Activity Analysis and Coordination in Continuous Responses to Music

Finn Upham

McGill University, Montreal, QC, Canada and New York University, New York, NY

Stephen McAdams

McGill University, Montreal, QC, Canada

Running title: Activity Analysis

Abstract

Music affects us physically and emotionally. Determining when changes in these reactions tend to manifest themselves can help us understand how and why. Activity Analysis quantifies alignment of response events across listeners and listenings through continuous responses to musical works. Its coordination tests allow us to determine if there is enough inter-response coherence to merit linking their summary time series to the musical event structure and to identify moments of exceptional alignment in response events. In this paper, we apply Activity Analysis to continuous ratings from several music experiments, using this wealth of data to compare its performance with that of statistics used in previous studies. We compare the Coordination Scores and nonparametric measures of local activity coordination to other coherence measures, including those derived from correlations and Cronbach's α. Activity Analysis reveals the variation in coordination of participants' responses for different musical works, picks out moments of coordination in response to different interpretations of the same music, and demonstrates that responses along the two dimensions in continuous 2D rating tasks can be independent.

Key words: continuous responses to music, statistics, time series analysis, agreement, response coordination

Continuous response measures are a promising means of investigating our experience of music. By sampling aspects of participants' responses as they listen, these measures can capture the development of behavior, understanding, and feeling through time. However, interpreting these traces of a listener's response is not a straightforward task. Whether continuous ratings, skin conductance, or heart rate variability, these time series contain a great deal of information with ambiguous relationships to the stimulus presented. A range of techniques have been borrowed from disciplines adjacent to music cognition, each capturing different patterns in responses based on their own assumptions. This paper presents an analytical framework we call Activity Analysis that is specifically designed to explore and evaluate *coordination* in continuous responses to music, evaluating when responses of different listeners are reliably active at the same time.

Many studies have focused on evaluating the relationship between continuous ratings and time-varying exogenous variables such as descriptions of loudness, tempo, tonality, and timbre on a second-to-second basis (Schubert, 2004; Korhonen, Clausi, & Jernigan, 2006; Coutinho & Cangelosi, 2009; Dean & Bailes, 2010; Dean, Bailes, & Schubert, 2011). Although the products of Activity Analysis could be used to similar ends, their primary function is to describe and evaluate the responses themselves, looking within a set of listenings for signs of coordinated response events that may be aligned with the stimulus. A set of continuous responses to the same stimulus, each of the same response measure, sample rate, and duration, are here referred to as a collection.

The systematic study of music's impact on listeners hinges on recognizing what response patterns are common and repeatable, indications of effects shared across listeners and listenings. Continuous ratings are often highly variable and idiosyncratic (Dean, Bailes, & Dunsmuir, 2014), and a collection of responses to any given piece of music may not always paint a clear picture of how it can influence listeners. All accessible traces of response are vulnerable to complications, some of interest to music cognition like a listener's musical history, others less so, such as the

mechanics of a rating interface or of the specific reporting task. Given a collection of responses from a group of participants to a piece of music, do they agree enough to justify relating their time course to what the listeners heard? Activity Analysis provides a means for testing this against a well-defined null hypothesis of uncoordinated responses. These tests of coordination are designed to answer the question of whether responses are active in such a way that we could presume that the music is having some repeatable coordinating effect on them.

Continuous ratings to music have inspired a number of methodological papers, each proposing techniques that prioritize different information in the collections of times series. Nielsen (1987) used only the *average* tension-rating time series to describe or connect ratings with the presented music. Others have clarified or reduced average series with *down-sampling* (Chapin, Jantzen, Kelso, Steinberg & Large, 2010) or by taking the *first-order difference* (Schubert, 2004). Time series analysis (Bailes & Dean, 2012), from the tradition of econometrics, and functional data analysis (Levitin, Nuzzo, Vines, & Ramsay, 2007), out of mathematical biology, have demonstrated the insights such tools support. And yet, across the many approaches proposed, inter-response agreement or coherence has continued to be difficult to interpret across studies.

The degree of agreement or coherence across a collection of responses can be assessed in several ways. Making a claim about the reliability of an average or other summary statistic depends on having a good idea of what uncoordinated or unreliable responses would look like. After introducing Activity Analysis in the first section of this paper, we explore and compare other techniques for assessing coherence in continuous responses to music in the section Activity Analysis in Context. For this, we use a large set of response collections, containing over a thousand unidimensional continuous ratings from five studies, to compare Activity Analysis tests with statistics used in other papers to quantify coherence. We estimate false-positive thresholds for these statistics, including Cronbach's α, average inter-response correlations, and correlations

between averages of different populations' responses to the same works.

Another recurring question is how to mark moments of distinct agreement in continuous responses to music. The statistical burden of identifying significant time points is not trivial. Activity Analysis evaluates local activity by employing a nonparametric strategy to quantify the unexpectedness of simultaneous activity events. Also in Activity Analysis in Context, we show how this method contrasts with the second-order standard deviation (Schubert, 2007) and a clever use of Wilcoxon's test proposed by Grewe, Nagel, Kopiez, and Altenmüller (2007).

Coordination between response events such as changes in ratings is worth measuring because it can expose useful information about reliable elements of our musical experience. The section on Activity Analysis on the Experimental Data demonstrates that musical stimuli encourage different degrees of coherence in the continuous ratings of participants, and some of these fail to inspire statistically significant coherence in rating changes. Different audiences of participants can respond similarly to musical stimuli, both in degrees of coherence within groups and in the time course of subjective emotion rating changes. However, different interpretations of the same piece of music can yield noticeably distinct patterns of local activity in subjective emotional responses. Addressing broader methodological questions for continuous ratings, we evaluate the impact of two-dimensional rating interfaces for emotion and find no systematic interference between the dimensions. We also find that emotional arousal ratings are significantly more coherent than those of emotional valence.

Although broad in scope, this paper aims to equip music cognition researchers with statistical tools and a richer understanding of what they offer to the investigation of *when* we are affected by music as we listen. Whether through Activity Analysis or careful application of alternatives (including those discussed below), we encourage fellow researchers to dig deeper into ratings and other continuous responses collected during music listening.

Activity Analysis

Continuous measures of experience, from sensors tracking blood pressure to digital interfaces for reporting felt emotion, are collected by taking measurements at a sampling rate that is appropriate for the response signal of interest. However, regularity of sampling does not imply that the monitored phenomenon behaves smoothly; in many time series, some time points are more salient or informative than their neighbors. For Activity Analysis, we focus on these salient moments, selecting one kind of activity at a time, and investigate these specific events rather than try to explain every sample point of the original time series.

Consider continuous ratings to music and how they are generated. The task of constant self-assessment is onerous, and it seems implausible that participants succeed in simultaneously attending to the music, their response to the music, and the process of reporting these through physical gesture, all without break. It is more likely that they reach a compromise between these demands, for example, by considering what has happened and what the music is likely to do next, and then communicating the dynamics of whatever they decide to report through a gesture that does not require constant monitoring to execute. Figure 1A shows a representative response of one listener's rating of perceived emotional valence to the Allegro movement of Liszt's Piano Concerto No. 1. This rating was collected via mouse cursor on a computer screen in a 2D arousal-valence rating task, sampled at 1 Hz, while the participant listened over headphones (Korhonen, 2004, details in Appendix 1). This response alternates between the positive and negative valence ranges, with moments of sudden change between intervals of slow or no change. How do we interpret the rating values in intervals with many quick changes, for example from 20 s to 70 s in Figure 1A? Are the smaller dips in emotional valence at 145 s (0.15 to 0.105) and 195 s (0.1 to -0.02) idiosyncratic adjustments by this one listener, or are these subtle reactions also reported by others?

[insert Figure 1]

The scientific investigation of continuous responses to music begins by studying robust repeatable patterns. Experiments collecting continuous ratings to music have tended to gather dozens of responses to look for what might be predictable and consistent. And with these many samples comes the need for effective summaries of these data.

The most common summary for continuous rating collections is the average time series, essentially the average rating value at each sampled time point across all responses to the stimulus. But, when aggregating time series, there is a risk that the variance at each sample point may not be normally distributed, and this cross-sectional dispersion may change over time (heteroscedasticity). Some kinds of variability can be alleviated by treating the responses with particular transformations, e.g. filtering for a particular rate of change or dynamic time warping when there exist sufficient criteria for realignment. The average time series very specifically discards the variability across responses, and the resultant time series is always compressed relative to the range employed by individual responses, as can be seen in Figure 2A for valence ratings to the same Liszt excerpt. Although most of the responses cover more than half of the valence scale, the mean is restricted to less than 2/3 of the median rating range used (42%). This aggregate also results in a loss of some interesting information, such as instances of disagreement, for example between 150 s and 175 s in this excerpt. This moment immediately follows a structural boundary, featuring modal mixture as instruments exchange melodic lines. The responses are numerous both in the positive and negative valence range and some ratings move in opposite directions around 170s, leaving the average flat in the neutral middle. In considering only the average, we cannot differentiate moments of disagreement from moments with popular middle rating values. For some types of analysis, this is not a problem, but when reactions are sparse over time or we are interested in capturing behaviors exhibited by a minority of participants, the average can obscure important information.

[insert Figure 2]

Activity Events

One limitation of continuous ratings is the uncertain accuracy of the responses reported, in terms of the use of the rating scale and the criteria used for reporting changes in value. An advantage of continuous ratings, on the other hand, is the specific timing of changes in the reported response. Focusing on response dynamics, timing can contribute to identifying influences of the time-varying stimulus itself. We can consider causal factors for these instances of active reporting by virtue of the sequence of information presented to participants. Activity Analysis is an event-based approach to the study of continuous response to music: we choose what kind of *response event* is of interest (*activity*) and then evaluate whether this activity coincides across the measured responses in a collection. The process depends on defining the event we want to track and on translating each response into a point process, a time series of 0s and 1s with 1 indicating an event occurring within a given time frame.

For a given response measure and event definition, we can assess how much event activity occurs within a response time series. Some events are very frequent, for example rating changes in perceived tension (Fredrickson, 1995), whereas other events are more rare, such as large-scale peaks in aesthetic experience (Capperella-Sheldon, 1992). A simple estimate of event frequency is the *activity rate*, calculated by the number of detected events divided by the duration of the stimulus. Participants' recorded responses may differ in rate for a given event, but beside this variation, activity rates are also dependent on the stimulus, the continuous response collected, and the event type.

The lower three panels of Figure 1 show such *activity time series*, each a different kind of response event in the individual rating of perceived emotional valence: moments with increases in

valence (1B), moments of decreases (1C), and moments when the rating moves between the positive and negative halves of the valence scale (1D). Many kinds of response events can be captured in this fashion. Just as the researchers must decide what kind of task to give the participants, the decision concerning what kind of event to analyze depends on the questions at hand. For the following analyses, the active events in question are changes in ratings, either *increases* (*inc*), *decreases* (*dec*) or both, as they are particularly relevant to continuous ratings of emotion and tension. Unless otherwise specified, responses showing an increase or decrease in rating values of at least 2.5% of the rating scale over a 2-s time frame are considered to be an active event at that moment. This definition of active events is discussed in more detail in the section on the parameters of coordination tests.

Although the event-activity time series for an individual response invites speculative comparison to the stimulating music, we cannot make much of these events in isolation; many extra-musical factors may influence the ratings, preventing us from drawing conclusions about what is related to the stimulus for a single listener. If we wish to generalize from responses collected to the impact of this or other pieces of music, it is important to distinguish which effects are reliable, showing up in a significant proportion of responses in a collection.

Activity Levels and Activity-Level Time Series

Given a collection of responses to the same stimulus, we can apply the same event assessment criteria to each response and then count how many responses show the same kind of event at approximately the same time. Response events are not expressed instantaneously, and responses to the same stimulus event may be reported by different participants at different delays, so we define a *window of synchrony*: an interval of time over which response events are counted as occurring together. The *activity level* is then the proportion of responses showing the same event within the window of synchrony. By definition, activity levels range from 0 to 1, by which 0 means

no responses contain the event in question within the time frame considered, and .5 means that half of the responses in the collection are active within the window of synchrony.

The *activity-level time series* comes from assessing activity levels in a sequence of time frames over the collection's time series. Depending on the goal of the analysis, frames may need to be nonoverlapping. Figure 2 shows multiple activity-level time series for the perceived valence ratings of the Liszt excerpt from the Korhonen (2004) experiment. In this paper, plots reporting activity-level time series with bar graphs depict nonoverlapping time frames, as seen in Figure 2D. The multiple smoothed histograms above it (2B and 2C) report activity levels for different minimum thresholds of rating change activity in overlapping frames with a hop-size of one sample (1 s). With these series, we can compare the rating changes in the single response of Figure 1 to those across the collection. The rapid changes from 20 s to 70 s are shared by many but not all participants, visible in the dark spikes in 2B (increases) and 2C (decreases). The shallow fall and rise shown in the single response of Figure 1 at 145 s is not common enough to be clearly distinguishable, however the slightly larger fall and rise from 195 s can be read in the lighter shades of 2C and 2B, reporting changes greater than .025 or .005 of the rating scale.

A popular rule of thumb for the spread of rating responses to a stimulus event is the assumption that most participants' reactions will be expressed within a 2-s interval (Schubert, 2010; 1-3 s after the stimulus event). There has not been direct investigation into the dispersion of latency in rating changes in reaction to complex ongoing musical stimuli, but precedence and tests of parameters depicted in Appendix 3 of this paper suggest that 2 s is not an unreasonable window of synchrony.

The activity-level time series is a summary of the responses in a collection, focused on the timing of a particular response behavior. This representation is complementary to the average time series; it yields distinct insights into what is typical of the individual responses in the collection. By

plotting the activity levels of decreases in ratings below zero, as in Figures 2D and 3B, we can see a pattern of alternation between moments in which there are coincidental rating changes first in one direction and then in the other. The same fall and rise around 200 s from the response in Figure 1 was shown to be popular across the collection in Figures 2B and 2C, and this can also be read as a wave in Figure 2D: the activity-levels for the decreases reaching a maximum of 0.4 before 200s and the increases peaking at 0.31 only 4 s later. The shift of the average valence rating at this moment, in Figure 2A, is small considering the proportion of participants actively reporting similar changes.

Although the participants do not consistently report the same amount of change, increases and decreases can be popular and coherent at some moments. At other times, rating responses show changes in opposite directions, and these become visible in the activity-level time series plots by the superposition of increases and decreases at a given moment. Differences in rating-change activity levels yield greater contrast between successive moments than average rating scale values, encouraging distinct conclusions. The activity-level time series each show three distinct peaks in the first 50 s, whereas the average reports only one. While the average valence rating creeps upward from 235 s to 245 s in Figure 2A, 2D shows that a substantial subset, at least 9 of the 35 participants, were also reporting decreases in perceived emotional valence. From here we do not know if they were reacting to harmonic modal mixture, a decrease in loudness, or other factors, but at 240 s these participants reported an experience contrary to the dominant narrative.

