In contrast to the first movement, the second movement of the *Hammerklavier* sonata (Example 7.5) begins quickly, with rapid harmonic change, jaunty dotted rhythms, a gradual melodic descent, a rapid alternation between crescendo and diminuendo, and a strong perfect authentic cadence to close the excerpt. Participants rated this excerpt as very high in happy/joyful, carefree, and contentment, while rating it very low in sad/depressed/tragic and lonely.

Three of the predicted ratings, carefree, suspense/anticipation, and striving/yearning, were within the 95% confidence interval of the actual participant ratings. Nevertheless, several predictions were quite different from actual ratings. The prediction for calm/serene was much higher than the listener ratings, possibly as an

![Assai vivace d.80](image)

Example 7.5. The excerpt from the opening of the second movement of the *Hammerklavier* sonata. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.
artifact of the major mode and the high number of harmonies, features the model correlated with calm/serene. Unsettled/anxious, lonely, and sad/depressed/tragic were also quite far from the actual ratings. Pearson’s correlation test was insignificant at \( R = +0.46 \) (\( p = 0.077 \)), and so the weak null hypothesis was not rejected.

5. Waldstein sonata (I)

Though it exhibits a very chorale-like texture, the second theme of the Waldstein sonata (Example 7.6) was suggested by two separate responders for its emotionally expressive qualities. The melody and the line in thirds to the melody are doubled at the octave, set against a simple contrapuntal bass line. Like most chorale settings, there is a fast harmonic rhythm (every attack is a new harmony), coupled with a very slow surface rhythm. Perhaps the most striking thing about this short excerpt is the deceptive motion from the dominant of the submediant to the subdominant harmony. This excerpt was rated as very high in calm/serene, contentment, and carefree by participants, and very low in unsettled/anxious, dark, and suspense/anticipation.

Unsettled/anxious, calm/serene, and carefree were all accurately predicted by the models, whereas the rest of the predictions were not in the confidence interval, but close to it. In this case, it seems that most of the predictions were higher than actual participant ratings, but they seemed to be highly correlated with participant ratings. The Pearson correlation test indicated that the ratings and predictions were significantly correlated \(+0.77\) (\( p = 0.0030 \)).
Example 7.6. The excerpt from the second theme in the first movement of the Waldstein sonata, mm. 35-37. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.

6. Waldstein sonata (III)

The third movement of the same sonata contains an excerpt with arpeggiated major-seventh chords. These two chords, shown in Example 7.7, are both over a pedal C in the bass, but the first chord is build on that note as root and the second chord shifts up by half-step so that this same note becomes the seventh of the chord. Participants rated this excerpt as high in dark, unsettled/anxious, and suspense/anticipation, and low in happy/joyful, carefree, and contentment.

Likewise, the models predicted high ratings for unsettled/anxious, suspense/anticipation and dark. However, the models also predicted relatively high ratings for both happy/joyful and sad/depressed/tragic. While on the surface this seems like a contradiction, recall that participant ratings of happy/joyful and sad/depressed/tragic
Example 7.7. The excerpt from the third movement of the *Waldstein* sonata, mm. 240-245. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.

In the *Pathétique* sonata were only negatively correlated by -0.635. This means that there is room for these two affective dimensions to be high simultaneously. In the case of the excerpt above, the high levels of dissonance, low density, and lack of a cadence may have led to higher ratings for sad/depressed/tragic, whereas the major quality of the chords (despite the dissonance), the high pitch, and the legato articulations may have led to higher ratings for happy/joyful. This passage might be construed as sending signals for both affective dimensions, even though for participants sad/depressed/tragic clearly won out as the perceived affective content of the excerpt. In post-experiment interviews, many participants described how they tried to be “consistent” with themselves, so if
they rated an excerpt as happy/joyful, they tried to remember that and rate the same excerpt as not sad/depressed/tragic. This effect may have augmented the discrepancy between the ratings for these two affective dimensions.

For this passage, eight of the eleven predicted affective dimensions fell within the 95% confidence interval around the response means. Of the three that did not, happy/joyful was much higher than participant responses and calm/serene was much lower than participant responses. Despite the large number of predictions that fell within the 95% confidence interval of participant ratings, the predictions were only correlated with the ratings at +0.47, which was not significant ($p = 0.072$). This surprising result reveals the weakness of using a correlation test on so few data points, and suggests that a measure of Euclidean distance might be a more relevant estimator of goodness-of-fit. However, there is no significance test for Euclidean distance, and so the correlation test has been used.

7. Appassionata sonata

The second theme in the first movement of the Appassionata sonata is a low melody, doubled at the octave, and opens up the major triad with arpeggiations. The chordal support, even lower, is muddied even further by a low-level hemiola. The harmonic motion is a simple tonic-dominant-tonic motion, expanding a root position tonic to its first inversion. This excerpt was rated as high in calm/serene and carefree and low in suspense/anticipation and unsettled/anxious.
Example 7.8. Target excerpt from the second theme in the first movement of the Appassionata sonata, mm. 34-36. The plot shows the means and confidence intervals around the participant ratings; model predictions are marked with red stars.

The predictions were the highest for carefree, calm/serene, and contentment for this excerpt and were lowest for unsettled/anxious, striving/yearning, and lonely. The reason for most of these affective predictions probably lies with the extremely low register, complete diatonicism, the major mode, and lack of harmonic dissonance. While five of the affective predictions fell within the 95% confidence intervals, three affective dimensions had predictions much different from the participant ratings. Specifically, dark, sad/depressed/tragic, and suspense/anticipation were all predicted to be
positive, but participants rated them as negative. Nevertheless, the predictions were found to be significantly correlated +0.72 with participant ratings ($p = 0.0061$).


The finale for Beethoven’s Op. 109 is a theme and variations, and this excerpt (Example 7.9) is taken from the sixth variation. The excerpt begins with a weak dominant-function chord in a major mode, notable for the sharp dissonance between the leading-tone and the tonic. There is also a stark rhythmic change that happens at the resolution; the rhythms move from thirty-second notes to the relatively slow triplets. However, this rhythm is contrasted against a low rumbling trill doubled in the right hand. The affective dimensions with the highest ratings for this excerpt are happy/joyful and carefree, and the lowest ratings are dark, sad/depressed/tragic, and unsettled/anxious.

The highest predictions for this excerpt are also for happy/joyful and carefree, but weighty was also predicted to be high. Seven of the eleven affective predictions fell within the 95% confidence interval of the ratings for this excerpt. Of the four that did not, the biggest difference was for calm/serene. The fast rhythmic values, strong dissonances, and loud dynamics likely contributed to the low prediction for calm/serene. In the ratings, perhaps the major mode and the use of a harmonic dissonance of a type that was not functionally dissonant (minor second) outweighed the other features to pull up the ratings for calm/serene. Despite the relatively accurate predictions, they were nonetheless not significantly correlated with the ratings, with a correlation of +0.50 ($p = 0.061$).
Example 7.9. The excerpt from Variation VI in the final movement of Op. 109, mm. 11b-12a. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.


This excerpt (Example 7.10) outlines a simple tonic-dominant motion in D minor. The bass line simply arpeggiates the harmonies, while the melody is embellished with turns and grace notes. The result is a slow, lyrical melody accompanied with a graceful, thin texture. The highest ratings were for lonely, sad/depressed/tragic, striving/yearning, and dark, whereas the lowest ratings were for happy/joyful and carefree.
Example 7.10. Measure 11 in the first movement of Op. 10, No. 3. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.