Figure 3 shows activity levels of increases and decreases in a collection of felt emotional intensity ratings to Mozart's Overture to *The Marriage of Figaro*, K492. While alternation between the two forms of rating-change activity is visible to our pattern-hungry eyes, the proportion of responses showing concurrent increases is hardly ever as much as half of those in the collection. Decreases are even more thinly spread, with at most a third of the participants reporting decreases in felt emotional intensity in any given 2-s time frame. Although activity-level time series for rating

changes are not always so sparse, this representation calls into question the robustness of the music's influence on these responses.

[insert Figure 3]

The moment with the highest concurrent activity for the collection in Figure 3B is in the first 20 s of the music. As discussed by Schubert (2013), rating responses often have their largest changes shortly after the beginning of the stimulus, and such changes can have disproportionate weight on subsequent analyses as they integrate the mechanics of orienting to a new stimulus with the task of reporting. We are reluctant to discard these early instances of response activity in this analysis. The relative magnitude of change has no effect on the event types used here, and the timing of changes may still show stimulus-related synchrony. For example, the initial peak in activity shown in Figure 3B aligns with the first fortissimo/tutti moment in the music.

Testing Coordination in Activity

A test is needed to evaluate whether these response changes are likely to be an accumulation of noise, rather than driven by the music. For this, the coordination tests of Activity Analysis use the distribution of activity levels in the activity-level time series, a model of random activity, and a probability estimate. By themselves, tests of coordination are not tests of whether a musical stimulus is effective, nor can they measure whether participants were performing the task. However, if, over time, a stimulus changes along a dimension relevant to the rating task and the participants communicate their experiences with sufficient accuracy, the rating-change activity should show significant coordination. This coordination would be expressed in the distribution of activity levels across the activity-level time series–many moments of relatively high activity levels as well as a great number of remarkably low activity levels. In contrast, if responses seem to be changing independently of each other (and the stimulus), the distribution of activity levels should

be less varied. With a plausible model of this "less varied" distribution of activity levels for random uncoordinated activity, we can test whether our experimental data are all that different.

For rating changes, we use a simple parametric model of random activity, as if the responses changed independently of each other. The average activity rate across the responses in the collection is fed into a binomial model to generate a distribution fit to the size and character of the experimental data (Appendix 2). Demonstrating this on the Emotional intensity ratings to the *Figaro Overture*, Figure 3 (C and D) shows the actual activity-level distributions for rating increases and decreases and the random activity model for each type of activity. The random activity, in black, is slightly more concentrated around their averages, and the experiment collection reports twice as many 2-s time frames with zero rating increases than the random activity model would suggest (C).

To evaluate the significance of these differences between this collection's rating changes and the model's, we apply Pearson's Chi-squared Goodness-of-Fit Test, a standard statistic for comparing experimental data to random distributions. The goodness-of-fit test evaluates the likelihood that sampling from the random model would yield distributions similar to, or more extreme than, that of the experimental data, defining a p-value to be compared with some acceptable Type I error rate, for example, $\alpha = .05$.

There is an important condition for applying Pearson's goodness-of-fit test: each category or bin of the expected distribution (the random model) must have at least five samples. However, the activity levels measured in a continuous response collection like that of Figure 3 (B, C, and D), do not cover all possible values, nor are they expected to in our random models. In fact, the highest activity levels, such as 100% of the ratings changing in the same 2 s window, never occur in most experimental data collections, and so the categories of possible activity levels cannot be used directly in this goodness-of-fit test. To get around this, we make larger bins of activity levels,

counting together time points of similar degrees of activity levels (low, middling, high), using a simple algorithm to divide the random model into a reasonable number of bins of near equal size (see Appendix 2). These criteria ensure consistency in the application of this test and limit the distorting effect of outlier data points by giving each time frame more equal impact on the final value. Figure 3 (E and F) presents the bins of a goodness-of-fit test comparing the activity-level distributions for increases and decreases against the random alternatives along with the resulting chi-squared values. The activity levels for each direction of rating changes are found to be significantly different from the random activity model.

Collections of responses that show a lot of concurrent activity will have greater differences with the random model, whereas the contrast will be smaller for those with a higher proportion of noise. We can treat extremeness of these differences as a measure of coordination. Like any statistic, the value would be only an estimate of how strongly the responses change together, but even so, it might be useful for distinguishing the coordinating effects of stimuli or the sensitivities of different audiences, among other possibilities.

From the activity-level distributions, we propose the Coordination Score. This number is calculated using the *p*-values from the parametric goodness-of-fit test via a simple formula similar to that used in Yeshurun, Carrasco, & Maloney (2008). The explicit construction of the score is outlined in Appendix 2. The implementation of this calculation in the Activity Analysis Toolbox (Upham, 2017) yields values from 0 to 16, with scores above 2 effectively equivalent to $p < .01$. An important advantage of basing this measure on the goodness-of-fit test is that the results are comparable across collections of various durations, and numbers of participants. (See Appendix 4 for more details on how parameters of response collections affect Coordination Scores.)

Coordination Between Collections

The question of coordination need not be restricted to response events in a single collection. If the stimulus is really driving these events in listeners' responses, we expect another group of similar listeners would show comparable activity levels at the same moments in the music. Using contingency tables, we can test the independence of activity levels per moment of music between two collections of responses. If the null hypothesis of independence is rejected, we have reason to interpret the shared stimulus as influencing the two collections in ways that are related.

The Boston Symphony Orchestra Project invites this comparison, with two audiences experiencing related stimuli: one group reporting felt emotional intensity while attending a live performance by the BSO and the other performing the same task while hearing and seeing a video recording of this performance projected in a recital hall (details in Appendix 1). Figure 4A shows activity-level time series for increases from both collections of responses to this performance of the Overture of *The Marriage of Figaro*, K492. The partial symmetry of these two series suggests agreement between these audiences. While the music does not prompt reports of change in all or even most participants, a similar proportion are active in each group at many moments of relatively "high" activity levels, such as at 10 s, 24 s, and 122 s, each directly following the dramatic mid-phrase tutti of the Overture's first theme.

[insert Figure 4]

Were the activity levels of these two collections independent, we would expect one group to show mostly middling or low activity levels when the other is high, and vice versa. Figure 4C presents a heat map of how many time frames occurred at all possible combinations of activity levels from these two audiences as they listened to the Overture. In contrast, Fig. 4D reports the expected distribution of joint activity levels, if the increases in ratings in one collection were

independent of the same activity events in the other. To test this difference, we divide these joint distributions according to those of either collection (Fig. 4B and 4E), each cut into three bins containing approximately the same number of time frames. The totals are reported in the contingency table (Fig. 4F), which has relatively little variation in darkness/values. The actual joint activity-level distribution is reported in Fig. 4G, and their differences are tested with a chi-squared test, once again using a parametric estimate of the difference against a null hypothesis of independent collections. The test tells us that there is a significant difference between the round shape in the corner of Figure 4D (expected independent joint-activity) and the form of the center plot (3C) that stretches a bit more along the diagonal. This latter shape tells us that the relationship between the two activity-level time series is roughly parallel.

Like the first test of coordination in a single type of activity on a single collection, the calculations of this test can be transformed into a Coordination Score, or rather a Bi-Coordination Score (Bi-C score). Together, these two collections of felt emotional intensity ratings to this performance of the *Figaro Overture*, in Figure 4, shared a Bi-C score of 5.5 for rating increases. This result suggests that there were shared influences on the response activity of the two groups, moment to moment, despite different response conditions and participants.

Nonparametric Coordination Test

Not all types of activity or aspects of response events can be assessed with parametric statistics like the chi-squared test with a binomial model. When we are not sure if a parametric model is close enough to the behavior in question, it is safer to use numerical approximations, taking the nonparametric approach, despite the extra computational cost.

For Activity Analysis, the effectiveness of the parametric tests used for Coordination Scores depends on the measured activity event. Rating increases and decreases in non-overlapping 2-s

time frames seem close enough to a random binary process that we can use it to model the null hypothesis of stimulus-independent activity. But if we want to consider the timing of inhalations across an audience, for example, we know that any given listener must breathe every three or four seconds, with mild adjustments in respiration period from breath to breath. The occurrence of alignments between inhalations across an audience can only be assessed against a null hypothesis of accidental coincidences that respect these signals' temporal characteristics. Rather than attempting to construct a parametric model of this type of behavior, we can use the responses themselves to define uncoordinated activity through permutations of the data.

We are primarily interested in the coincidence of events across responses in relation to the timeline of the common stimulus. If the activity is coordinated by the music, breaking the temporal alignment between responses should result in less extreme activity levels. If we shift each individual response by a random amount of time (say some interval sampled between 0 to 30 s, hereafter referred to as the *shuffling range*), this alternative alignment of our real responses would result in a physically plausible activity-level time series but without the potentially coordinating effects of the stimulus appearing in the activity-level distribution. Bootstrapping the alignment of the original data, we generate 2000 uncoordinated alternative activity-level distributions and then test how the stimulus-aligned collection's activity-level distribution ranks in distance from their average. Appendix 2 explains the calculations of the nonparametric test in more detail.

Nonparametric Coordination Test of Local Activity

Besides providing alternative activity-level distributions for the nonparametric coordination test, these shuffled alternative alignments also generate a distribution of alternative activity-levels for each time frame of the series. We can assess the expectedness of the collection's stimulus-aligned activity levels against a distribution of uncoordinated activity levels tailored to each second of the music (Grün, 2009). Frame-by-frame, their rank against the non-aligned

alternative activity-levels identify those moments of extreme high activity levels, say above 98.5% of the alternate distribution, or extreme low activity-levels, below 2.5%. Figure 5 reports moments of extreme high and low activity levels for rating increases in a collection of emotional arousal ratings to "Morning" from Grieg's *Peer Gynt Suite*. With such strong alignment between responses (Fig. 5A), a great many moments are marked as showing extreme local activity (Fig. 5B). These moments of salient activity levels are not defined in relation to a fixed threshold. Instead, the activity patterns of all responses from the surrounding minute determine the expectedness of each time frame's measured activity. Therefore, the increase of activity before 20 s is selected as a moment of notable alignment when later points with higher activity levels are not. Over the course of a 3-min stimulus, we expect some time frames to reach locally extreme activity levels by accident, but there is still great exploratory opportunity for well-defined criteria to investigate active moments in relation to the music. When no other criteria are available, we recommend studying the results of the local-activity coordination test only if the collection of responses shows significant coordination as a whole according to the associated nonparametric activity test. Appendix 2 explains further the details of assessing local-activity coordination.

[insert Figure 5]

Tuning Parameters of Tests for Rating Increases and Decreases

The application of Activity Analysis requires a few parameters to be fixed, parameters that depend on what is being measured. First, what qualifies as a response event? For rating changes, this amounts to the size of change that is counted as activity within some time interval, namely exceeding the minimum rating change threshold. Second, within what time interval might events be counted as happening at the same time, i.e., how big is the window of synchrony? The nonparametric test and local-activity test also require a third parameter: the duration of the shuffling range that best distinguishes coordination from coincidence. These parameters should be

set so that the tests perform as expected: detecting uncoordinated collections with a reasonable false-positive rate.

One way of evaluating reasonable parameter values is with previously collected experimental data. Across the five different experiments on listener responses made available for this study, we have 42 collections of one-dimensional continuous ratings to western concert music, from perceived emotional valence to thematic familiarity. With these 1350 individual one-dimensional ratings, we can generate unrelated-response collections: combinations of ratings to different stimuli that together should only look coordinated by coincidence. The details of their composition are shared in Appendix 1. Generating examples of uncoordinated collections with experimental data ensures that our random collections still hold some important qualities connected to the style of stimuli and the task of rating, including the natural variability in participants' rating strategies.

The activity levels of rating increases and decreases in 2000 unrelated-response collections were assessed for different values of each parameter. After exploring many combinations of the window of synchrony and the minimum change threshold for the Coordination Scores, the 2-s window of synchrony and a minimum change of .025 of the rating scale resulted in reasonable false-positive rates. For the parametric within-collection coordination tests on rating increases and decreases, ~1% of these unrelated-response collections reached or exceeded the Coordination Score of 2, our target given the construction of these scores. The nonparametric evaluation of these activities found 2.5-4.5 % of these unrelated-response collections ranked at or above the 99th percentile of the alternatives for these same parameter values.

To evaluate the impact of these parameters on between-collection measures of coordination, we construct unrelated pairs using the original response collections, excluding combinations that are related by stimulus. From 40 collections of adequate size (more than 14

ratings), we have 753 pairs that should not be measured as coordinated except by coincidence. The performance of the coordination test depends on its capacity to differentiate between similarity from the common shapes of ratings and the specific stimulus influences on the timing of rating changes. The implications of the window of synchrony and minimum change threshold are very different for this test but again, a window size of 2 s and a threshold of .025 yielded a false-positive rate of ~1% (.93% for Increases, 1.3% for Decreases).

In the parameter spaces evaluated, larger time frames yield similarly acceptable false-positive rates on the unrelated-response collections. However, larger windows of synchrony result in fewer time frames over which to assess the Coordination Scores. This increases the minimum response length onto which these can be applied without violating the limitations of the goodness-of-fit test. With a 2-s window of synchrony, the parametric coordination tests can run on collections of continuous ratings of 120s or more.

A last parameter to consider is the shuffling range for the nonparametric assessment of activity coordination and local activity. This was evaluated by comparing the proportion of remarkably high and low activity-level moments identified in experimental collections and uncoordinated response collections. The difference between these collections grows with the shuffling range and stabilizes with substantial advantage for the experimental response collections (2.5 times as many high activity moments, 5 times as many low activity moments) from 30 s onwards.

Unless otherwise specified, the parameters for all rating changes activity assessments in this paper use these values: minimum rating-change thresholds of .025 of the rating scale, 2-s window of synchrony, and 30-s shuffling ranges. More details on these evaluations can be found in Appendix 3.

Conclusion

Activity Analysis focuses on a specific kind of agreement between continuous responses, matching not in overall tone or mood but rather through the simultaneity of response events. It loosens the expectation of agreement between responses over time without giving up the power to identify repeatable stimulus-related patterns. A response can actively agree with one subset of the collection at some moment and be unmoved when most of the same group change 20 s later, but that inconsistency is not a problem. Rather, activity-levels are agnostic to which responses are active, while the coordination tests focus on finding exceptional inter-response agreement as measured in moments.