The highest prediction for this excerpt was for unsettled/anxious, whereas the lowest predicted ratings were for contentment, calm/serene, carefree, happy/joyful, and weighty. Four of the eleven predictions were within the 95% confidence interval, and of the seven that were not, the biggest discrepancies were for unsettled/anxious, and calm/serene: specifically, listeners judged the excerpt to sound much less unsettled/anxious and more calm/serene than predicted. The correlation for this excerpt was +0.59, not statistically significant ($p = 0.028$).
10. Op. 10, No. 3, Largo e mesto, ms. 21

This excerpt (Example 7.11), taken from ten measures later in the movement, tonicizes the dominant with a first-inversion dominant-seventh chord. The chord resolves deceptively, however, to a very low F major triad decorated by a half-step ornamental figure of the dominant of the chord. The dynamic of this passage is very loud, and the low register serves to make the timbre much darker. The highest ratings for this excerpt were for weighty, lonely, dark, and sad/depressed/tragic, and the lowest ratings were for happy/joyful, contentment, and carefree.

None of the predictions were within the confidence intervals of the participant ratings. The highest predictions were for unsettled/anxious, suspense/anticipation,

Example 7.11. Measure 21 in the first movement of Op. 10, No. 3. The plot shows the model predicted much higher ratings than participants offered. Nevertheless, the predictions were only correlated +0.38 with the response means, not statistically significant ($p = 0.126$).
dark, and striving/yearning. The lowest predictions were for calm/serene, carefree, and happy/joyful. Despite failing to accurately predict the participant ratings, evidenced by a lack of any predictions to fall within the confidence intervals, most of the predictions were very close to the ratings. The biggest exception is unsettled/anxious, in which the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.


This excerpt (Example 7.12), taken from the first movement, third variation, exhibits a transformation of the minor subdominant into the leading-tone fully-diminished seventh of the dominant. This chord resolves to the dominant with a suspension/retardation figure on the downbeat. The highest ratings for this excerpt are sad/depressed/tragic, striving/yearning, suspense/anticipation, weighty, lonely, and dark. The lowest ratings are for carefree and happy/joyful.

Eight of the eleven predictions were within the 95% confidence interval of the ratings. The highest predictions for this excerpt were for dark, sad/depressed/tragic, lonely, and weighty, whereas the lowest predictions were for happy/joyful and carefree. Of the three predictions that were outside of the confidence intervals, all three were close to the participant ratings. The predictions were significantly correlated with the ratings +0.79 (p = 0.002).
Example 7.12. The excerpt from the first movement of Op. 26, mm. 7-8. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.

12. Op. 26 (IV)

This major-mode passage (Example 7.13) is very rhythmically active with a walking bass line and a heavily figured melody. There is also a very strong perfect authentic cadence in the middle of the excerpt. The highest ratings for this excerpt were for happy/joyful, carefree, and contentment. The lowest ratings for this excerpt were sad/depressed/tragic, lonely, dark, striving/yearning, and weighty.

The highest predictions for this excerpt were for calm/serene, contentment, carefree, and happy/joyful, whereas the lowest predictions were for unsettled/anxious, suspense/anticipation, and dark. Four of the predictions fell within the 95% confidence intervals, and the rest were close with the exception of unsettled/anxious, calm/serene,
Example 7.13. The excerpt from the fourth movement of Op. 26, mm. 119-124. The plot shows the means and confidence intervals around the participant ratings and the models' predictions, marked with red stars.

The *Pathétique* sonata that the model was built on. The fast harmonic rhythm of this excerpt may possibly account for this inflated value. Due in part to these differences, the predictions were only correlated +0.51 with the ratings, not statistically significant ($p = 0.053$).

13. *Les Adieux* sonata

This excerpt (Example 7.14) was one of the two excerpts suggested principally because of the contrast between the excerpt and the surrounding context. After a rhythmically and chromatically dense movement, the tempo and rhythm slow down markedly and
Example 7.14. The excerpt from the *Les Adieux* sonata, mm. 182-183. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.

Simple arpeggios outlining major triads are introduced. This excerpt simply outlines the tonic triad in a slow tempo in contrary thirds, with a crescendo at the end of the excerpt. The highest ratings for this excerpt are for carefree, calm/serene, happy/joyful, and contentment. The lowest ratings are for dark, sad/depressed/tragic, unsettled/anxious, and suspense/anticipation.

Even though none of the ratings for this excerpt included 0 in their confidence intervals, most of the predictions stayed close to 0. Only the prediction for weighty was within the confidence interval. The rest, though matching the sign of the ratings (except for calm/serene), were very close to 0. Nevertheless, the predictions were significantly
correlated with the ratings +0.74 ($p = 0.0048$). The only rating that was more than 0.5 standard deviations from the mean was dark.


This excerpt (Example 7.15) comes from the slow introduction to Op. 111. It is marked by loud, sudden attacks, marked fortissimo, and long periods of silence. Moreover, the harmony of the excerpt clearly articulates the dominant of a minor key, outlined by the fully-diminished seventh chord and the diminished-seventh leap in the bass. Unsurprisingly, the highest ratings for this excerpt are for dark, weighty, suspense/...
anticipation, sad/depressed/tragic, and unsettled/anxious. The lowest ratings are for carefree, happy/joyful, contentment, and calm/serene.

Three of the predictions were within the 95% confidence intervals of the ratings. Of those that were not, most of the predictions were stronger than the corresponding ratings. This is consistent with the predictions being scaled to a different variance. This is borne out by the correlation of +0.85, which was significant ($p = 0.00045$). The highest predictions for this excerpt were unsettled/anxious, suspense/anticipation, striving/yearning, dark, and weighty. Conversely, the lowest predictions for this excerpt were calm/serene, contentment, and carefree.

15. Op. 2, No. 1

The final excerpt (Example 7.16) is taken from the final movement of Beethoven’s Op. 2, No. 1. This excerpt is characterized by an insistent triplet figuration in the left hand, outlining very low, almost closed-position triads. There is a strong cadential second-inversion tonic-dominant perfect authentic cadence in the middle of this minor-mode excerpt. After this cadence, the right-hand jumps up a tenth and plays a descending line doubled at the octave. This passage had high ratings for dark and unsettled/anxious, and exhibited low ratings for calm/serene, carefree, and contentment.

The predictions for this excerpt were not very accurate. Although the predictions for suspense/anticipation, weighty, and striving/yearning were nearly identical to the mean ratings for the excerpt, the predictions for unsettled/anxious, calm/serene, carefree, contentment, and happy/joyful were all inverted from the ratings given. Not
Example 7.16. The excerpt from Op. 2, No. 1 finale, mm. 32-35. The plot shows the means and confidence intervals around the participant ratings and the models’ predictions, marked with red stars.

surprisingly, predictions were not significantly correlated with ratings ($p = 0.61$), and was in fact -0.099.

Summary

In the discussion above, each excerpt was tested individually for accuracy in predictions for all eleven affective dimensions. The predictions for some of the excerpts were much more accurate than others, in consideration of the strong version of the hypothesis. For example, the third movement of the Waldstein sonata and the first movement from Op. 26 each had 8 predictions fall in the 95% confidence interval and seven predictions for the Op. 109 excerpt fell within the confidence interval. Additionally, five predictions fell within the confidence interval for both the first movement from the Hammerklavier
sonata and the *Appassionata* sonata. It may be that these excerpts were more similar to the *Pathétique* sonata, and so the models built from the music in that excerpt may have been better at predicting listener responses.

Additionally, from a visual inspection of the plots above, it seems that the closer the observed responses were to the mean the more accurate the predictions were. This suggests that the more extreme the perceived surface-level affective content of an excerpt, the weaker the predictions are. A likely explanation for this effect is that the model is not powerful enough. If the variance in the excerpts tested was larger than the variance in the excerpts used to build the model, then there would be a loss of accuracy in predictions in ratings further away from the mean.

Many of the other excerpts exhibited significant correlations between the predictions and the actual listener ratings. With only eleven affective dimensions per excerpt, significant correlations are difficult to obtain. Nevertheless, the predictions for six of the fifteen excerpts were significantly correlated with the listener ratings, and only two of those were excerpts mentioned above with high numbers of strong accurate predictions. Visual inspections also indicated that many of the predictions were close to the ratings observed. This is consistent with the weak version of the hypothesis that the models are predicting something of the perceived affective content of the excerpts, even if the scale may be different.

Direct comparisons were also made between the different affective dimensions for each excerpt and the highest and lowest predictions and ratings were discussed for every excerpt. However, recall that these ratings and predictions are based on
normalized data. The ratings for each affective dimension are therefore based on the means and the standard deviations of ratings for each scale. While it is interesting to compare affective dimensions across an excerpt in looking for accuracy of predictions, the interpretations of localized high and low affective dimensions is somewhat limited.

**Affective dimension effects**

A less problematic comparison would be to conduct a correlation test across affective dimension rather than across excerpt. If the models are accurately predicting the perceived surface-level affective content of an excerpt, but are scaled differently, then the rank order of excerpts on each affective dimension would be the same between the predictions and the ratings. In this way, only one affective model would be tested at a time. If the results are significant, then this would be consistent with the hypothesis that, given a number of different excerpts, the model would accurately be able to rank them for the given affective dimension.

After performing the test, the predictions for the rank ordering of excerpts for seven of the eleven affective dimensions were significantly positively correlated with the listener ratings for the rank order of those excerpts. Specifically, significantly positively correlated affective dimensions included *carefree* (+0.79, \( p = 0.00026 \)), *contentment* (+0.75, \( p = 0.00064 \)), *dark* (+0.82, \( p = 0.000092 \)), *happy/joyful* (+0.82, \( p = .000088 \)), *suspense/anticipation* (+0.78, \( p = 0.00033 \)), *unsettled/anxious* (+0.64, \( p = 0.0053 \)), and *weighty* (+0.76, \( p = 0.00050 \)). Although skewed in the predicted direction, no significant correlations were
found for calm/serene (+0.25, \( p = 0.19 \)), lonely (+0.21, \( p = 0.23 \)), sad/depressed/tragic (+0.29, \( p = 0.15 \)), or striving/yearning (+0.43, \( p = 0.057 \)).

**Conclusion**

In this study, a model of perceived surface-level affective content was built for short musical excerpts in the *Pathétique* sonata, second movement by regressing sixteen musical predictors onto the participant ratings from the main study. Despite the plentiful potential sources of error, the models built were able to account for between 20.0% to 36.2% of the variance in participant responses for excerpts from the movement. A unique model was built for each of the eleven affective dimensions tested in this study.

Following the construction of the models, an experiment was described in which excerpts solicited from music theorists and pianists were tested against the models. The variability in the 112 excerpts used in the building of the models was limited, due to the lack of data independence that comes when all of the excerpts used are from the same movement. Nevertheless, a study was conducted to test the applicability of the model to other excerpts from different Beethoven sonata movements. It was found that, of the eleven excerpts in which the affective content was described by the responder, seven of the descriptions were in line with professional musicians’ open-ended descriptions of the affective content.

Following this, the fifteen provided excerpts were played for participants who rated the perceived surface-level affective content of the excerpts along the affective
dimensions in the models. The predictions performed fairly well, although performance decreased with stronger affective ratings. The predictions for some excerpts, perhaps closer in style and affective content to the *Pathétique*, were much more accurate in comparison with participant responses than others. Even in the case where the prediction did not fall within the 95% confidence interval around the rating mean, in many cases the rating was close or the pattern of predictions was significantly correlated with participant ratings.

Finally, the different affective dimension models were tested to determine if the predictions of the various excerpts along each affective scale were consistent with participant ratings. For seven of the eleven affective models tested, the correlations between the rank-order of excerpts of the predictions were significantly correlated with participant ratings.

While the models built seem to a certain extent to be generalizable to other excerpts in other movements, it seems that more work is needed to improve the accuracy of the models. It is difficult to find significant results with very few observations, and so more excerpts should be tested. Conversely, a more generalizable model could be built that begins with excerpts from a broad range of works. The larger variability in musical features should serve to strengthen the model even further to be more accurate for excerpts with musical features that go beyond the ranges of the *Pathétique* sonata.
Chapter 8: Coda

The investigation of the relationship between composed music and its affective content is an old and venerable pursuit, and continues to this day. There are many ways in which this relationship can be investigated. Within the past century, several different approaches have been used to try to discover the mechanisms by which music can evoke emotion in listeners or by which music is recognized as portraying emotion. Each approach has its own unique sets of strengths and weaknesses, and to date no methodology has proved entirely satisfactory in its explanatory power or investigative approach.

In this dissertation, the progressive exposure method, a new approach in investigating the relationship between music and its perceived affective content, has been described, tested, and discussed, using Beethoven’s *Pathétique* sonata, second movement (No. 8, Op. 13) as a model for affective analysis. The method employs the strategy of dividing up long passages or works into small, discrete excerpts that fade-in and fade-out. The method can be used for either a *diachronic* presentation, in which the order of the excerpts preserves the order of the original work, or a *mosaic* presentation, which presents the excerpts in random order. Since the intent of this study was to focus on surface-level affect, the mosaic method was employed to study Beethoven’s
Pathétique sonata, second movement. Participants listen to the excerpts, and rate the perceived amount of each tested affective dimension for each excerpt. After the data is collected, the individual snapshots portraying the perceived affective content of each excerpt are amalgamated into a mosaic portrait that maps how each affective dimension changes throughout a work.

In part, the method was suggested by methodological difficulties with the existing paradigms. Specifically, the retrospective response paradigm does not permit the ability to track how the affective content of the movement changes or develops over time. The continuous data collection paradigm, while tracking changes in affect over the entire passage and providing a listening experience with more ecological validity, exhibits low reliability within and between participant trials and results in data that is very difficult to interpret. The progressive exposure method utilizes the strategy of allowing participants to listen to discrete excerpts and introspect as long as deemed necessary, resulting in cleaner and more reliable data. However, dividing one long passage into several small excerpts also provides the opportunity to investigate how affects change and develop over the course of the passage.

Many current studies that measure emotion in music tend to use one of three categorizations of emotion. One approach, based on the empirical work of Ekman (1972), identifies four to six emotions that are considered somehow “basic.” Common choices for these emotions are anger, disgust, fear, happiness, sadness, and surprise. While there is evidence that these emotions are facially expressible and recognizable, it is not clear how relevant these emotions are to the expression of music. An alternate
approach, often associated with Russell’s “circumplex model” (1980), identifies purportedly orthogonal affective dimensions, such as arousal and valence, in which any affect can theoretically be plotted. While this approach is theoretically very versatile, it is also rather abstract, and it is not clear how well the dimensions correspond to how participants actually conceive of and rate perceived affects in music. Finally, the GEMS model, developed by Zentner, et al. (2008), used a bottom-up empirical approach to reduce a list of 515 affective terms to sixty-six that were explicitly appropriate for music. However, the list was constructed to be applicable to a wide variety of music. Therefore, although GEMS has expanded the list of possible affective terms, especially compared with Ekman’s basic emotions and the valence/arousal model, it is unlikely that this list has managed to identify all pertinent emotions related to music. Even if there existed a list of (for example) 200 emotion terms, one might well imagine a piece of music for which none of the listed terms provided an apt description.