Certainly, coordination in activity is not the only clue to a stimulus influencing responses. The existence of rating changes and the shifts in rate of change would, in most cases, suffice as evidence of response to music if the alternative is silence. However, coordination is a strong argument for the repeatability of effects on listeners' experience of music and timing cues can distinguish responses to different pieces. How, then, does this approach to issues of coherence in responses compare to other statistical tools used for studying continuous ratings to music?

Activity Analysis in Context

The question of whether continuous ratings agree is not new. Several statistical tests have been employed to assess how well these traces of experience confirm or contradict each other. Studies of responses to music have used well-established calculations such as Pearson correlations and Cronbach's α along with statistics modified for particular purposes like capturing moments of affect. Each measure of coherence treats some information as important and other information as noise to be discarded, prioritizing different aspects of these responses. We will be using the term "coherence" as an umbrella over these different flavors of inter-response agreement.

In this section, we assess three types of coherence measures and compare them to those of Activity Analysis: coherence between ratings within a collection, coherence between collections of ratings to the same musical stimulus, and the local coherence within the time course of a collection. Some of the first and second types of coherence have been used as significance tests, but evaluating the significance of these statistics on continuous response collections can be problematic. Instead of parametric estimates of significance employed elsewhere, we use the unrelated-response collections and pairs of collections drawn from the effort to tune Activity Analysis parameters to identify plausible thresholds for α_{crit} = .05 and .01 (significance) for each of these other statistics. As in the previous section, these serve as examples of collections that should not qualify as coherent, except by accident. From these analyses and example applications on experimental data, we argue that the coordination tests of Activity Analysis capture important and distinct qualities of collection coherence.

Coherence Within Collections of Continuous Ratings

Thirty-five adults listened to the first three minutes of the Adagio movement of J Rodrigo's *Concierto de Aranjuez* and continuously rated emotion in a square interface with dimensions of Arousal x Valence (Korhonen, 2004). Figure 6 reports their ratings along each dimension along with the average rating responses (6A and 6C) and the corresponding rating-change activity-level time series (6B and 6D). Do these continuous ratings suggest some coherence in their evaluations of this music? Were changes sufficiently coordinated that we might expect another group of listeners to show similarly timed effects? What values of inter-response correlations or Coordination Scores support the claim of significant coherence within such a collection?

[insert Figure 6]

Treating each dimension as its own collection, ratings along the Arousal dimension show

cohesion according to most of the measures. According to the tests of Activity Analysis, the increases in ratings are highly coordinated (14.0), an obvious claim according to activity levels in the second plot, with many increases between 35 s and 60 s and near complete quiet after 120 s. The decreases in ratings are not as coordinated, C score = 3.0. The distribution of activity levels for decreases is not that of a random distribution, but the number of frames with low-to-middling activity levels demonstrates that these participants are not consistently reacting together with decreases in their Emotional Arousal ratings at specific moments of the music. The ratings of emotional valence to this piece show much lower values: increases (1.9) and decreases (.9). In Figure 6C, the responses are split on whether the music is positively or negatively valenced, and many moments show some ratings increasing while others report the opposite. The three other types of within-collection cohesion measures discussed here, Cronbach's α, inter-response correlations, and ratios of deviation, may not draw the same conclusions as the coordination scores of rating changes for these two rating dimensions.

Cronbach's α. This measure of agreement ranges from -1 to 1 and was developed to assess how effectively a test (the aggregate of a set of test items) captures the variation of a population along a single dimension (Cronbach, 1951). It is most often used to evaluate the length and quality of tests for assessing psychological traits, comparing the agreement of each test item, say a rating of agreement to a statement, and the final score. In this context, values of .8 are considered to be good. Translated to a continuous rating experiment, Cronbach's α evaluates the relationship between the variance in rating values across time samples for each participant and the variance in the average time series across time samples. The test population is constituted by the moments of the music, second by second, and in that sense, Cronbach's α can be used to assess the effectiveness of the average response of these participants' ratings at capturing variation along the feature of interest, say emotionality, within this musical piece and others with a similar range over time.

One study has used this statistic to compare continuous ratings on different scales and between different pieces for participants attending a live performance (Torres-Eliard, Labbé & Grandjean, 2012). Fourteen participants' continuous ratings of "Power" to the live performance of the 2^{nd} movement of the String Quartet n°3 in A major op. 41 by R. Schumann (395 s) produced a Cronbach's α = .94, whereas their continuous ratings of "Sadness" during the 3^{rd} movement of the String Quartet n°4 in C major by B. Bartok resulted in a Cronbach's α = .71. Do these values suggest that the participants were reporting similar judgments?

To estimate which values of Cronbach's α imply coherence or the lack thereof in collections of continuous ratings, we applied this statistic to the 2000 unrelated-response collections: the resultant distribution gives us a sense of the numbers likely to occur when there isn't a common stimulus guiding or driving listeners reactions. Table 1 reports the most important part of the distribution: the 95th and 99th percentiles in these distributions, which estimate the 5% and 1% false-positive rates. Over the range of collection parameters of the unrelated-response collections, Cronbach's α values over .82 and over .85 effectively exceed $\boldsymbol{\alpha}_{crit}$ = .05 and .01 significance thresholds, respectively. From this information alone, it seems the Power ratings in the Torres-Eliard et al. (2012) study were significantly coherent, whereas the Sadness ratings were not. However, the number of responses in a collection and the duration of ratings both affect Cronbach's α values. According to explorations described in Appendix 4, the value of .71 for a collection of 14 responses may instead fall between the 90th and 95th percentiles of values on incoherent collections.

Looking back to the example of the Arousal and Valence ratings for the Rodrigo excerpt (Fig. 6), their respective Cronbach's α values were .97 and .48. Cronbach's α for the valence ratings falls below the median across all the unrelated-response collections (.59), whereas the Coordination Score for rating increases was on the edge of the 99th percentile with 1.9. Considering the

relationship between Cronbach's α and the average response time series, this seems appropriate: the participants' ratings seemed to disagree, and the average score per moment has little relationship to any of them. And yet a certain degree of response coordination seems to have been hidden in the dataset, entirely overlooked by this measure of coherence.

Average inter-response correlations. Many analyses of continuous ratings have employed combinations of correlations to assess the agreement between participants' responses. One common version of this inter-response correlation, here referred to as *InterCorr,* reports the average pairwise correlation coefficient between all responses (e.g., Krumhansl, 1996; Toiviainen & Krumhansl, 2003; Williams, Frederickson & Atkinson, 2011). Another possible measure of within-collection correlation is the average correlation between each response and their average response time series, here referred to as *MeanCorr*. If a collection of responses shows sufficient agreement in changes over time to produce an average time series that shares their large-scale contour, then the MeanCorr will be much closer to 1 than a collection with little agreement. Like Cronbach's α, this statistic presumes that individual responses in a collection are the sum of a singular underlying response and noise.

The significance of either average inter-response correlation measure is not easily calculated for rating time series. Aside from concerns about inflation from serial correlation in time series (Schubert, 2004), many published *p*-values for this statistic are necessarily false because the data do not comply with the standard significance estimation conditions for correlations. Significance testing depends on knowing how much independent information is present in each set, something not easily assessed in time series (Pyper & Peterman 1998). Several techniques make comparisons between time series more reasonable, but these methods often change what is being compared. For example, a correlation between differenced time series does not assess the same kind of inter-response consistency as a correlation between filtered and downsampled versions of

the same (Upham, 2012). However, correlation coefficients can be very informative outside of the context of the common Pearson significance test (Rodgers & Nicewander, 1988). If responses go through periods of high and low values, they can end up comparing the large-scale contours of rating time series, gaining power if the more extreme values co-occur, while the middle distribution of sample values have relatively little impact. With appropriate reference values, these calculations should tell us something important about the similarity between participants' rating responses. However, their performance on real continuous ratings to music is not very consistent.

In rows three and four of Table 1 are the 95th and 99th percentile values of these statistics over the unrelated-response collections, our estimate of threshold α_{crit} values for .05 and .01. Additionally, both statistics are also very sensitive to the duration of responses. Across all 2000 collections of diverse parameters, the 95th percentile for Intercorr was r = .35, but threshold may change by as much as 0.1 for a difference in duration of 240 s, whereas MeanCorr may shift to a similar extent. MeanCorr is also sensitive to the number of responses in a collection, with the 95th percentile shifting from .52 to .38 for collections of 12 to 36 ratings. Like other applications of correlations, degrees of freedom in both response duration and the number of responses per collection change the statistical implications of these correlation values. See Appendix 4 for analysis details.

With these numbers in mind, consider the InterCorr values reported in Krumhansl's (1996) landmark study on continuous ratings of tension. From her first experiment, the 15 participants' ratings of tension over 314 s of music had an average pairwise Pearson inter-correlation of r = .42. This is close to the 99th percentile value of the InterCorr statistic on the 2000 unrelated-response collections, suggesting a p value near .01. Add to this the relatively long duration of these responses and the correlation argument for stimulus-related cohesion appears even stronger. In the fourth experiment reported in her paper, InterCorr values on tension ratings by 24 participants to 224 s of

music were around r = .18, indistinguishable from the unrelated-response collections. And yet, the averages of these responses correlated well with those to similar stimuli. Given the combination of these results, the Intercorr statistic seems particularly ineffective as a measure of cohesion between a collection's continuous ratings.

The utility of any coherence measure hinges on whether it can distinguish collections of responses with shared stimulus effects from those without. Thus far, we have discussed the false-negative rate, setting threshold values per statistic to limit the likelihood of mistaking incoherent response collections for coherent ones, but also of great importance is the false-negative rate: the likelihood of dismissing a collection of stimulus-coordinated responses as incoherent. We cannot assume that all of our 40 experiment collections are actually coherent, but if their coherence measure statistics distinguish them from the unrelated-response collections, that is a good indication of stimulus-inspired coherence. The third and fifth columns of Table 1 report these results for all within-collection coherence measures discussed here. Of these, Intercorr and Meancorr detect the fewest as more coherent than the unrelated-response collections. For InterCorr, only 14 experiment collections of 40 generate values greater than the 95th percentile, 12 for 99th percentile. For MeanCorr, the experimental collections performed better, with 26 of 40 exceeding the .05 equivalent, and 17 exceeding .01. If a correlation is the preferred measure of similarity between responses, then by these results MeanCorr seems to be more useful than the more widely published InterCorr calculation.

[insert Table 1]

Returning to the ratings of emotion to the Rodrigo excerpt in Figure 6, the respective MeanCorr r values for the valence and arousal ratings (Fig. 6) were .72 and .33, significantly coherent and incoherent, respectively, like the results of Cronbach's α. The InterCorr coherence measures, on the other hand, were .54 and .41, counting both Arousal and Valence ratings among

the 12 exceeding the 99th percentile of values from the unrelated-response collections.

Variance ratio. The Variance ratio, also referred to here as *VarRatio*, is a measure of coherence that uses the scale on which ratings are initially collected. The statistic is calculated by dividing the variance of the average time series (over time) by the mean of the variances of the individual responses (derived from the standard deviation ratio, Upham, 2012). It is an easily computed criterion for assessing the degree to which disagreement between responses flattens the average response time series. One advantage of this statistic over the Coordination Scores is that it is suitable for evaluating responses to stimuli that are not very dynamic alongside those with more tumultuous time profiles. However, it is still sensitive to the number of responses in a collection and the duration of responses, decreasing inversely to each.

The continuous ratings of emotion in response to the Rodrigo excerpt have a VarRatio for the arousal dimension of .53 and a value so low in the valence dimension, .05, that it falls below 7th percentile of the unrelated-response collections.

Parametric and nonparametric C Scores. The Activity Analysis tests have already been described in detail, and the 95th and 99th percentile values from parametric scores on the unrelated-response collections are as expected according to the tuning of the window of synchrony and the minimum rating-change threshold. The 95th and 99th percentile values for rating increases in the nonparametric coordination test are higher that its parametric cousin and should be interpreted accordingly.

Unlike the previous statistics discussed here, the coordination score thresholds set stable false-positive rates because the degrees of freedom from the duration of responses and the number of responses are both included in the calculations used to construct them. Above a minimum duration (120 s), these statistics are comparable between response collections of long and short

stimuli, and similarly for large and small numbers of responses. The nonparametric coordination score, to the contrary, appears to increase in value with larger collections and longer durations (see Appendix 4.)

Most remarkable is the sensitivity of these Coordination Scores to coherence in collections. The three rating-change Activity Analysis tests reported in Table 1 distinguish more experiment collections from the unrelated-response collections than even the strongest alternative: the Coordination Scores on increases in ratings identified 28 collections of 40 as significantly coherent for $\boldsymbol{\alpha}_{crit} = .01$, compared to 20 of 40 marked by Cronbach's α.

Correlations Between Coherence Measures

According to the descriptions above, the calculation of each coherence measure seems to prioritize only some aspects of what we might consider to be agreement between continuous ratings to music. Table 2 reports the consequences of these differences in how they evaluate our 40 experiment collections of responses. Spearman's ρ rank correlation was computed between three specific measures of within-collection coherence and all others discussed above, as was the number of experiment collections both measures would report as significantly coherent, by the 99th percentile thresholds reported in Table 1. Featured are Cronbach's α, the average mean-to-response correlation (MeanCorr), and the parametric Coordination score for activity in rating increases.

[insert Table 2]

The correlations between these measures over the 40 experiment collections of unidimensional ratings to music are positive and significant, but that does not make them interchangeable. Although Cronbach's α correlates with VarRatio at $\rho = .95$, VarRatio would miss a quarter of the collections identified by the first statistic as coherent. The Activity Analysis

coordination measures have lower correlations than all but InterCorr because they focus on distinct qualities of continuous responses. It is possible for a set of ratings to be very coordinated in their increases, but less so in their decreases, while an average-focused measurement of coherence depends on the alignment of both. Despite lower correlations with the other measures, C Scores of rating increases reveal more collections and report higher agreement percentages than the other measures discussed.

There may be instances when a specific measure of coherence is relevant based on the qualities of interest, and all of these measures may have their uses, save perhaps InterCorr. If, however, the purpose of the statistic is simply to identify whether or not the stimulus has had some common influence on the ratings gathered, the coordination scores of Activity Analysis may be the clearest and most broadly sensitive tests available.