For the purposes of the current research, an initial study was conducted to ascertain what affective dimensions would be appropriate for an analysis of Beethoven’s *Pathétique* sonata, second movement. Five experienced musicians were recruited to listen to fifteen representative excerpts from the movement and freely describe what emotions they thought the movement was expressing or conveying. An informal content analysis was conducted on the 592 discrete comments the participants provided in response to the fifteen excerpts they listened to. Fifteen affective dimensions arose from this analysis for use with the main study: calm/serene, carefree, cheeky/sassy, contentment, dark, emotional/moody, happy/joyful, important/serious, lonely, sad/
depressed/tragic, sincerity/truthful, striving/yearning, suspense/anticipation, weighty, and unsettled/anxious.

In the main study, 110 participants from Ohio State University and Westminster Choir College were recruited to listen to the movement using the progressive exposure method. In order to attempt to mitigate the effects of musically arbitrary excerpt boundaries, two offset groups were created – 0” and 2.5” offset groups – that overlapped one another by 1.25” on either side of each excerpt. Participants were randomly assigned to one of the two offset groups and to one of five different affect groups that rated three affective dimensions throughout the study. They listened to each of the 56 excerpts in their assigned offset group and rated each of the three affective dimensions in their assigned affect group for each excerpt.

Additionally, each participant rated five repeated excerpts for each of the affective dimensions in their affect group. This data was collected to be able to investigate the intrasubjective reliability of these affective dimensions. Typically, intrasubjective reliability criteria are used to eliminate unreliable participants. In the case of this study, however, the affective scales used had not previously been tested for reliability. Therefore, data sets were eliminated by subject-scale and not by participant, because even diligent participants could have had trouble with affective dimensions that were not reliable. Of the 330 subject-scales tested, 76 failed to meet the exclusion criterion of an intrasubjective correlation of +0.25, and so were eliminated. Some affective scales were more heavily represented in this set of 76 subject-scales. Specifically, dark, emotional/moody, and sincerity/truthful were most highly
represented, with nine of the 22 subject-scales in each category being eliminated.

Cheeky/sassy was discarded from further study, because one intrasubjective correlation was exactly 0 and three were exactly 1, suggesting that it was not an appropriate affective dimension for this movement.

In order to further test the reliability of the scales, the intersubjective reliability of these scales was also tested. The 56 judgments for each subject-scale not eliminated due to low intrasubjective reliability were correlated with each of the other comparable sets, and subject-scales were eliminated that failed to meet the exclusion criterion of an average correlation of +0.2 with the comparable sets. Eighteen additional subject-scales were eliminated; eight of these were sincerity/truthful, three were important/serious, and two were emotional/moody. Sincerity/truthful, emotional/moody, and important/serious were also discarded from further consideration because these dimensions had a large number of unreliable subject-scales eliminated.

The remaining eleven affective dimensions all exhibited a high degree of both intra- and intersubjective reliability. The highest mean intrasubjective correlation for these eleven dimensions was weighty (mean correlation of +0.814), and the lowest mean intrasubjective correlation was sad/depressed/tragic (mean correlation of +0.558). The highest mean intersubjective correlation for these eleven dimensions was calm/serene, (mean correlation of +0.661), and the lowest mean intersubjective correlation was striving/yearning (mean correlation of +0.461). These results are consistent with the idea that the progressive exposure method produces reliable and consistent ratings of
the perceived surface affective content of music, both in repeated trials of the same participant and across participants.

Following the investigation of reliability, the participant data was amalgamated to form a mosaic portrait of each affective dimension. The results from this amalgamation are summarized in Appendix A. Several passages were highlighted for affective analysis, and a discussion of the relationship between ratings of perceived affect and the musical features followed. Additionally, the five different statements of the theme were compared with one another, to determine if there were effects of statement. Each of the five statements were similar, although there were differences between all of the statements except for the first and third. By collapsing responses for each affective dimension across all excerpts within each statement, effects of the differences between passages were analyzed. Specifically, there was an effect of higher register for carefree, dark, happy/joyful, and weighty, an effect of faster surface rhythms for sad/depressed/tragic and unsettled/anxious, and an interaction effect of rhythm and register for lonely, sad/depressed/tragic, and weighty. The overall tendency throughout the statements of the theme in the movement is for a progressive move toward more positively-valenced, higher-arousal affects, and toward less negatively-valence, lower-arousal affects.

Additionally, passages with “extreme” ratings of more than one standard deviation from the mean were investigated. It was found a posteriori that excerpts in refrain passages had 9.17 extreme ratings per second, whereas excerpts in episode or coda passages had 16.41 extreme ratings per second. These results are consistent with
the hypothesis, originally advanced by Leonard Meyer (1957) and reiterated by Eugene Narmour (1990, 1992) and David Huron (2006) that less musically predictable formal sections – in this case episode and coda sections in Rondo form – evoke stronger affective evaluations by listeners. Additionally, the five-second excerpts that evoked the largest number of extreme ratings were investigated, and it was found that the four excerpts with the largest number of extreme ratings were ms. 23-24 and ms. 48-49, the two retransitions to the refrain sections. Retransitions are often described as the moments of the most musical tension in a movement, as the chord of continuation is most strongly expected at these moments.

The affective dimensions that had the highest and lowest mean ratings across the entire movement were also investigated. The four affective dimensions with the highest mean ratings were calm, contentment, striving/yearning, and suspense/anticipation. While the juxtaposition of these affective dimensions seems counterintuitive, these four affective dimensions were all represented in the theoretical literature describing the emotional content of the movement. It may be that the juxtaposition of conflicting emotional characters is what contributes to the commonly described emotionally powerful effect of the movement.

In order to investigate complex mixes of affect, a correlation table was calculated between all of the tested affective dimensions. It was found that unsettled/anxious was highly negatively correlated with calm/serene and contentment, but highly correlated with suspense/anticipation. Sad/depressed/tragic was highly positively correlated with lonely, happy/joyful was highly positively correlated with carefree, and calm/
serene was highly positively correlated with contentment. Dark was highly negatively correlated with carefree, contentment, and happy/joyful, and suspense/anticipation was highly negatively correlated with contentment.

In order to determine how surface musical features were influencing ratings, sixteen musical features were measured for each of the 112 excerpts used in the study. These included such features as number of harmonies, highest and lowest pitch, presence of dissonance, mode, and articulations. A multiple regression analysis was conducted using these sixteen musical features as predictors of the eleven affective dimensions. The result was a formula for each affective dimension, using between six and nine predictors. The amount of the variance accounted for by these models ranged from 20.0% and 36.2%.

A study was then conducted to determine if these models were generalizable to other excerpts from Beethoven piano sonatas. Professional music theorists from the Society for Music Theory listserv and professional piano performers and educators from the Piano Street online forum were polled for the most emotionally expressive Beethoven piano sonata moments. The result was fifteen excerpts, eleven of which were described by the responders with affective descriptions. Each of the same sixteen musical features were measured for each of the excerpts, and predictions were made for each of the eleven affective dimensions for each excerpt.

In order to translate the affective descriptions into the eleven terms used in the models, a study was conducted in which participants “translated” the descriptions by providing a rating of -1, 0, or 1 for each term used in the models. A Pearson’s correlation
test was conducted to determine if the predictions made by the model were significantly correlated with the translations given by participants. Of the eleven descriptions provided by responders, seven of the excerpts were significantly correlated with the predictions from the model.

The predictions were further tested by playing each five-second excerpt to participants and asking them to rate the excerpts on the eleven affective dimensions corresponding to the eleven models. The resulting participant ratings were compared to the predictions made by the models. Although many of the predictions failed to fall within the 95% confidence intervals for the participant ratings, most of the predictions exhibited significant correlations with the ratings. This may result from a difference in scale; because the model was making predictions based on means of excerpts from the *Pathétique* and participants’ scores were based on means of the excerpts they heard, there may be a difference in scale. Nevertheless, the significant correlations are consistent with the model explaining something of the affective content of the excerpts, even if measured with a different scale.

The series of studies presented in this dissertation chronicled an attempt to approach perceived surface affective content in a novel way. The progressive exposure method provides a new way of measuring perceived surface affective content of complete musical works or movements that provides relatively clean data that is straightforward to interpret. It also provides the opportunity of assembling snapshots of a long work together to observe differences in perceived surface affective content over the course of the work.
It is difficult to determine exactly how large the effect is of removing context from these analyses. Further work could investigate this difference by playing selected excerpts used in this study with all of its preceding context. Perhaps a light could come on that would tell the participant to begin evaluating the affective content of the passage. By comparing the results from this study with that future study, the bottom-up effects of low-level musical features could be factored out, leaving just the effects of large-scale context.

Another future study could build a model truly intended for the purpose of generalization. By sampling broadly across all of Beethoven’s piano sonatas, a model could be constructed that should be measured in a more generalizable way. This model would be more aptly suited to making exact predictions in other sampled excerpts from Beethoven’s piano sonatas.

The current study used interleaved 2.5” excerpts across Beethoven’s *Pathétique*. While the resolution of such an approach was fairly high, a more precise approach could discover the effects of each consecutive onset event. Instead of using excerpt windows according to elapsed time, windows could be used that include new onset events. The result would be direct comparisons between excerpts that share much more in common. Similarities could be factored out leaving a residual of how each new onset event effects the music’s perceived affective content. In this way, a more precise affective analysis could be carried out. Much more work remains to be done. Nevertheless, the results from this study offer promise for the value that using the progressive exposure method could have for future analyses of the perceived affective content of music.
Selected Bibliography


James, William. (1884). What is an emotion? Mind, 9, 188-205.


Appendix A: Complete Surface-Level Affective Analysis of the Second Movement of Beethoven’s *Pathétique* Sonata Using the Progressive Method

The following pages contain the complete surface-level affective analysis of Beethoven’s *Pathétique* sonata, second movement, obtained from the data collected from the main study. This data was obtained by using the progressive exposure method with the eleven affective dimensions for an affective analysis of the music’s surface.

The graph is meant to be read with both opposing pages simultaneously. The positively-valenced affective dimensions are printed above the music on the left-hand page, whereas the negatively-valenced affective dimensions are printed below the music on the right-hand page. Affective dimensions increase in arousal with greater distance from the music.

The data from both offset-groups have been combined into a single graph. In order to better amalgamate data across participants, the data for each participant was first normalized across subject-scale. The dotted lines indicate the five-second excerpts heard by participants. The points between the dotted lines represent the averages of the evaluations of the excerpts between the dotted lines, judged by the 2.5”-offset group. The points on the dotted lines represent the averages of the evaluations of the excerpts centered on the dotted lines, judged by the 0”-offset groups. In other words, there was a 1.25” overlap with the other offset group at the beginning and end of every stimulus.
Appendix B: Anova Summary Tables for Treatment Effect of Different Statements of the Theme by Affective Dimension

The following pages show the results of the ANOVA analyses for each affective dimension that were conducted to investigate the effect of theme “treatment” on perceived affective content. ANOVA summaries are first provided in a table for each affective dimension, followed by a plot of the means for each statement of a theme (with a 90% confidence interval), and finally a table showing Tukey’s HSD between the first statement of the theme and the other statements, and between the fourth and fifth statement of the theme. Family-wise $\alpha$ level was set to 0.1, and so the corrected $\alpha$ was 0.007.

The first statement of the theme presents the theme in its original form. The second statement of the theme adds an extra inner voice and transposes the melody up an octave. The third statement of the theme is exactly the same as the first statement, offering a control group. The fourth statement of the theme presents the melody with a faster sixteenth-triplet accompaniment. The fifth statement of the theme is the same as the fourth statement of the theme transposed up an octave. The fifth statement of the theme also ends on a perfect authentic cadence on the downbeat rather than with a suspension figure.
Table B.1. ANOVA summary table for Calm/Serene ($p < 0.007$).

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>21.103</td>
<td>5.2758</td>
<td>9.8807</td>
<td>1.137e-07*</td>
</tr>
<tr>
<td>Residuals</td>
<td>451</td>
<td>240.811</td>
<td>0.5339</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure B.1. The ANOVA plot for calm/serene. The biggest effect seems to come with the increased rhythmic activity of the triplets.

Table B.2. Tukey’s HSD for calm/serene between the first theme and the other themes and between the fourth and fifth themes. The themes played with the triplet rhythms are significantly lower in calm/serene at $p < 0.1$.

<table>
<thead>
<tr>
<th>Calm/Serene</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>-0.01467</td>
<td>-0.05132</td>
<td>-0.31922*</td>
<td>-0.55440*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td></td>
<td>-0.23518</td>
</tr>
</tbody>
</table>
### Table B.3. ANOVA summary table for Carefree (p < 0.007).

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>37.39</td>
<td>9.3482</td>
<td>15.864</td>
<td>2.777e-12*</td>
</tr>
<tr>
<td>Residuals</td>
<td>537</td>
<td>316.43</td>
<td>0.5893</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure B.2.** The ANOVA plot for carefree. There is a significant effect for higher register and a significant effect for faster rhythms.

### Table B.4. Tukey’s HSD for carefree between the first theme and the other themes and between the fourth and fifth themes. The themes played with the triplet rhythms are significantly higher in carefree than their slower-rhythm counterparts and the themes played up the octave are significantly higher in carefree at p < 0.1.

<table>
<thead>
<tr>
<th></th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>-0.01792</td>
<td>0.24500*</td>
<td>0.19413</td>
<td>0.71496*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td></td>
<td>0.52083*</td>
</tr>
</tbody>
</table>

The themes played with the triplet rhythms are significantly higher in carefree than their slower-rhythm counterparts and the themes played up the octave are significantly higher in carefree at p < 0.1.
Table B.5. ANOVA summary table for Contentment ($p < 0.007$).

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>13.693</td>
<td>3.4233</td>
<td>6.4051</td>
<td>4.802e-05*</td>
</tr>
<tr>
<td>Residuals</td>
<td>565</td>
<td>301.968</td>
<td>0.5345</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure B.3.** The ANOVA plot for contentment. The faster rhythms are significantly lower in carefree.

Table B.6. Tukey’s HSD for contentment. The statements of the theme with faster rhythms are significantly lower than the slower rhythms at $p < 0.1$.
Table B.7. ANOVA summary table for Dark ($p < 0.007$).

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>19.118</td>
<td>4.7796</td>
<td>8.8064</td>
<td>8.508e-07*</td>
</tr>
<tr>
<td>Residuals</td>
<td>365</td>
<td>198.098</td>
<td>0.5427</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure B.4. The ANOVA plot for dark. The biggest effect is seen in the higher register, which is significantly lower for dark.