Coherence Measures Between Collections of Responses

Another important question for continuous responses to music is whether there is agreement between two collections of responses to the same piece, say between the ratings from two groups of participants, or between the first and second listenings to a piece by one audience. Correlations of various types have been used to make comparisons between average times series of collections. Activity Analysis also proposes a test for this purpose. Figure 7 presents responses to another stimulus from the Boston Symphony Orchestra project with a live audience (Fig. 7A) and an audience watching a video recording of the same performance (Fig. 7D). Looking at the average time series (Fig. 7B), it is hard to know whether these two sets of responses really agree after the first 20 s of music.

[insert Figure 7]

Before drawing conclusions about the relationship between these ratings of emotional

intensity to a specific performance of the Jupiter Symphony's Finale (K551), we need a reference distribution for each between-collection coherence measure. As with the within-collection measures, it is essential to know what numbers are expected when there is no possible causal influence on coherence.

The 753 pairs of experiment collections unrelated by stimulus, initially described above (*Tuning Parameters of tests for rating increases and decreases*), can produce such a distribution and estimate thresholds for suitable false-positive rates. Table 3 reports the 95th and 99th percentile values on these unrelated pairings for these measures of coherence, along with the number of stimulus-related pairings found to exceed these thresholds. Like the within-collection coherence measures, we can get a snapshot of the sensitivity of these measures by comparing their assessments of experiment collection pairs with some possibility of similarity. In our set of 40 experiment collections, only four pairings are of responses to the same stimulus on the same rating scale. The 23 other stimulus-related pairings report different aspects of responses, say Emotional Arousal and Emotional Valence and are not expected to be strongly coherent. However, the opportunity for coincidental activity should be higher in these pairs than those unrelated by stimulus.

[insert Table 3]

Correlations. Several variations on correlations have been used to compare collections of continuous ratings, via their average rating time series. In this section, we consider three variants of correlation measures of cohesion between collections: Pearson or Spearman correlations, correlating the average rating time series directly or taking the first order difference (1-s step), and excluding the first dozen seconds of ratings to reduce the impact of the orientation time for rating tasks.

Schubert (2013) describes the time interval during which the accuracy of a participant's rating of a musical stimulus are confounded by the time needed to orient to the music and the mechanics of the rating interface, called the initial orientation time. Another study described this period on a similar interface as the integration time, reporting an average of 8.3 s before participants' ratings settled in a region of the rating range (Bachorik et al., 2009). Some of the largest rating changes can occur in this interval, and this is expected to compromise average rating time series. To remove the impact of this period, we cut the first 12 s, the median orientation time reported by Schubert (2013), from the average responses. On the distribution of correlations on the unrelated pairs, this shifted the means from r = .14 and ρ = .05 to r = .01 and ρ = .005. However, the 95th percentile values on these unrelated collection pairs is still very high: r = .68 with the whole time series, r = .57 without these first 12 s, ρ = .55 with, and ρ = .54 without. The 99th percentile values are also reported in Table 3. Either way, the difference does not compensate for the overestimation of degrees of freedom in many continuous rating studies that have estimated this threshold for $\boldsymbol{\alpha}_{\text{crit}}$ = .05 as less than .17 for continuous ratings 90 s in duration and longer.

A common approach to compensating for the serial effects of continuous ratings is to take the first-order difference of the ratings. Correlations between the *first-order difference* series of two collections average around zero for both Spearman and Pearson correlations, and the 95th percentile values over the unrelated pairs are r = .25, and ρ = .17. The distribution of values in these first-order difference time series is rarely normal. Differenced ratings and differenced averaged ratings tend to be composed of many small values and a few large negative and positive values within the units of the rating scale, resulting in very high kurtosis: a median of 20.5 for the first-order difference of our continuous ratings to music and 7.5 for their average time series per collection, whereas the plain average ratings have a median kurtosis value of 3.3. As such, it may be more practical to use the nonparametric Spearman's ρ if the goal is to capture coherence on differenced ratings. Without the added transformation of first-order differencing, the Spearman

correlation does not seem to be preferable to Pearson for analytic reasons, and it fails to distinguish those few stimulus- and task-related collection pairings above the 99th percentile of unrelated collection pair values.

As reported in Table 3, these correlation measures identified few stimulus-related pairs as more coherent than the unrelated pairs. The two collections described in Figure 7 seem more or less coherent depending on which type of correlation is applied. The complete averages correlate at r = .85, above the 99th percentile of unrelated pairs, less the first 12 seconds, r = .60, just above the 95th percentile, whereas difference series correlate at r = .25, on the line for the 95th percentile. Using Spearman's ρ, the means correlate at ρ = .62, or ρ = .59 when excluding an initial orientation period, and ρ = .17 on the differenced means. Over all, these numbers suggest these two collections are more coherent than might be expected by chance, but the effect is minor.

Bi-Coordination Scores. The Activity Analysis test between collections (Bi-C Score in Table 3) compares only one type of activity event at a time, so we have used it on three activity events: increases, decreases, and all changes of at least 2.5% of the range of ratings in 2-s windows of synchrony. These all show much higher sensitivity than the correlation based measures, finding a number of significant time-aligned events between collections related by stimulus but not rating scale. The broadest of the three tests, counting all kinds of rating changes regardless of direction, picks up on coherence in half of all possible pairs. This demonstrates the openness of these Activity Analysis-based coherence measures, picking up on patterns in time without demanding that participants report the same experience from beginning to end.

On the felt emotional intensity ratings in Figure 7 and the activity levels reported in 7C, the increases in ratings have a weak Bi-Coordination Score of 2.1, whereas decreases and rating changes score only .9 and 1.1, respectively. By these measures of between-collection coherence, the two sets of responses show very little coordinated activity–barely a trace of shared stimulus effects.

The details and implications of this are discussed in conjunction with their within-collection coordination scores, in the section on Activity Analysis on Experimental Data.

As significance tests, all of these statistics are meant to detect the likelihood that whatever is shared exceeds what might happen by chance. Should that similarity extend through the whole average rating time series? The Coordination Scores of Activity Analysis offer a greater sensitivity to the influences of music without requiring a common contour. They are also more robust to the collection parameter differences than the other statistics considered here, and express degrees of coherence in a scale that is comparable from one collection of responses to the next. But all of these measures evaluate collections of responses over their entire duration, whereas the most promising opportunities of continuous ratings are in their description of responses from second to second.

Alternatives to Local Activity Coordination

With the variability present in continuous response collections, the agreement between responses cannot be assumed to be constant over the course of the stimulating music. The question of which moments are particularly interesting or coherent in continuous responses has been around since the beginning of digital continuous rating experiments (e.g., Capperella-Sheldon, 1992). Here we compare the local-activity coordination test with two other measures proposed in recent years: Schubert's (2007) second-order standard deviation (second-order SD) and a modified Wilcoxon test defined by Grewe et al. (2007). These measures consider different aspects of continuous response collections, whereas their utility depends on their reliability as indicators of interesting response behaviors.

Second-order SD. To assess the validity of the average time series from one moment to the next, Schubert (2007) proposed a quantification of local inter-response coherence for a collection of continuous ratings. By this method, ratings are considered to be in "good agreement" if the variance

of rating values at that moment is lower than some collection-specific threshold. This threshold is defined in terms of the distribution of a collection's standard deviation across all time points in the responses. Publications employing this assessment of local reliability have used different thresholds to identify moments with little dispersion across responses: from one standard deviation below the mean of this distribution (Schubert, 2007) to a standard deviation above the mean (Schubert, 2013).

This calculation is not a statistical test of significance: it does not evaluate how the data collected compare to some defined random alternative. As a measure of local coherence, it is entirely relative to the collection in question; the variability of one moment in a particularly noisy collection can be counted as being in good agreement, whereas the same degree of dispersion would be a cause for exclusion in a more consistent collection. Still it is a simple calculation to highlight moments showing higher inter-response agreement on rating values and can be used to localize moments at which ratings converge to some extent after periods of higher disagreement.

Figure 5D shows the moments in an excerpt of "Morning Mood" from Grieg's *Peer Gynt Suite* selected as falling within good agreement, given a few different definitions of "good." The first moments are tagged as highly coordinated (black dots), because the response interface resets the rating marker to the midpoint of the rating range before the stimulus begins to play. Participants move the marker away from the origin once they have the opportunity and sufficient reason to report another value. As mentioned above, the impact of this initial orientation period can be misleading for some calculations (including the second-order SD). In this case, however, excluding the first 12 s of ratings does not change the point at which all responses concentrate to fall within the mean plus one standard deviation. A lower threshold of the mean minus one standard deviation produces three distinct intervals, marked by asterisks: shortly following the full orchestra entry on the theme at 51 s, the following instance of the horns suggesting the return of the full orchestra at

93 s, and a fleeting moment at 146 s.

Modified Wilcoxon test. Looking for moments of affect change in physiological measures of response and continuous ratings of felt emotion, Grewe et al. (2007) suggested a test for moments of extreme agreement. The 90th percentile of each first-order differenced time series sampled at 1 Hz was determined and then the median of these values across listeners was taken as a threshold. Subsequently, moments with median values above the threshold were collected, and the differenced rating values at these moments were then evaluated using Wilcoxon's signed rank test.

Considering the variability of responses in most of these collections, this initial criterion is strikingly stringent: only moments in which the majority of responses show changes in the same direction give rise to nonzero values in the median of the first-order differenced ratings. Of those, selecting only the time points with values over the median of the differenced responses' 90th percentiles, we are guaranteed that very few moments will qualify.

This selection criterion attempted to reduce the risks associated with multiple comparisons: in applying the Wilcoxon test to every time point according to an uncorrected significance estimate, we would have to expect false positives, and researchers often choose instead to be selective about the number of moments tested. However, choosing to test only the outliers of the distribution of a related dependent variable is effectively equivalent to applying said test to all samples. Even in random distributions, there are extremes, and these extremes are those most likely to give false positives in significance tests.

Following the methodology outlined in the initial use of this test (Grewe et al., 2007), and as demonstrated in Figure 5C, very few moments qualify as significant events, even for highly coordinated collections according to other measures of coherence. Only one moment, moving into

the first tutti, is selected as an exceptional event, the same moment that ushers in the Second Order SD's intervals of good agreement. A lower threshold at the 80th percentile is more informative, marked with grey circles in figure 5C. Given the number of responses in the collection, the probabilities involved suggest that even one such moment is highly unexpected, $p < .001$. This looser threshold identifies many moments of increased arousal reported to this excerpt of "Morning Mood," with highlights aligning with many moments marked as having extreme high activity levels as well.

Local-activity coordination. Compared to the percentile threshold described above, the local-activity coordination test of Activity Analysis points to even more distinct moments of activity behavior. Figure 5B reports moments of both locally extreme high and low activity levels. The black dots mark time points at which fewer responses showed increases than we would expect given the rating-increase patterns in the surrounding 60 s. This lack of increasing may be related to active decreases, say before 80 s, or stability in ratings, following 60 s. Another difference between the Modified Wilcoxon tests and the local activity estimate can be seen around 17 s. Here the Wilcoxon test would necessarily fail because the median is, in fact, zero, but the time regional sensitivity of the local activity coordination test picks up on the minority of responses showing coherent rating increases with the end of the oboe's first reply to the main theme.

These different ways of evaluating moments of cohesion over the time course of a collection of ratings each serve slightly different purposes. However, even with adjustments made to enhance the amount of information conveyed through distinct intervals of rating qualities detected, the Local-Activity Coordination test of Activity Analysis appears to be the more sensitive and discerning technique.

Activity Analysis on Experimental Data

Activity Analysis provides a statistical foundation for evaluating whether the ratings in a collection show synchronous changes in values, which can then be related more reliably to stimulus features at specific moments in time. As demonstrated in the previous section, Activity Analysis and its coherence measures capture useful information about coherence in collections of responses to music. These can be applied to investigate a number of different questions about these responses and the music that inspires them. Although none of the experimental data sets were collected with the explicit hypotheses of testing activity coordination, they can be useful for demonstrating how Activity Analysis can be applied, and what kinds of hypotheses might be tested in future work. Here we use some of the experiment collections to demonstrate possible applications, with examples of the variation in activity coordination related to the music, the participants, the musical interpretation, and the rating task.

Variation in Coordination Related to Stimulus

It is easy to accept that some collections of continuous ratings are more coherent than others, but where does this variation come from? We expect individual pieces of music to provoke distinct experiences in listeners; perhaps they also vary in the uniformity of these experiences reported by participants. This section explores the rating change coordination in a publicly accessible data set: continuous ratings of perceived emotion in two dimensions (Arousal × Valence) to six popular classical music excerpts, collected by Mark Korhonen (2004). A discussion on the independence of rating dimensions can be found below; for this analysis, we will treat the ratings in each dimension separately. These data have been used in multiple papers, including Korhonen et al. (2006) and Coutinho and Cangelosi (2009), to train and evaluate models of the average emotional valence and arousal time series using continuous stimulus features. The following analysis suggests that efforts to model averaged emotion ratings may be compromised by stimulus-dependent

variability in inter-participant agreement.

Table 4 reports the rating change activity coordination scores for increases and decreases along each dimension of the perceived emotion ratings for the stimuli of this data set. The Coordination Scores across the 35 participants vary a great deal from piece to piece, from 0.9 to 16, the maximum value for this implementation of coordination scores, and the coordination of one dimension does not seem to determine the coordination of the other.

The most dramatic contrast between the Arousal and Valence dimensions of ratings in the data set is to an excerpt from Rodrigo's *Concierto de Aranjuez*, plotted in Figure 6. The reported perceived emotional arousal (see Fig. 6A) gives rise to reasonably high coordination in arousal increases (14) with clear alternations between activity levels of increases and decreases (Fig. 6B). The valence dimension of these ratings (see Fig. 6C), however, has very weak coordination: C Scores of only 1.9 for increases and 0.9 for decreases. Many moments in the activity-level time series show participants' responses simultaneously moving toward opposite ends of this bipolar scale (Fig. 6D). These ratings suggest that participants were split on whether they heard the excerpt as positively or negatively valenced, and different people interpreted certain moments as having opposite implications for their respective assessments.

[Insert Table 4]

In the case of valence, the average rating does not give a representative description of these responses. The flat line, shown in black in Figure 6C, is a misleading descriptor of the emotional timeline of this piece, which is both dynamic and ambiguous. In other cases, a flat average may be a genuine representation of a set of responses that are simply very stable over time, like the arousal ratings of the excerpt of Copland's *Fanfare for the Common Man* (not shown). Here the rating changes are significant but do not have very high C Scores (Stimulus 3 in Table 4), because the

shifts in the music's emotional character were not sufficiently dramatic to provoke many simultaneous rating changes in most participants.