Table B.8. Tukey’s HSD for dark. The effect of the higher octave is to significantly lower dark ratings at $p < 0.1$. 

<table>
<thead>
<tr>
<th></th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>0.06589</td>
<td>-.46065*</td>
<td>-0.08126</td>
<td>-0.45241*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td></td>
<td>-0.37115*</td>
</tr>
</tbody>
</table>
### Happy/Joyful

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>36.528</td>
<td>9.1319</td>
<td>15.244</td>
<td>1.065e-11*</td>
</tr>
<tr>
<td>Residuals</td>
<td>450</td>
<td>269.572</td>
<td>0.5990</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table B.9.** ANOVA summary table for Happy/Joyful ($p < 0.007$).

**Figure B.5.** The ANOVA plot for happy/joyful. The biggest effect is for taking the melody up an octave, though the faster rhythmic values also significantly raise the ratings for happy/joyful.

<table>
<thead>
<tr>
<th>Happy/Joyful</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>-0.00314</td>
<td>0.46854*</td>
<td>0.33011*</td>
<td>0.74307*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td>0.33011*</td>
<td>0.41296*</td>
</tr>
</tbody>
</table>

**Table B.10.** Tukey’s HSD for happy/joyful, showing significant effects for both the higher register and faster rhythmic values at $p < 0.1$. 

211
Table B.11. ANOVA summary table for Lonely ($p < 0.007$).

<table>
<thead>
<tr>
<th>Lonely</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>74.635</td>
<td>18.6588</td>
<td>34.789</td>
<td>&lt; 2.2 e-16*</td>
</tr>
<tr>
<td>Residuals</td>
<td>451</td>
<td>241.892</td>
<td>0.5363</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure B.6. The ANOVA plot for lonely. There appears to be an interaction between rhythm and register on lonely. There is no significant difference for register unless the rhythm is faster.

Table B.12. Tukey’s HSD for lonely, showing significant effects for both the higher register and faster rhythmic values at $p < 0.1$.  

<table>
<thead>
<tr>
<th>Lonely</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>0.04957</td>
<td>-0.17757</td>
<td>-0.68196*</td>
<td>-0.98939*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td>-0.68196*</td>
<td>-0.30742*</td>
</tr>
</tbody>
</table>
Table B.13. ANOVA summary table for Sad/Depressed/Tragic ($p < 0.007$).

<table>
<thead>
<tr>
<th>Sad/Depressed/Tragic</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>59.669</td>
<td>14.9172</td>
<td>26.698</td>
<td>&lt; 2.2 e-16*</td>
</tr>
<tr>
<td>Residuals</td>
<td>423</td>
<td>236.346</td>
<td>0.5587</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure B.7. The ANOVA plot for sad/depressed/tragic. There appears to be an interaction between rhythm and register. There is no significant difference for register unless the rhythm is faster.

Table B.14. Tukey’s HSD for sad/depressed/tragic, showing significant effects for both the higher register and faster rhythmic values at $p < 0.1$. 

<table>
<thead>
<tr>
<th>Sad/Depressed/Tragic</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>0.01289</td>
<td>-0.15190</td>
<td>-0.50706*</td>
<td>-0.97668*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td></td>
<td>-0.46962*</td>
</tr>
<tr>
<td>Striving/Yearning</td>
<td>Df</td>
<td>Sum Sq</td>
<td>Mean Sq</td>
<td>F value</td>
</tr>
<tr>
<td>------------------</td>
<td>----</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>treatment</td>
<td>4</td>
<td>2.24</td>
<td>0.55987</td>
<td>0.8049</td>
</tr>
<tr>
<td>Residuals</td>
<td>508</td>
<td>353.34</td>
<td>0.69555</td>
<td></td>
</tr>
</tbody>
</table>

Table B.15. ANOVA summary table for Striving/Yearning ($p = 0.5224$).

![ANOVA results for Striving/Yearning with 90% CI](image)

Figure B.8. The ANOVA plot for striving/yearning. There are no significant differences between the different themes.

<table>
<thead>
<tr>
<th>Striving/Yearning</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>0.01667</td>
<td>0.11675</td>
<td>0.03692</td>
<td>-0.08830</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td></td>
<td>-0.12522</td>
</tr>
</tbody>
</table>

Table B.16. Tukey’s HSD for striving/yearning. There are no significant differences at $p < 0.1$. 

214
<table>
<thead>
<tr>
<th>Suspense/Anticipation</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>2.68</td>
<td>0.66932</td>
<td>0.9074</td>
<td>0.4593</td>
</tr>
<tr>
<td>Residuals</td>
<td>509</td>
<td>375.44</td>
<td>0.73760</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.17. ANOVA summary table for Suspense/Anticipation ($p > 0.007$).

![ANOVA results for Suspense/Anticipation with 90% CI](image)

Figure B.9. The ANOVA plot for suspense/anticipation. There are no significant differences between the different themes.

<table>
<thead>
<tr>
<th>Suspense/Anticipation</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>0.05749</td>
<td>-0.01516</td>
<td>0.13994</td>
<td>-0.0771</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td>0</td>
<td>-0.2170</td>
</tr>
</tbody>
</table>

Table B.18. Tukey’s HSD for suspense/anticipation. There are no significant differences at $p < 0.1$.  

215
### Table B.19. ANOVA summary table for Unsettled / Anxious ($p < 0.007$).

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>22.398</td>
<td>5.5996</td>
<td>9.7852</td>
<td>1.244e-07*</td>
</tr>
<tr>
<td>Residuals</td>
<td>508</td>
<td>290.703</td>
<td>0.5722</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure B.10.** The ANOVA plot for unsettled/anxious. Faster rhythmic values significantly increase the level of unsettled/anxious.

### Table B.20. Tukey’s HSD for unsettled/anxious. There is a significant difference for faster rhythmic values at $p < 0.1$.

<table>
<thead>
<tr>
<th>Unsettled/Anxious</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>0.06019</td>
<td>0.17905</td>
<td>0.45885*</td>
<td>0.52532*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td>0.06647</td>
<td></td>
</tr>
</tbody>
</table>
### Table B.21. ANOVA summary table for Weighty ($p < 0.007$).

<table>
<thead>
<tr>
<th>Weighty</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>4</td>
<td>58.23</td>
<td>14.5576</td>
<td>24.292</td>
<td>&lt; 2.2 e-16*</td>
</tr>
<tr>
<td>Residuals</td>
<td>536</td>
<td>321.21</td>
<td>5993</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure B.11.** The ANOVA plot for weighty. There is no effect of faster rhythms. The higher register, however, significantly reduces weighty ratings.

### Table B.22. Tukey’s HSD for weighty. The melody taken up an octave is significantly lower in weighty ratings at $p < 0.1$.

<table>
<thead>
<tr>
<th>Weighty</th>
<th>3rd</th>
<th>2nd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>0.12711</td>
<td>-0.45807*</td>
<td>0.11172</td>
<td>-0.68344*</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td></td>
<td>-0.79516*</td>
</tr>
</tbody>
</table>
Appendix C: The Distributions of Ratings for Each Affective Dimension

The following graphs show the histograms for the ratings for each affective dimension across Beethoven’s *Pathétique* sonata, second movement. There are two plots for each affective dimension. The plot on the left shows the raw scores, whereas the plot on the right shows the normalized scores by subject-scale. Bimodal distributions can be seen in many of the raw score plots. This is probably an artifact of the nature of the interface. The slider began by being centered in the middle of the scale. The spike in the middle of the graphs corresponds to participants who leave the slider in the middle. For many of these affective dimensions there is a localized peak both in the upper half and the lower half, representing a mean of participant ratings for that subset of the ratings.
Figure C.1. The result distribution for calm/serene. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.2. The result distribution for carefree. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.3. The result distribution for contentment. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.4. The result distribution for dark. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.5. The result distribution for happy/joyful. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.6. The result distribution for lonely. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.7. The result distribution for sad/depressed/tragic. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.8. The result distribution for striving/yearning. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.9. The result distribution for suspense/anticipation. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.10. The result distribution for unsettled/anxious. The left graph shows the raw scores, and the right graph shows the normalized scores.
Figure C.11. The result distribution for weighty. The left graph shows the raw scores, and the right graph shows the normalized scores.
Appendix D: The Regression Summaries for the Eleven Affective Dimensions Tested in the Main Study

The following tables show the regression equations for each affective dimension tested in the main study. The left columns of the table provide an estimate for the coefficient terms for each predictor variable, and the right columns provide a significance test for the variable. Finally, the bottom shows the $R^2$ and adjusted $R^2$ values, indicative of the amount of variance the model is able to account for in the listener responses.

### Calm/Serene

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.38556</td>
<td>-0.53900</td>
<td>0.06615</td>
<td>0.51759</td>
<td>3.10890</td>
</tr>
</tbody>
</table>

Coefficients:

|         | Estimate | Std. Error | t value | Pr(>|t|) |
|---------|----------|------------|---------|----------|
| (Intercept) | 0.406709 | 0.120126   | 3.386   | 0.000741 *** |
| artic     | -0.420454 | 0.052139  | -8.064  | 2.38e-15 *** |
| mode      | 0.249618  | 0.040149   | 6.217   | 7.78e-10 *** |
| Htemp     | 0.326527  | 0.030793   | 10.604  | < 2e-16 *** |
| hpit      | -0.025400 | 0.003124   | -8.131  | 1.43e-15 *** |
| cres      | -0.135493 | 0.038300   | -3.538  | 0.000425 *** |
| close     | 0.086664  | 0.033118   | 2.617   | 0.009027 **  |
| dis       | -0.327718 | 0.065684   | -4.989  | 7.29e-07 *** |
| surp      | -0.068984 | 0.015294   | -4.511  | 7.34e-06 *** |
| tend      | -0.310629 | 0.105059   | -2.957  | 0.003192 **  |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7834 on 886 degrees of freedom
Multiple R-squared: 0.368, Adjusted R-squared: 0.3616
F-statistic: 57.32 on 9 and 886 DF, p-value: < 2.2e-16

Table D.1. The regression table for calm/serene.
### Carefree

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>Median</th>
<th>30</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-2.48124</td>
<td>-0.64443</td>
<td>0.02013</td>
<td>0.63774</td>
</tr>
<tr>
<td>Max</td>
<td>2.47001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients:

|                                | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------------------|----------|------------|---------|----------|
| (Intercept)                  | -0.591247| 0.155417   | -3.804 | 0.000150 ***|
| mode                         | 0.487224 | 0.041849   | 11.642 | < 2e-16 ***|
| hpi                          | 0.017452 | 0.003433   | 5.083  | 4.39e-07 ***|
| cres                         | -0.185314| 0.038761   | -4.781 | 1.99e-06 ***|
| speed                        | 0.162732 | 0.036530   | 4.455  | 2.62e-06 ***|
| artic                       | -0.242221| 0.051271   | -4.724 | 0.0001222 **|
| lpi                          | 0.014045 | 0.004332   | 3.243  | 0.001222 **|
| close                        | 0.212359 | 0.039294   | 6.454  | 8.51e-11 ***|
| dis                          | -0.171566| 0.059189   | -2.899 | 0.003826 **|
| Htemp                        | 0.077361 | 0.009523   | 2.636  | 0.00523 **  |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8564 on 1054 degrees of freedom
Multiple R-squared: 0.2485,  Adjusted R-squared: 0.2421
F-statistic: 38.73 on 9 and 1054 DF,  p-value: < 2.2e-16

**Table D.2.** The regression table for carefree.

### Contentment

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>Median</th>
<th>30</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-2.45545</td>
<td>-0.55625</td>
<td>0.05711</td>
<td>0.55377</td>
</tr>
<tr>
<td>Max</td>
<td>2.10375</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients:

|                                | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------------------|----------|------------|---------|----------|
| (Intercept)                  | 0.22299  | 0.11676    | 1.910   | 0.0564   |
| mode                         | 0.44774  | 0.03634    | 12.319  | < 2e-16 ***|
| artic                       | -0.40082 | 0.04635   | -8.647  | < 2e-16 ***|
| Htemp                        | 0.29491  | 0.02603    | 11.330  | < 2e-16 ***|
| tend                        | -0.43070 | 0.09727    | -4.428  | 1.04e-05 ***|
| speed                       | -0.10341 | 0.03322    | -3.113  | 0.0019 **|
| dis                          | -0.25815 | 0.05940    | -4.346  | 1.51e-05 ***|
| close                        | 0.19517  | 0.02978    | 6.555   | 8.51e-11 ***|

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8028 on 1112 degrees of freedom
Multiple R-squared: 0.3419,  Adjusted R-squared: 0.3378
F-statistic: 82.54 on 7 and 1112 DF,  p-value: < 2.2e-16

**Table D.3.** The regression table for contentment.
### Dark

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.64734</td>
<td>-0.55925</td>
<td>-0.01183</td>
<td>0.61236</td>
<td>2.30233</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | -0.21013 | 0.22082 | -0.952  | 0.341631 |
| mode | -0.44209 | 0.04704 | -9.398  | < 2e-16 *** |
| hpit | -0.03049 | 0.00389 | -7.837  | 1.66e-14 *** |
| artic | 0.26062 | 0.05302 | 4.916   | 1.10e-06 *** |
| tempo | 0.28932 | 0.10323 | 2.803   | 0.005205 ** |
| tend | 0.28084 | 0.12023 | 2.336   | 0.019781 * |
| harmS | -0.20773 | 0.08099 | -2.565  | 0.010523 * |
| dis | 0.18707 | 0.07918 | 2.362   | 0.018421 * |
| close | -0.13336 | 0.03838 | -3.475  | 0.000542 *** |
| Htemp | -0.18037 | 0.03673 | -4.911  | 1.12e-06 *** |

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.806 on 718 degrees of freedom
Multiple R-squared: 0.3272, Adjusted R-squared: 0.3187
F-statistic: 38.79 on 9 and 718 DF, p-value: < 2.2e-16

Table D.4. The regression table for dark.