In contrast, the most highly coordinated collection in this set is the perceived emotional arousal ratings to the Liszt excerpt shown in Fig. 2A. Here the average varies widely over the range of the rating scale and the activity-level distributions for increases and decreases in this rating collection yield maximum Coordination Scores of 16. And yet, even in these activity-level times series (Fig. 2D), it is still very rare that a majority of responses show concurrent supra-threshold rating changes in either direction.

That activity levels fail to reach unanimity in moments of change reflects a reality of continuous rating data: participants rarely report changes at precisely the same time, nor in the same way, and the ambiguity of a task like rating perceived emotion continuously via the position of a mouse cursor cannot be distinguished from differences in perception. Consider the interval between 225 s to 250 s of the Liszt excerpt. The average rating time series is monotonically increasing (Fig 2A) and activity-levels for rating increases in 2-s time frame include 20% or more throughout; however, a subset of participants also report some decreases against this trend (Fig. 2D). Did these participants notice something in the music the others missed, or were they inclined to report smaller and faster changes than their peers? Even in the most coordinated collections, there is variety in the responses.

The Coordination Scores for collections in the Korhonen data set show that we cannot assume that continuous ratings to a musical stimulus are coherent: one set of participants can vary dramatically in the coordination of their ratings from piece to piece. If coherence is low, the relevance of an average rating time series is questionable. Variability or consistency in the rating responses reported may be a quality of the music itself.

Coordination Between Participant Groups

The question of rating change coordination came out of a very challenging data set involving two audiences reporting felt emotional intensity to related stimuli. As described previously, the Boston Symphony Orchestra experiment collected continuous ratings of felt emotional intensity from participants attending a live concert and others gathered in a recital hall to hear and watch a high-definition video recording of the performance. Here we have a test case for assessing consistency in responses to music as these two groups experienced very similar stimuli. Interpretation of these data encountered the usual variability in the responses collected. Some listeners reported no change in felt emotion for some of the stimuli, leaving 30 to 32 dynamic respondents in the live-concert audience, depending on the piece, and 21 to 23 in the recorded-concert group. The intention had been to compare average time series of the two groups. However, it was difficult to know what differences might be significant given the seeming incoherence in the remaining ratings.

In contrast to data sets like Korhonen's, the coordination of these collections ranged from low to medium. Table 5 shows the Coordination Scores for rating increases and decreases in felt emotional intensity reported by each group to four Mozart excerpts. Considering their respective within-collection C Scores, ratings to the Overture to *The Marriage of Figaro* seem to be the most coordinated. The *Jupiter Symphony* Finale (K551) also reaches similar degrees of activity coordination for the live audience, but not the other one, and the middle two excerpts are near or below significance thresholds for decreases or changes in both directions. Across these excerpts, it seems the ratings to the video recording were less coordinated than those collected during the live session. Such an analysis approach could lead to explorations for the reasons behind this difference.

[insert Table 6]

The coordination *between* these collections, reported under Bi-C Scores in the last two columns of Table 5, can be interpreted in conjunction with the within-collection rating-change activity. Ratings to the *Figaro* Overture (K492) were coordinated in rating changes within each audience's collection of responses and across them. At the other extreme, there was no significant rating-change activity coordination between the audiences in either direction for K16, the Rondo to Mozart's First Symphony, as might be expected given the low scores for this collection. The low but significant coordination in only rating increases within the collections for the *Clarinet Concerto* excerpt (K622, Adagio) was reflected in the between-collection coordination. However, although the between-collection scores for the Jupiter Symphony Finale looked like those of the preceding piece, the within-collection coordination scores were quite different, with the live audience showing markedly more agreement in rating change activity than the other one. This combination of activity coordination scores suggests some substantial differences between the experiences reported by these two groups.

Figure 7 shows the two collections of responses to the *Jupiter* Finale and their rating-change activity. The highest activity moments for both audiences are in the first 25 s, with activity-levels of increases peaking shortly after 15 s (see Fig. 7C). Besides this moment, both collections' activity levels are quite low, with rarely as much as a quarter of responses reporting either type of rating change at once. Although there are a few shared moments of rating decreases with relatively high activity levels, say above .15, there are many more with similar activity levels in one but not the other audience, such as at 70 s or 298 s. Altogether, these activity levels look relatively independent, producing a low Bi-Coordination Score of 0.9 (dec). The curves of the average rating time series reflect these small disagreements in activity between the two collections. After the first large increase at the beginning, the shifts in these times series are quite shallow, and they rarely align.

There are many factors that could contribute to these activity coordination results. As with the Korhonen data set, there are important differences between the stimuli: K492 is dynamic and easy to follow, whereas the Rondo of the First Symphony is trite and emotionally flat. These collections also report lower activity Coordination Scores than other data sets, perhaps a result of setting: musical experiences in a concert hall may be quite different from those experiencing a recorded video in a recital hall without the excitement of live musicians. The two groups were also composed of different populations, and only for the live audience group were children present (it was a family-oriented concert). And yet, despite a number of complications, Activity Analysis has found coordination in the dynamics of participants' felt emotional intensity, both within and between groups, coordination that could only have come from the music presented.

Comparing Performances with Local Activity Coordination

The question of when a piece of music moves listeners is important. We study continuous ratings because they can capture these shifts in perception and reaction on a second-by-second basis, and many of these shifts of feeling depend on musicians' specific interpretations of a work. Activity Analysis can be used to identify when the music presented has affected a significant proportion of collected responses, and also to consider what distinguishes the impacts of individual takes on the same musical material (Farbood & Upham, 2013).

From the CIRMMT Audience Response System (CARS) experiment (see Appendix 1), we have ratings of felt emotions from different participant groups to two interpretations of a madrigal by Arcadelt, one recorded and one live. The arousal dimension of these felt emotion ratings was significantly coordinated for increases to both versions: parametric coordination scores of 11 to the recording and 3.0 to the live performance. The nonparametric activity coordination test, which uses the same statistical assumptions as the local-activity coordination test, also points to significant coordination with 3.3 (maximum value for 2000 iterations) on rating increases in both collections.

With this assurance of coordinated activity for increases of at least .025 of the felt emotional arousal rating scale over 2-s time frames, counted in 2-s windows of synchrony, we can use the significance estimates on the activity levels per time frame to pick out moments of exceptional high and low activity level.

Comparison of interpretations requires alignment in time. With recordings of both interpretations, we hand annotated every note onset and linearly interpolated these to get the timing of 16th notes over the duration of the music. The rate of these 16th notes averaged 6 Hz and never fell below 2 Hz. A 16th-note sampling of the ratings and the activity-level time series were then taken from the original 10 Hz times series of responses using nearest neighbor values. Figures 8A and 8C plot each collection of ratings in this shared metrical time, Rsp (Rec) and Rsp (Live), and Figure 8B shows their activity levels for increases in overlapping 2-s time frames sampled in this shared musical time line. The responses to the King's Singers' interpretation is plotted above zero with moments of significant local activity levels (Rec X-Act), and responses to the live performance of the Orpheus Singers are plotted below (Live X-Act). These moments of locally extreme activity levels rank above the 97.5 percentile or below the 2.5 percentile of the random alternatives for each specific 2-s time frame from the 2000 alternative activity levels generated by breaking stimulus alignment

[insert Figure 8]

Looking at Figure 8B, there are a few moments of similar behavior: popular increases in arousal in mm6-8, in mm13-14, and a similar lack of increases around m 21 and m 46. An obvious difference is the rush of increases in emotional arousal ratings to the live performance after m 10. Here the choir did a rapid *crescendo* and *subito piano* at the beginning of a line, whereas the King's Singer's version stayed quiet. A similar contrast in intensity aligns with the flourish of increases reported by those hearing this recording after m 15. At other times, the differences are not so

obvious, say mm22-27, when smaller subsets of participants (<27% and <36%) are concurrently increasing rating values in each collection. The details of these performances and resultant ratings deserve much more detailed analysis than is relevant to the scope of this methodological paper.

By focusing on a specific active component of continuous ratings, it is relatively easy to identify when the performances are drawing distinct reactions from these participants. Activity Analysis makes use of proportions and popularity to pick up moments of importance, potentially moments the performers intended to cause specific reactions. Despite the distinct timing of each performance and different groups of participants, we have aligned their activity levels and the results of the local-activity coordination estimates and could do the same for more subtle contrasts as well.

Continuous Ratings in Two Dimensions

Concerns have been raised about the quality of ratings collected using two-dimensional interfaces: Can participants assess and report their responses on two scales simultaneously? What is the influence of one dimension on the other? Activity Analysis can evaluate the interactions between the dimensions to get a sense of whether participants can treat each dimension independently. Of the data sets included in this study, two employed continuous ratings of emotions in two dimensions: Korhonen's, as discussed previously, and the CARS experiment, run over three sessions at McGill University in 2009.

In the first McGill session, responses to three recorded pieces were collected from a group of 45 participants, of which one-third rated felt arousal, one-third felt valence, and the last third rated both dimensions simultaneously on a 2D interface. Although only three musical stimuli were presented, this data set is a useful starting point for exploring how the task of rating two dimensions compares to that of rating only one.

First, we look at whether the activities within these collections are comparable. Figure 9A reports the Coordination Scores for rating increases and decreases of each dimension per response condition (gray bars) and the coordination between the ratings collected on 1D and 2D interfaces (Bi-C score, black bars) per stimulus. With only 15 ratings per collection, individual participants' rating techniques can add a lot of noise to the C Scores, and yet there is still substantial similarity in these results with a Pearson correlation of $r(10) = .77$, $p < .005$, across the 12 dimension X stimulus combinations. Additionally, the degrees of coordination seem fairly similar: Valence dimension ratings for S1 and S3 seem uncoordinated, with barely a trace of agreement in activity, whereas the increases and decreases in the arousal dimensions range between 3.5 and 6.7. And these dimensions of emotion are no more or less coordinated when collected simultaneously, according to a paired t-test, $p = .62$, df = 11.

The activity within each 1D collection of 15 ratings also seems to share a reasonable amount of activity with their 2D counterparts. Taking the average C Scores per stimulus, dimension, and direction of rating change, these correlate strongly with the Bi-C score values between conditions (black bars of Fig 9A), $r(10) = .87$, $p < .001$. Across three very different stimuli, a Renaissance madrigal, a Romantic string quartet movement, and a semi-structured improvisation on an electronic instrument, there is a notable amount of consistency in the coordination reported by stimulus and dimension of felt emotion rating, regardless of whether these ratings were collected using a one- or two-dimensional interface.

[Insert Figure 9]

But the task of rating two dimensions simultaneously might still result in unexpected interactions between them. Perhaps there is a chance of participants confusing the dimensions or collapsing them into a simple one-dimensional combination such as along one of the diagonals of the 2D interface. To check on this, Figure 9B reports the between-collection coordination scores

(Bi-C Scores) calculated between the dimensions of arousal and valence reported by those rating both at once (black bars) and those rated by separate participants on 1D scales (gray bars). Ignoring direction of rating change (far left of Fig. 9B), ratings from 2D raters show more concurrent activity in two of the three stimuli (S2 and S3) than do the 1D groups (grey bars). Thus, these participants often reported changes along some diagonal across these emotion axes, easy to do on the handheld touch-screen interfaces used in the CARS experiments. The interface did encourage more concurrent changes in both dimensions than happened in the physically independent ratings.

However, these simultaneous changes did not produce a systematic alignment or confusion between the felt emotion dimensions in the 2D ratings. The oriented rating-change activity Bi-C Scores (the other four conditions in Fig. 9B) mostly fall below significance. Given the overall low coordination of the valence dimension of ratings in this data set, this might not be convincing on its own. Thus we also report the results of between-dimension coordination for the 2D ratings in the Korhonen dataset in Figure 9C. That a few stimuli show oriented alignment, say between decreases in both arousal and valence in the Strauss excerpt (A/V dec, in black), may be a quality of that specific piece of music, rather than a systematic problem with the task of reporting two dimensions at once. These results support the view that participants are capable of reporting felt and perceived emotions on these two dimensions concurrently, making independent assessments of each.

Continuous Ratings of Emotional Valence and Arousal

As discussed above, continuous ratings of emotion to music are often collected in a 2D Arousal × Valence emotion space. In emotion and music research, these two dimensions are very common representations of emotion (Eerola & Vuoskoski, 2011), but they may not be equally important or consistent in listeners' perceived or felt responses to music. It has been argued that emotional arousal is sensitive to universal cues (Becker, 2010) such as loudness and pulse rate.

Many common musical cues for emotional valence appear to be learned, such as culturally specific scales and lyrics. Such a distinction between a universal arousal response and a culture-specific valence response has been reported in a cross-cultural comparison between Canadians and Congolese Pygmies (Egermann, Fernando, Chuen & McAdams, 2015).

Besides the dataset used in the previous section, the last CARS session also collected continuous responses along these two dimensions, although some responses were lost due to technical complications. Between the Korhonen and the CARS data sets, we have continuous emotional arousal and valence ratings for 11 stimuli. A one-tailed *t*-test shows the arousal dimension ratings to have significantly higher within-collection Coordination Scores than the valence dimension ratings, $t(10) = 3.18$, $p = .005$, consistent with visual comparison of the left and right halves of Figure 9A. Although the lower coordination of alternating activity suggests that this difference is due to greater ambiguity in the valence dimension, as discussed with the ratings to the Rodrigo excerpt in Fig. 6, another important factor may affect these scores. The coordination measured here depends on the quantity of change, and the dimension with more noticeable changes will show more stimulus synchronous coordination. It may be that valence is less dynamic, changing less frequently and less dramatically, than emotional arousal for these stimuli. Both ambiguity and rate of change in these dimensions of emotion need to be explored further if we hope to understand and anticipate how listeners might respond to other pieces of music.

Conclusions

Activity Analysis takes a distinct approach to continuous responses to music by evaluating specifically when participants show changes in their experience of what they hear. It focuses on one kind of reaction or action at a time. We are not all affected by music in the same way, nor are our responses necessarily the same each time we hear a recording or performer present a piece. This ambiguity is expressed in the variation across collections of continuous responses. With activity-

level time series comes an opportunity to consider the popularity of these reactions to a piece of music. The coordination tests of Activity Analysis work around the ambiguity in responses to a recording or performance to determine whether they show sufficient coherence within the shared timeline of the stimulus, whether there might be similar patterns of activity between collections of responses to the same stimulus, and when during these continuous responses a remarkable number of responses agree in their active reactions to the music presented.

We find that coordination varies from piece to piece. Some music prompts timely increases and decreases in ratings; other pieces fail to produce rating changes distinguishable from those of responses that are independent of the shared stimulus. Activity-level time series expose contradictory reactions to pieces, challenging the common practice of averaging ratings before exploring relationships to musical features. With coordination scores, we show that participants are capable of simultaneously rating emotional arousal and emotional valence as independent dimensions to concert music. Also with the local-activity coordination test, we demonstrate the potential for exploring different performances of a single piece through the ratings of listeners.

The concept of Activity Analysis is simple: isolate some kind of response event, count when it co-occurs across responses to the same music, and, for statistical assessment, compare these counts to what might happen by chance. To help with implementation, we have released an Activity Analysis Toolbox in MatLab (Upham, 2017) with demonstration scripts using the public domain Korhonen data set. Anyone interested in using these functions in other data analysis platforms are welcome to get in touch.

There is a great deal to explore in continuous responses to music with Activity Analysis. Looking at ratings, we can consider the differences between different kinds of events. Are decreases of perceived emotional arousal more or less coordinated than increases of ratings to classical music? We've treated them here as symmetric, but many perceptual processes are more sensitive in

one direction of change. Populations may differ in the coordination of their rating change activity, say between musicians and nonmusicians, or people more or less familiar with the genre or piece. Analyses of populations have also found differences in their sensitivity to specific cues. With the right stimuli, such differences should be expressed in the activity of tonal tension rating changes at specific moments. Perhaps the impact of program notes and lyric translations (Hackworth & Fredrickson, 2010) or combinations of audio and video recordings might be better explored with measures of coordination. The added information may increase inter-response coherence without changing the contour of the average response.

More could be learned with the replication of continuous rating experiments. Is it a mere coincidence that the Overture to *The Marriage of Figaro* produced similar coordination scores in two separate groups of raters? In technical terms, this means that there are similar mixes of distinctly coordinated moments and potentially random coincidences of rating changes. We are interested in the degree to which the music nudges or pushes responses to change beyond the variability of individual ratings affected by extra-musical factors specific to each listener. A couple of studies reported here suggest that Coordination Scores may be relatively consistent, or at least rank-ordered per stimulus, but more examples would be welcome. For instance, Korhonen's (2004) data were collected as a replication of Schubert's (1999) study of continuous ratings of emotion, with a similar interface and identical stimuli. In this paper, we report the Coordination Scores for rating increases and decreases for this group of 35 participants. It would be interesting to know whether the responses in the original experiment resulted in the same pattern of coordinated activity across stimuli.

The local-activity coordination test may be particularly useful for exploring the influences of specific features of music on listeners, be they factors of performance practice, composition, or even cross-modal influences. As shown in the responses to the Arcadelt madrigal, specific interpretations

can prompt different moments of locally extreme high or low activity levels but the causes for these reactions warrant more explanation.

Activity Analysis can also be applied to other continuous measures of experience. The nonparametric tests offer a practical view into the timing of events in psychophysiological responses that are also subject to complicated temporal characteristics, such as the cyclical demands of respiration or the changing responsiveness of skin conductivity. With the appropriate adjustments of analysis parameters (the definition of events, the size of the window of synchrony, and the shuffling range), Activity Analysis can point to coordination across collections of involuntary processes. Appendix 3 demonstrates the process of tuning analysis parameters to specific signal qualities of experimental data.

Activity Analysis does have limitations. If the responses are too short, say under 100 s, some of the statistical tests cannot be applied. (The drop off point depends on the event rate and the number of responses in the collection.) The activity-level time series may be useful to look at patterns of concurrent events across responses, but we recommend reserving the tests of coordination for rating responses of 2 minutes or more. And as there is still more to be learned about the behavior of Coordination Scores on ratings to music, interpretation should always be pursued with caution.

We hope that the analysis of other continuous rating coherence measures is informative, both for interpreting other published work with the numerical estimates of false-positive thresholds and for its discussion of factors defining the advantages and disadvantages of the respective calculations. Depending on one's interests, evaluating coherence with two different measures may be a reasonable course of action. The construction of Coordination Scores and the nonparametric local-activity coordination test seem to focus on aspects of response coherence not easily handled by other coherence measures. It may be a good complement to measures such as the

variance ratio (VarRatio), Cronbach's α, or mean correlation (MeanCorr).

A systematic study of coherence in continuous ratings to music would not be possible without access to the response collections in several experiments by a number of researchers. Mark Korhonen set a precedent by making his data openly available and his example is an important one to follow. When working with complex stimuli and responses, we can learn a great deal from other data sets. Until we have more complete models to describe the behaviors of individual responses to the wonderful stuff of music heard, these measures can give us perspective on the real variability involved and make it possible to tune analysis parameters to the qualities of continuous ratings. The exploration of other types of coherence and their sensitivity to stimulus duration and the number of responses could be elaborated with more examples of ratings to music. As such, we encourage the sharing of data sets.

Activity Analysis can answer standing questions about continuous responses to music and raise others. With the release of the Activity Analysis Toolbox, we encourage other researchers to consider applying these descriptive and inferential statistics to data previously collected as well as to future experiments.

References

BACHORIK, J., BANGERT, M., LOUI, P., LARKE, K., BERGER, J., ROWE, R., & SCHLAUG, G. (2009). Emotion in motion: Investigating the time-course of emotional judgments of musical stimuli. Music Perception, 26, 355–364.

BAILES, F. & DEAN, R. T. (2012). Comparative time series analysis of perceptual responses to electroacoustic music. Music Perception, 29, 359–375.

CAPPERELLA-SHELDON, D. A. (1992). Self-perception of aesthetic experience among musicians and non-musicians in response to wind band music. Journal of Band Research, 28, 57–71.

CHAPIN, H., JANTZEN, K., KELSO, J. S., STEINBERG, F., & LARGE, E. (2010). Dynamic emotional and neural responses to music depend on performance expression and listener experience. PloS ONE, 5(12), e13812. doi: 10.1371/journal.pone.0013812

COUTINHO, E. & CANGELOSI, A. (2009). The use of spatio-temporal connectionist models in psychological studies of musical emotions. Music Perception, 27, 1–15. doi: 10.1525/mp.2009.27.1.1

CRONBACH, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334.

DEAN, R. & BAILES, F. (2010). Time series analysis as a method to examine acoustical influences on real-time perception of music. Empirical Musicology Review, 5(4), 152– 175.

DEAN, R., BAILES, F. & DUNSMUIR, W. (2014). Time series analysis of real-time music perception: approaches to the assessment of individual and expertise differences in perception of expressed affect, Journal of Mathematics and Music, 8, 183-205. doi:

10.1080/17459737.2014.928752.

DEAN, R., BAILES, F., & SCHUBERT, E. (2011). Acoustic intensity causes perceived changes in arousal levels in music: An experimental investigation. PloS one, 6(4):e18591.

DUDEWICZ, E. & MISHRA, S. (1988). Modern mathematical statistics. New York: John Wiley & Sons.

EGERMANN, H., FERNANDO, N., CHUEN, L. & MCADAMS, S. (2015). Music induces universal emotion-related psychophysiological responses: Comparing Canadian listeners to Congolese Pygmies. Frontiers in Psychology, 5:1341. doi: 10.3389/fpsyg.2014.01341

EEROLA, T. & VUOSKOSKI, J. (2011). A comparison of the discrete and dimensional models of emotion in music. Psychology of Music, 39, 18–49. doi:10.1177/0305735610362821

FARBOOD, M. M., HEEGER, D. J., MARCUS, G., HASSON, U., & LERNER, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. Frontiers in Neuroscience, 9:157.

FARBOOD, M. M. & UPHAM, F. (2013). Interpreting expressive performance through listener judgments of musical tension. Frontiers in Psychology, 4:998. doi: 10.3389/fpsyg.2013.00998

FREDRICKSON, W. (1995). A comparison of perceived musical tension and aesthetic response. Psychology of Music, 23, 81–87.

GREWE, O., NAGEL, F., KOPIEZ, R., & ALTENMÜLLER, E. (2007). Emotions over time: Synchronicity and development of subjective, physiological, and facial affective reactions to music. Emotion, 7, 774–788. doi: 10.1037/1528-3542.7.4.774

GRÜN, S. (2009). Data-driven significance estimation for precise spike correlation. Journal of

Neurophysiology, 101, 1126–1140. doi: 10.1152/jn.00093.2008

HACKWORTH, R. & FREDRICKSON, W. (2010). The effect of text translation on perceived musical tension in Debussy's Noël des enfants qui n'ont plus de maisons. Journal of Research in Music Education, 58, 184-195. doi: 10.1177/0022429410369835

KORHONEN, M. D. (2004). Modeling continuous emotional appraisals of music using system identification. (Unpublished Master's thesis). University of Waterloo.

KORHONEN, M., CLAUSI, D., & JERNIGAN, M. (2006). Modeling emotional content of music using system identification. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 36(3), 588–599. doi: 0.1109/TSMCB.2005.862491

KRUMHANSL, C. L. (1996). A perceptual analysis of Mozart's piano sonata K. 282: Segmentation, tension, and musical ideas. Music Perception, 13, 401–432. doi: 10.2307/40286177

LEVITIN, D., NUZZO, R., VINES, B., & RAMSAY, J. (2007). Introduction to functional data analysis. Canadian Psychology, 48(3):135.

MARRIN NAKRA, T., & BUSHA, B. F. (2014). Synchronous sympathy at the symphony: Conductor and audience accord. Music Perception, 32, 109-124.

MCADAMS, S., VINES, B., VIEILLARD, S., SMITH, B., & REYNOLDS, R. (2004). Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. Music Perception, 22, 297–350.

NIELSEN, F. V. (1987). Musical 'tension' and related concepts. In T. Sebeok & J. Umiker-Seboek (Eds.), The semiotic web '86: An international yearbook (pp. 491–513). Berlin: Mouton de Gruyter.

PYPER, B. J. & PETERMAN, R. M. (1998). Comparison of methods to account for autocorrelation in

correlation analyses of fish data. Canadian Journal of Fisheries and Aquatic Sciences, 55, 2127–2140.

RODGERS, J. & NICEWANDER, W. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1), 59–66. doi: 10.1525/mp.2004.22.2.297

SCHUBERT, E. (1999). Measurement and time series analysis of emotion in music (Unpublished doctoral dissertation). University of New South Wales.

SCHUBERT, E. (2004). Modeling perceived emotion with continuous musical features. Music Perception, 21, 561–585. doi: 10.1525/mp.2004.21.4.561

SCHUBERT, E. (2007). When is an event in a time-series significant? In Schubert, E., Buckley, K., Eliott, R., Koboroff, B., Chen, J., and Stevens, C. J. (Eds.), Proceedings of the inaugural International Conference on Music Communication Science (pp. 135–138). Australia: ARC Research Network in Human Communication Science.

SCHUBERT, E. (2013). Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music. Psychology of Music, 41, 350-371. doi: 10.1177/0305735611430079.

TOIVIAINEN, P. & KRUMHANSL, C. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. Perception, 32, 741–766. doi: 10.1068/p3312

TORRES-ELIARD, K., LABBE, C., & GRANDJEAN, D. (2012). Towards a dynamic approach to the study of emotions expressed by music. In A. Camurri, C. Costa, & G. Volpe (Eds.), Intelligent Technologies for Interactive Entertainment, LNICST 78 (pp. 252–259). Berlin: Springer.

UPHAM, F. (2012). Limits on the application of statistical correlations to continuous response data.

In E. Cambouropoulos, C. Tsougras, P. Mavromatis, & K. Pastiades (Eds.), Proceedings of the 12th International Conference on Music Perception and Cognition (pp. 1037-1041). Thessaloniki: Aristotle University of Thessaloniki.

UPHAM, F. (2017). Activity Analysis MatLab Toolbox [Code repository]. Retrieved from https://github.com/finn42/ActivityAnalysisToolbox_2.0 (last accessed June 1, 2017).

WILLIAMS, L., FREDRICKSON, W., & ATKINSON, S. (2011). Focus of attention to melody or harmony and perception of music tension: An exploratory study. International Journal of Music Education, 29(1), 72. doi: 10.1177/0255761410372725

YESHURUN, Y., CARRASCO, M., & MALONEY, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. Vision Research, 48, 1837– 1851.

Appendix 1: Data Sets

The development of Activity Analysis and this study of continuous ratings to music were made possible with the ratings of a number of experiments. This appendix shares details on the different sets of response collections used in this paper, first the experiment data sets and finally the randomly generated collections of unrelated responses.

The Angel of Death Project

Eight collections of continuous responses were gathered as part of an intricate experiment involving two premier concerts on different continents, each presenting two versions of Roger Reynolds' *The Angel of Death*, a 35-minute piece for piano, orchestra, and computer-processed sound (McAdams, Vines, Vieillard, Smith & Reynolds, 2004). The first concert was presented in 2001 at the Grande Salle of the Centre Georges Pompidou in Paris, France with the Ensemble Court Circuit, the other at the Mandeville Auditorium of the University of California, San Diego in La Jolla, California in 2002 with the Sonor Ensemble. In both concerts, one set of participants continuously reported their felt emotional intensity (*force émotionnelle* was the term in the Paris concert/experiment) and the other group reported the resemblance (*familiarité*) of the current musical material as compared to anything heard from the beginning of the piece. Participants were diverse in age and musical expertise, and recorded their responses on handheld slider potentiometers wired to their seats in each concert hall. These responses were sampled at 2 Hz, and some outlier responses, such as those that reported no change for several minutes, were discarded prior to analysis. The details of the original collections are presented in Table A.1.

Compared to most experiments on continuous responses to music, these responses are exceptionally long. For the purposes of this paper, responses in each collection were truncated to their first 500 s. The script generating the unrelated response collections had access to all of the

original responses listed in Table A.1.

[insert Table A.1]

Boston Symphony Orchestra Project

The Boston Symphony Orchestra project collected continuous responses from two sets of participants at a concert in 2006 by the Boston Symphony Orchestra under the direction of Maestro Keith Lockhart. One set of participants were part of the audience at a live performance in Boston, Massachusetts, and the other collectively watched and heard a video recording of the performance in the Tanna Schulich recital hall in Montreal, Quebec. All participants reported felt emotional intensity on the same handheld potentiometer sliders used in *The Angel of Death* project. They rated their experience of all pieces in this program celebrating the works of W. A. Mozart (see also Marrin Nakra & BuSha, 2014). From this set, eight response collections were drawn, each groups' ratings of the four orchestral pieces by Mozart. Responses that showed little to no change over each piece were discarded from each collection. The details of each collection are listed in Table A.2.

[insert Table A.2]

Ratings from this data set are discussed in more detail as examples in Figures 3, 4, and 7, and their rating change activity coordination scores, both within and between collections, can be found in Table 5.