### Happy/Joyful

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.768818</td>
<td>-0.593772</td>
<td>0.001321</td>
<td>0.576114</td>
<td>2.650941</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | -0.873708 | 0.114237 | -7.648  | 5.29e-14 *** |
| mode | 0.485557 | 0.042726 | 11.364  | < 2e-16 *** |
| hpit | 0.033477 | 0.003341 | 10.021  | < 2e-16 *** |
| artic | -0.093481 | 0.049530 | -1.887  | 0.05944 |
| Htemp | 0.097498 | 0.030927 | 3.153   | 0.00167 ** |
| close | 0.083510 | 0.034867 | 2.395   | 0.01682 * |
| cres | -0.060168 | 0.041398 | -1.453  | 0.14646 |

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8534 on 889 degrees of freedom
Multiple R-squared: 0.254, Adjusted R-squared: 0.249
F-statistic: 50.44 on 6 and 889 DF, p-value: < 2.2e-16

Table D.5. The regression table for happy/joyful.
**Lonely**

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>10</th>
<th>Median</th>
<th>30</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.58842</td>
<td>-0.58723</td>
<td>0.06976</td>
<td>0.59543</td>
<td>2.16778</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 1.515150   | 0.248244 | 6.103 1.55e-09 *** |
| speed     | -0.324606  | 0.033323 | -9.741 < 2e-16 *** |
| dense     | -0.171538  | 0.038286 | -4.480 8.42e-06 *** |
| close     | -0.270493  | 0.032631 | -8.290 4.17e-16 *** |
| mode      | -0.298505  | 0.042298 | -7.057 3.42e-12 *** |
| hpit      | -0.018898  | 0.003668 | -5.152 3.17e-07 *** |
| tempo     | 0.358148   | 0.089962 | 3.981 7.42e-05 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8154 on 889 degrees of freedom
Multiple R-squared: 0.318, Adjusted R-squared: 0.3134
F-statistic: 69.08 on 6 and 889 DF,  p-value: < 2.2e-16

**Sad/Depressed/Tragic**

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.4358</td>
<td>-0.6147</td>
<td>0.1052</td>
<td>0.6204</td>
<td>2.7265</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.490693   | 0.268875 | 1.825 0.068362 . |
| close     | -0.269143  | 0.034661 | -7.765 2.40e-14 *** |
| mode      | -0.293242  | 0.045108 | -6.501 1.38e-10 *** |
| dense     | -0.150409  | 0.040444 | -3.719 0.000214 *** |
| surp      | -0.109428  | 0.014591 | -7.500 1.64e-13 *** |
| tempo     | 0.342682   | 0.094934 | 3.610 0.000325 *** |
| tend      | 0.215267   | 0.096087 | 2.240 0.025333 * |
| hpit      | -0.023323  | 0.003837 | -6.079 1.84e-09 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8316 on 832 degrees of freedom
Multiple R-squared: 0.2874, Adjusted R-squared: 0.2814
F-statistic: 47.93 on 7 and 832 DF,  p-value: < 2.2e-16

**Table D.6.** The regression table for lonely.

**Table D.7.** The regression table for sad/depressed/tragic.
### Striving/Yearning

Residuals:

Min       1Q   Median       3Q      Max  
-2.40288 -0.64531  0.09752  0.65748  2.26979

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | -0.449029 | 0.168970 | -2.657  0.00800 ** |
| close     | -0.345817 | 0.032569 | -10.618 < 2e-16 *** |
| mode      | -0.307449 | 0.043345 | -7.093  2.48e-12 *** |
| cres      | 0.182840  | 0.040876 | 4.473  8.59e-06 *** |
| dyn       | 0.068860  | 0.034400 | 2.002  0.04558 * |
| tend      | 0.396468  | 0.092853 | 4.270  2.14e-05 *** |
| lpit      | -0.012201 | 0.004145 | -2.943  0.00332 ** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.884 on 1001 degrees of freedom  
Multiple R-squared: 0.2051,  Adjusted R-squared: 0.2003  
F-statistic: 43.04 on 6 and 1001 DF,  p-value: < 2.2e-16

**Table D.8.** The regression table for striving/yearning.

### Suspense/Anticipation

Residuals:

Min       1Q   Median       3Q      Max  
-2.61074 -0.59045  0.08873  0.59803  2.72032

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | -1.22276 | 0.22195 | -5.509  4.59e-08 *** |
| mode      | -0.29313  | 0.04210 | -6.963  6.01e-12 *** |
| tend      | 0.75793   | 0.10442 | 7.258  7.87e-13 *** |
| artic     | 0.31564   | 0.04455 | 7.085  2.62e-12 *** |
| tempo     | 0.36878   | 0.08068 | 4.571  5.47e-06 *** |
| close     | -0.27068  | 0.03243 | -8.347  2.31e-16 *** |
| Htemp     | -0.27888  | 0.02906 | -9.597  < 2e-16 *** |
| dis       | 0.17946   | 0.06649 | 2.699  0.00707 ** |
| dense     | 0.09198   | 0.03390 | 2.713  0.00677 ** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8288 on 999 degrees of freedom  
Multiple R-squared: 0.2939,  Adjusted R-squared: 0.2883  
F-statistic: 51.98 on 8 and 999 DF,  p-value: < 2.2e-16

**Table D.9.** The regression table for suspense/anticipation.
### Unsettled/Anxious

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.20504</td>
<td>-0.55289</td>
<td>-0.05869</td>
<td>0.57574</td>
<td>2.39528</td>
</tr>
</tbody>
</table>

Coefficients:

|         | Estimate  | Std. Error | t value | Pr(>|t|) |
|---------|-----------|------------|---------|----------|
| (Intercept) | -0.059348 | 0.115036 | -0.516  | 0.606030 |
| hpit     | 0.015186  | 0.002992  | 5.076   | 4.59e-07 *** |
| cres     | 0.130287  | 0.036677  | 3.552   | 0.000400 *** |
| mode     | -0.425512 | 0.038448  | -11.067 | < 2e-16 *** |
| tend     | 0.502481  | 0.100607  | 4.995   | 6.96e-07 *** |
| close    | -0.185554 | 0.031715  | -5.851  | 6.63e-09 *** |
| artic    | 0.365671  | 0.049930  | 7.324   | 4.96e-13 *** |
| surp     | 0.050679  | 0.014646  | 3.460   | 0.000562 *** |
| Htemp    | -0.327139 | 0.029488  | -11.094 | < 2e-16 *** |
| dis      | 0.209610  | 0.062901  | 3.332   | 0.000893 *** |

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7957 on 998 degrees of freedom  
Multiple R-squared: 0.3588,  
Adjusted R-squared: 0.353  
F-statistic: 62.06 on 9 and 998 DF,  
p-value: < 2.2e-16

**Table D.10.** The regression table for unsettled/anxious.

### Weighty

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.48390</td>
<td>-0.63016</td>
<td>-0.01599</td>
<td>0.63789</td>
<td>2.82417</td>
</tr>
</tbody>
</table>

Coefficients:

|         | Estimate  | Std. Error | t value | Pr(>|t|) |
|---------|-----------|------------|---------|----------|
| (Intercept) | -0.986135 | 0.159373  | -6.188  | 8.73e-10 *** |
| hpit     | -0.030713 | 0.003393  | -9.052  | < 2e-16 *** |
| lpit     | -0.034691 | 0.004196  | -8.267  | 4.10e-16 *** |
| cres     | 0.251838  | 0.038409  | 6.557   | 8.59e-11 *** |
| close    | -0.106926 | 0.030904  | -3.460  | 0.000562 *** |
| dyn      | 0.121831  | 0.034821  | 3.499   | 0.000487 *** |
| tend     | 0.161227  | 0.098017  | 1.645   | 0.100290 |
| mode     | -0.188630 | 0.039995  | -4.716  | 2.72e-06 *** |
| dis      | 0.179576  | 0.065209  | 2.754   | 0.005991 ** |

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8456 on 1055 degrees of freedom  
Multiple R-squared: 0.262, Adjusted R-squared: 0.2564  
F-statistic: 46.82 on 8 and 1055 DF,  
p-value: < 2.2e-16

**Table D.11.** The regression table for weighty.