Korhonen's Perceived Emotion Ratings

In an experiment following the paradigm and material of Emery Schubert's earlier experiments (Schubert, 1999), continuous ratings to music were collected using the EmotionSpace Lab, a 2D graphical interface on a computer screen for reporting evaluations of emotional arousal and valence jointly as a position in a square with a mouse cursor (Korhonen, 2004). Thirty-five

participants listened and rated edited excerpts of popular classical music taken from the Naxos CD *Discover the Classics*, Vol. 1. These participants varied in age, musical expertise, and familiarity with the genre of music, termed instrumental art music. Korhonen has made these response collections openly available for subsequent analysis. The characteristics of these ratings and their relationship to the stimuli have been discussed in previous publications (Korhonen, Clausi, & Jernigan, 2006; Coutinho & Cangelosi, 2009). Table A.3 provides more details.

[insert Table A.3]

Ratings from this data set are discussed in more detail as examples in Figures 1, 3, 5, and 7, and their rating change activity coordination scores can be found in Table 4 and Figure 9.

McAdams' CARS Felt Emotion Ratings

As part of an experiment on continuous responses to music from an audience of participants, the CIRMMT Audience Response System (CARS) was used to gather continuous ratings of felt emotion in two different conditions. Participants in the March 2009 session were drawn from the Schulich School of Music. One third of them rated felt emotional arousal, another third rated felt emotional valence, and the last third rated both dimensions simultaneously, all via touch screen interfaces on iPod Touch devices while listening to recordings of concert music. Participants in the October 2009 session were members of the public and music theorists in town for a conference. A subset of participants reported felt emotions on the two-dimensional interface (Arousal X Valence) during the live performance of the same musical pieces. Of the collections reported in Table A.4, one pair was used in the generation of unrelated response collections, but not as part of the experiment response collections. A number of ratings to the live performances of the Schumann string quartet excerpt were lost due to equipment malfunction, leaving too few responses for some coherence measures to yield comparable results to the other collections.

[insert Table A.4]

Ratings to the two versions the Arcadelt madrigal are presented in Figure 8, while the coordination scores of ratings to the recorded stimuli are reported in Figure 9.

Farbood's Liking Ratings to Scrambled Brahms

Continuous ratings of liking were collected as part of a larger study on the neural processing of musical structure (Farbood, Heeger, Marcus, Hasson, & Lerner, 2015). Participants heard multiple versions of an excerpt of a Brahms piano concerto that had been segmented at many hierarchical levels, sections, phrases, and bars, and was scrambled at each. Musician and nonmusician participants performed the experiment individually in a sound-proof booth, and they reported their liking for the music using a horizontal slider presented on a computer screen using a mouse. Although the scrambling of smaller blocks had a marked effect on how participants responded, in particular how often they reported changes in liking, the gentler reordering of the music at the segment level produced ratings of similar quality to that of the unscrambled version. Both collections of ratings were used to generate the unrelated response collections and as experiment response collections. Table A.5 reports the details of these collections.

[insert Table A.5]

Unrelated Response Collections

Two thousand collections of responses unrelated by stimulus and rating scale were generated by sampling randomly from the 42 experimental response collections described above with a total of 1350 continuous ratings. Following the distributions in the experimental data sets, these constructed collections varied in parameters typical of the collections available here. The number of ratings per collection were sampled around the median (31) and standard deviation

(9.4) while bounded between 15 to 40. The duration of each collections was centered on M = 251 s with SD = 150 s, and bound between 100 s and 400 s, and the sampling rate was set at 1 Hz, 2Hz, 4 Hz, or 10 Hz with equal probability. Each of these collections was then populated from experimental collections of the same duration or greater by first selecting the collection (equal probably, with replacement), and then selecting a response (equal probability). Selected ratings longer than that of the collection were truncated by cutting off the end of the response. Responses were resampled using linear interpolation to match the unrelated response collection's sample rate. These unrelated response collections are used in the figures of Appendices 3 and 4.

Appendix 2: Activity Coordination Test Procedures

This appendix describes the statistical tests as implemented in the Matlab functions of the *Activity Analysis Toolbox* (Upham, 2017).

Parametric Activity-Coordination Test: Random Alternatives

Activity Analysis offers a number of measures to assess the coordination of events across responses within and between them. All of those producing Coordination Scores are parametric; they use simple parametric models of random behavior to assess the difference between what has actually measured and what might have otherwise happened without any coordinating influence.

The within-collection Coordination Score makes use of two parametric assumptions: a parametric model of the alternative activity-level distribution and the parametric Pearson Chi-squared Goodness-of-Fit test to compare this to the measured distribution. The between-collection Bi-Coordination score uses only one, the Pearson Goodness-of-Fit test to evaluate the alternative independent joint-distribution to the measured joint-distribution of activity levels.

For the Coordination Score, the parametric model of the alternative activity-level

distribution is generated with the average event activity rate per time frame and a binomial distribution (or Poisson distribution for collections with a lot of ratings.) The activity rate of a single response is estimated from the proportion of non-overlapping time frames in the series containing an activity event. The average of this rate across all responses in the collection estimates the likelihood of any single response containing the event in any single time frame. This probability in a binomial probability function (or Poisson) can then calculate the likelihood of any activity level as a proportion of all responses active at once. The Activity Analysis Toolbox defaults to the binomial calculation. The measured activity level distribution is matched by a closest random alternative by multiplying these probabilities by the number of non-overlapping time frames in this measured activity-level time series.

This model of randomly occurring activity levels makes some assumptions about activity events. It treats activity in any given response as independent from one time frame to the next. By using time frames of size suitable for the specific event and only evaluating activity-levels in non-overlapping time frames, autocorrelation in activity series decreases. Still, not all responses in a collection show the same degree of autocorrelation, making this assumption a stretch of we looked at each alone, just as their activity rates vary from response to response.

Depending on the response and activity event, these response-wise complications get blurred out when aggregated to the activity-level distribution. This seems to be the case for rating change activity to music over 2 s time frames. But for periodic events like inspiration onsets in respiration measurements, the parametric assumptions in the binomial model of activity levels is not acceptable. For these kinds of responses and events, the non-parametric tests of activity analysis are necessary.

The random alternative for the between-collection coordination tests and Bi-Coordination Scores does not introduce any parameters. Instead, the contingency table is constructed using the

actual activity level distributions of both collections and assumes the level in one collection is independent from those in the other (i.e., not influenced by their shared musical stimulus). For any combination of activity levels of collection 1 and 2, their joint activity-level probability is simply the product of either collection's actual activity-level occurrence rates. This usually results in an even blob of common joint activity levels in the low to middle range, with very little chance of the very low and very high activity levels happening concurrently in these two collections.

Binning Algorithm for Pearson Chi-Squared Tests

There are a few ways to evaluate the differences between a measured distribution and a random model. In the parametric activity coordination tests, we use the Pearson's Chi-Square test because of its simplicity, flexibility, and popularity in data analysis code libraries.

The degrees of freedom in a Pearson Chi-Squared is not the number of samples but rather the number of categories or bins of comparison between the distributions. The more bins, the smaller the differences to be exposed, and the more likely such small differences are inconsequential. The activity-level distributions are discrete with many values (the number of responses in the collections plus one) but these are related by degree. To reduce complexity and ensure sufficient number of samples for comparison, the activity level distributions are divided into a few contiguous bins (Dudewicz & Mishra, 1988).

In the Activity Analysis Toolbox, the process of reducing the activity-level distributions to a few bins is performed by an algorithm in the function *equiSplit.m*. This algorithm cuts the distributions into a specified number of bins, usually 3 to 5, according to two objectives: (a) all bins contain a minimum of 5 samples (time frames) according to the random alternative distribution; (b) the number of samples in each bin is as close to equal as possible, again according to the alternative distribution.

If these conditions cannot be satisfied, the algorithm repeats the exercise on N-1 bins. To find a maximally even bin, the function starts with the cumulative distribution of the random alternative, segmenting it according to the number of samples, and selecting the best option from the finite list of bin edge combinations.

The binning of the joint probabilities in the between-collection coordination test applies this algorithm to cut each collection's actual activity-level distribution into three bins: low, middling, and high activity levels. If the product in any of these nine bin combinations leaves less than five expected samples, the test cannot be performed. For this reason, the between-collection test of activity coordination cannot report on responses to very short musical excerpts. Stimuli of less than 110 s are not likely to be distributed sufficiently smoothly to satisfy the statistical test criteria when using time frames of at least 2 s for rating changes. With these reductions, we calculate the Chi-Squared value of the difference between the actual and alternative distributions of activity-levels, and the resultant *p*-value given N-1 d.f. (within-collection) or 4 d.f. (between-collection).

Construction of the Coordination Score

Beyond a test result of whether a collection shows more coordination than would be expected by chance, it would be useful to be able to make some claim as to the degree of coordination. The *p*-value is a measure of deviance from expected random behavior. As such, we can use these values as the basis of a *Coordination Score* for some activity type within a collection of continuous responses. We propose then that the Coordination Score be calculated from the *p*-values of goodness-of-fit tests via a simple formula similar to that used in Yeshurun, Carrasco, and Maloney (2008): $c = -\log_{10}(p)$. Under this transformation, scores above 2 would be equivalent to $p < \boldsymbol{\alpha}_{crit} = .01$. The maximum value for a Coordination Score is set to 16, for numerical convenience, by adding 10^{-16} to all calculated *p* values.

An advantage of the goodness-of-fit test is its flexibility for different numbers of samples and sample values: the results are comparable between collections of various sizes, in duration and number of participants. Unfortunately, the calculation is sensitive to the necessary simplification of the distribution through the binning process. Also besides the issue of how many bins are used, the nonoverlapping time frames that divide the continuous responses may separate different participants' activity in response to the same musical moment. However, this temporal segmentation can be turned to our advantage. The coordination test uses nonoverlapping time frames, and when these frames are larger than the sampling period, it is therefore possible to repeat the test on the same collection at different phases of the framing. For example, four different phases of nonoverlapping 2-s time frames are possible over responses sampled at 2 Hz, and a Coordination Score value can be calculated for each. When these are averaged, the resulting Coordination Score is more reliable as it blends away some of the consequences of segmentation. The Coordination Scores reported here are the averages of the Coordination Scores across these multiple frame alignments over the collections, using the Activity Analysis Toolbox functions *coordScoreSimple.m* and *coordScoreRelated.m*. Given that these values are not independent, many methods of combining *p*-values are not applicable. That these Coordination Scores successfully differentiate stimulus-related and -unrelated collections of continuous ratings suggests the calculation works well enough.

We use the same logarithmic transformation to report the results of the nonparametric coordination test, NPC Scores. As this uses overlapping time frames, there is no advantage to applying it to different offsets. Within some range of practicality, confidence in these numbers can be increased with the number of random iterations used to define the distribution of alternative uncoordinated activity. The iteration count also determines the range of the NPC Scores: the 2000 iterations used for most of this paper only yield maximum values of 3.31.

NonParametric Coordination Test

The tests described thus far are specifically for activity events that are sufficiently rare and flexible to have a simple sequential structure, specifically a point process for which a binomial model is reasonable. There are many kinds of activity that are more complicated and responses for which the activity of interest cannot be stripped of sustained or recurrent patterns without interfering with the measurement of synchrony. If we want to consider the timing of inhalations in an audience, we know that any given listener must breathe again and again, but only so often. Such temporal limits or quasi-regularities require a different means of assessing the significance of coordination, with null hypotheses that respect these characteristics. Rather than attempting to model qualities only partially understood, a convenient way to do this is to use the structure of the responses themselves.

We are primarily interested in the coincidence of these events across responses in relation to the timeline of the common stimulus. If the activity is coordinated by the music, breaking the temporal alignment between responses should result in lower activity levels. Using the simple idea of randomly shifting the entire series of each response by a value sampled uniformly from some range of shuffling times (say 0–30 s), it is possible to generate alternative activity-level time series and activity-level distributions. Repeating this process many times simulates a distribution of activity-level distributions as well as distributions of activity levels for each time frame (Grün, 2009).

To evaluate the coordination in the experimental data, we can compare their distribution of activity levels to those of the shuffled alternatives. If the collection's original Coordination Score is more extreme than 95% of the alternatives generated, we have estimated $p < .05$ via this nonparametric test of event coordination. In this case, the comparison was made using the Euclidian distance between each cumulative distribution of activity levels (shuffled alternative

and experimental) to that of the average cumulative distribution of the shuffled alternatives. The rank position of the sample collection against the alternatives is converted to the NPC score as described above.

This analysis of activity is of course sensitive to a few parameters defining these calculations: the qualities of the activity event under investigation, the window of synchrony, and the range of time over which the responses are shuffled to generate the nonparametric alternative distributions. The issue around activity events are much the same as for the parametric coordination tests: better a relevant definition of an intermittent behavior than something so rare that it happens no more than once per listening, or so common that it is a near-constant state for most responses. The shuffling range is more particular and should be chosen with consideration of the temporal structure of the activity being assessed. If the events are periodic to some degree (say heart beats), the shuffling window should be larger than at least one average period. But if the activity changes character over time, such as skin conductance that loses sensitivity over the course of minutes, the time interval should not be so large that this trend is erased. Appropriate shuffling ranges for rating-change activity are evaluated systematically in the following section, and we recommend 15 to 45 seconds for the types of collections considered here.

A last issue to consider for assessment using these alternative alignments of a collection of responses is how to deal with the ends of the series when shuffling. By displacing each series by some interval of up to several seconds, the beginning and end of each collection no longer have a full collection of responses as some are moved away from these end points. We have considered two solutions for this, which should be applied carefully depending on the data. In one case, if response behavior does not change overly much over time, it is safe to loop each response time series, filling in the gap made at the beginning or end with the section of the series cut off at the other. If the characteristics of the series are more evidently changing, another option is

reflection, filling the gap with the series running backwards from the shifted end point. This is particularly useful for periodic events with changing recurrence rates. If neither of these methods can be applied without producing artifacts, the option remains to exclude the beginning and end of the responses by the half-length of the shuffling interval and accept the loss. For these analyses in this paper, we have chosen to loop the ends.

Local Activity Test

Once a collection has been identified as having a high level of coordination in activity across the responses, we can investigate which moments make the activity-level distribution distinct from the alternative by testing which have much higher or lower activity levels than would be expected. Such a test requires a distribution of activity levels for the null hypothesis of unrelated events. With the nonparametric coordination test described previously, we have calculated many alternative activity levels for each time frame through the repeated random shifts, breaking alignment between responses and the stimulus. Our shuffled alternative distribution reports the possible activity levels of these responses, were they not synchronized by the music, while retaining all of characteristics of these binary time series including serial recurrences and changes in activity rate over time.

Each moment is then judged by the rank of the actual activity level against the distribution of alternative activity levels, a nonparametric *p*-value for the experimental coordination at that moment. In this assessment, we are testing many moments and so we must assume that some exceed threshold through coincidences. As previously stated, the existence of a locally extreme high or low activity level over the course of a three-minute piece is not, by itself, a sign of significant coordination. In Appendix 3, the parameter search for a reasonable shuffling range reports how unrelated-response collections average around 5% extreme high activity levels and 2% extreme low activity-levels. For that reason, it is important to first evaluate the coordination of activity over the full activity-level time series of the stimulus. If a collection of responses does not have

significant coordination, most, if not all of these extreme activity levels are likely to be the result of spurious coincidences.

As was the case for the activity-level distributions over the full stimulus, depending on the rate of activity, time frames may display notable coincidences through high activity or low activity. Figure 5B, discussed in the text, marks both the moments when responses are quite active and those in which they are notably still in these ratings of perceived emotional arousal in the orchestral music excerpt "Morning" from Grieg's *Peer Gynt* suite, from the Korhonen (2004) data set. Extreme high and low activity moments are also marked in the rating increases for two collections' responses to different interpretations of the Arcadelt madrigal in Fig 8.

Appendix 3: Evaluating Activity Analysis Parameters

Specific parameters of Activity Analysis and the coordination tests must be defined in ways relevant to the experimental data and hypotheses being considered. What might count as simultaneous changes in ratings of emotion, blurred by attention and introspection, may be too loose for synchronous taping or other reactions to the music. We also need to define the activity event itself, e.g., how large a change in rating value gets counted as an increase in ratings? Depending on the sensitivity of the device collecting responses, there may be small changes from vibrations of the hand or noise generated by the sensor itself. Other rating changes might be part of expressive gestures by the participant, but only as the tail end of a larger reaction to a specific event in the music. These parameters or definitions of activity must not only encompass plausible alignment with the stimulus but also allow the reactions in responses to be differentiated from irrelevant coincidences. No definition of an activity event is guaranteed to capture all and only relevant responses, but we can use the empirical data at hand to identify values of parameters that are measurably effective by some criteria.

Rather than presume to know which collections of ratings contain examples of coordinated rating changes, we can learn the reciprocal: identify parameter values that allow us to reliably reject examples of uncoordinated activity. This perspective is the basis of frequentist tests: defining thresholds to reject the null hypothesis of randomness, according to some articulated false positive rate.

In this section we use continuous ratings to music from many experiments (described in Appendix 1) to establish a reasonable combination of Activity Analysis parameters so that the tests and Coordination Scores are most easily interpreted. To reiterate, the parameters of minimum rating change (defining active events such as increases and decreases in ratings) and the window of synchrony are being chosen to satisfy false positive rates of $\boldsymbol{\alpha}_{crit} = .01$ for

- continuous ratings in one dimension;
- to Western classical or concert music;
- by participants familiar with Western music (involved in experiments run in Canada, France, and the USA);
- in collections of 15 to 40 ratings of duration 120s to 400 s sampled at 1-10 Hz.

We evaluate the optimal combination of parameters by applying the coherence measures to collections we know are not coordinated by construction. To determine values that exceed critical **α** in within-collection coherence measures, we use the 2000 unrelated response collections described in Appendix 1. These are composed of continuous ratings to music, preserving all the characteristics of theses types of stimuli and tasks, combined to form collections without a single common stimulus influencing the timing of activity events of responses.

Maintaining the threshold values of 2 for Coordination Scores, Figure A.1 reports the proportion of uncoordinated collections or collection pairs that exceed threshold per combination

of activity parameters. Each plot reports the results for a different test or event, and the trends of change over different windows of synchrony (ranging from 1 s to 5 s) and minimum rating scale change for increases or decreases in ratings (roughly logarithmic from .3% to 20%).

[insert Figure A.1]

The top two plots report the false coordination rates for the parametric mono-activity tests for increases (Fig. A.1A) and decreases (Fig. A.1B). Windows of synchrony that are too short (1 s) and too long (5 s) both yield excess false positives. At the short end, the activity-levels would generally be very low, and more coincidences across two or three responses would influence the calculation. At the long end, the coordination test would suffer from having too few time frames with little variation in activity levels. Note that the functions used for the parametric calculations automatically reduce the number of bins if it is not possible to split the expected distribution of activity levels according to the reported binning criteria. The rarity of rating changes greater than 20% challenges these parametric tests in a similar manner.

Selecting parameters through this kind of search allows us to find values that suite the significance tests in their capacity to reject false positives, without any reference to true positives or false negatives. The results of these coherence measures do not determine if a musical stimulus influenced the listeners or their responses. Rather they only report whether the coincidences in activity events across responses exceed $\boldsymbol{\alpha}_{\text{crit}}$ against a well-defined alternative of accidental alignment. A rating of liking to a Brahms piano concerto may sometimes increase at the same time as a rating of emotional valence to Grieg's *In the Hall of the Mountain King*, but such changes would necessarily be coincidences. Additionally if a collection of emotional intensity ratings to the first movement of Mozart's First Symphony showed no more coincidental activity than most of these collections of unrelated responses, this is a good reason to acknowledge that the little rondo might not have had a significant coordinating effect on the emotional experience of these listeners.

For the between-collection measures, experimental collections that are not related by stimulus serve to provide a distribution of statistical values without cause for significant coherence. If the correlation between the averages of two audiences responses to a performance of Mozart's Clarinet Concerto in A Minor is no better than most of those between responses to two different pieces, we should be hesitant to read too much into the concurrent swoops of these collections' summaries. Across the 40 experiment collections, 748 pairs are not related by stimulus or rating scale. To compare these collections, the longer of each pair was truncated to match the shorter, and one was resampled to match the other before applying a between-collection coherence measure. The Nonparametric Coordination Test (Fig. A.1C) is slower to focus than the parametric scores. A 2-s synchrony window and minimum rating change of 2.5% yields a false positive rate of 4.6%, rather more than the preferred 1%. In this case, a smaller threshold or larger window of synchrony would come closer to the target rate. Here we compromise for the combination of tests to support a standard set of parameters for rating change active events.

In the false coordination rates of between-collection coordination (Fig. A.1D), the window of synchrony is the most important parameter. Unless the window is large enough to allow a range of joint activity levels, this test defaults to excessive false coordination rates. For this test, the parameter combination of a 2-s window of synchrony and a 2.5% activity threshold produces a slightly higher rate of false positives than the target. However, given the combination of tests and other criteria, this pair of parameters appears relatively effective.

Nonparametric Coordination Parameters

From the Coordination Scores, we have reason to use 2 s as a window of synchrony and 2.5% rating change minimum over 2 s as criterion for defining an activity event. The nonparametric coordination test and the local activity test involve one more parameter. The shuffling window is evaluated here by tracking the proportion of time frames marked with exceptionally high or low

activity for different shuffling ranges, both on experimental collections and a random subset of unrelated response collections. A time frame is counted as having exceptionally high activity if the experimental activity level for that frame is greater than 97.5% of the activity levels from the alternative shufflings on that frame. Exceptionally low activity time frames are correspondingly below 97.5% of those from the alternative alignments of responses. As the shuffling range increases (sampled logarithmically), these ratios reach a plateau, maintaining stable average rates of 10% extreme low activity levels and 15% extreme high activity-levels for the experiment collections, each approximately 8% higher than the rates for unrelated-response collections. For this type of activity event, increases and decreases of at least 2.5% of the rating scales in 2-s time frames, counted in 2-s windows of synchrony, a shuffling range of at least 30 s is suitable.

[insert Figure A.2]

Appendix 4: Collection Dimensions and Coherence Measures

Collections of continuous ratings vary in size, specifically the duration of responses collected and the number of responses gathered to the common stimulus. As suggested in the discussion of Within-Collection Coherence measures, the dimensions of a collection of ratings can have impacts of the value of a measure, raising or lower measure value for reasons mostly independent of the questions of coherence. This can complicate efforts to compare the coherences of response collections. Collection dimensions relate to the amount of information on which an assessment of coherence can be made, the sample-defined degrees of freedom by which to interpret these measures as frequentist tests of coherence. We cannot offer analytic derivations of false-positive thresholds given the number of ratings in a collection or their duration in seconds, but we can demonstrate the direction of these dependences with some careful manipulation of the unrelated response collections. Here we explain the evaluation of collection sizes on the within-collection coherence measures described in the Activity Analysis in Context section of the main text.

To assess the impact on the number of responses in a collection, we used 200 unrelated-response collections of 36 ratings each, and applied the within-collection coherence measures on subsets of nine different sizes from 12 to 36 ratings. We then considered the distribution of measure values generated on collections of each size, testing whether and how the variance and means changed with increased sizes. Differences in variance were evaluated using the non-parametric Brown-Forsythe test, and differences in means were tested with a single factor ANOVA. The outcomes for the measures discussed in the Analysis in Context section are listed in columns 2 and 3 of Table A.6.

[insert Table A.6]

The number of continuous ratings in a collection has impacts on these measures of coherence in different ways. As shown in Fig. A.3A, the distribution of Cronbach's α changes substantially: the variance decreases as the number of responses increases, while the mean value of these incoherent collections rises significantly. This has the effect of squeezing the range of coherent values even further. The MeanCorr measure of coherence does the opposite: these unrelated-response collections concentrate in lower measure values with more responses to evaluate (see Fig A.3C). As expected, the C Scores are not affected by the number of responses in a collection.

[insert Figure A.3]

To evaluate at the impact of response duration on within-collection coherence measures, we generated new unrelated-response collections from the longer responses in the experimental data sets and applied the measures to these, truncated to various durations. Sixteen of the available experiment collections were of ratings longer than 360s, containing a total of 573 individual unidimensional ratings. Of these 200 collections were randomly generated by the same process as

described in Appendix 1. These collections are necessarily less diverse than those sampled from more experiment collections, increasing the likelihood of coherent rating changes. While they may not be composed of completed unrelated responses, their distributions in relation to response duration are still informative. Columns 4 and 5 of Table A.6 report the direction of influence of longer ratings on these measures of within collection coherence. Again, the rating change activity Coordination Scores are not affected by this parameter (see Fig A.3F). Cronbach's **α**'s variance on these collections does not change significantly from 120s to 360s, but the average shifts significantly downwards (see Fig A.3B), and the MeanCorr measure behaves much the same way as it did to increases in the number of responses (see Fig A.3D).

Lastly, we should consider sample rate. Above a minimum frequency, sample rate does not change the amount of stimulus-related information in continuous ratings, but it can still affect these measures. Unfortunately, our collections do not vary sufficiently widely to evaluate this numerically. However, we expect that within a certain range, the coordination scores would decrease in variance with an increase in sample rate, as they have more opportunity to be smoothed with a greater number of samples per window of synchrony.

Author Note

Our sincere thanks to David Temperley and the anonymous reviewers for their considerate (and considerable) suggestions for elaboration and clarification. Thanks also to the many present and past members of McGill University's Music Perception and Cognition Lab who supported the development of Activity Analysis with challenging questions and novel applications over the course of several years. Our appreciation is also due to Mary Farbood for sharing liking ratings from her experiments, to Mark Korhonen for making his data publicly available, to Dan Levitin and Teresa Marrin Nakra for organizing the data collection for the Boston Symphony Orchestra project, and Bennett K. Smith and Julien Boissinot for help with the technical setup for the CIRMMT Audience Response System. This research was made possible by funding to SMc from the Canadian Social Sciences and Humanities Research Council (410-2009-2201), the Canadian Natural Sciences and Engineering Research Council (RGPIN 312774-2010), and the Canada Research Chairs program (950-223484). Much of this work was completed in fulfillment of the requirements of a Master of Arts thesis at McGill University (Upham, 2011). Correspondence concerning this article should be addressed to Finn Upham [finn@nyu.edu].

Tables

Table 1. Performance of within-collection coherence measures on unrelated-response collections and experiment collections.

Within-collection Coherence Measures	95th% on the Unrelated-response Collections	Experiment Collections over 95th % Threshold	99th% on the Unrelated-response Collections	Experiment Collections over 99th % Threshold
Cronbach's α	0.82	25	0.85	20
InterCorr	0.35	14	0.39	12
MeanCorr (r)	0.42	26	0.47	17
VarRatio	0.17	23	0.21	16
C Score, inc	1.33	33	1.9	28
C Score, dec	1.24	30	1.9	25
NPC Score, inc	1.92	36	2.8	28

Note: In columns two and four we report the threshold values that would yield 5% and 1% false-positive rates on the unrelated-response collections, i.e., 2000 randomly assembled collections of real continuous ratings that do not share a common musical stimulus. Columns three and five report the number of experiment collections, out of 40, that exceed these false-positive rate thresholds, our numerical approximation of $p < \boldsymbol{\alpha}_{crit} = .05$ and $p < \boldsymbol{\alpha}_{crit} = .01$.

Table 2. Direct comparisons between within-collection coherence measures by their evaluation of the experiment collections.

Coherence Measures	Corr with Cronbach's α	Agree over 99th %	Corr with MeanCorr	Agree over 99th %	Corr with C score, Inc	Agree over 99th %
Cronbach's α	1	20 (100%)	.89 ***	16 (80%)	.62 ***	17 (85%)
InterCorr	.4 *	7 (58%)	.53 ***	8 (67%)	.42 **	9 (75%)
MeanCorr (r)	.89 ***	16 (94%)	1	17 (100%)	.61 ***	17 (100%)
VarRatio	.95 ***	15 (93%)	.95 ***	16 (100%)	.57 ***	16 (100%)
C Score, inc	.62 ***	17 (61%)	.61 ***	17 (61%)	1	28 (100%)
C Score, dec	.74 ***	18 (72%)	.63 ***	15 (60%)	.58 ***	19 (76%)
NPC Score, inc	.41 **	17 (61%)	.43 **	16 (57%)	.66 ***	25 (89%)

Note: * p<.05, ** p<.01, *** p<.001. For three measures, Cronbach's α, MeanCorr, and the Activity Analysis Coordination Score on rating increases, we list how they correlate (Spearman's ρ) with the other studied measures (df = 38). For each pairing of measures, we also report the number of the 40 collections that exceed both 99th percentile value sets for each measure on the unrelated-response collections, along with the percentage of the second measure's supra-threshold collections are covered in this overlap.

Table 3: Between-collection coherence measures on combinations of experiment collections.

Between-collection Coherence Statistics	95th% on Unrelated Pairs of Collections	Stimulus-related Collection Pairs over 95th % Threshold	99th% on Unrelated Pairs of Collections	Stimulus-related Collection Pairs over 99th % Threshold
Bi-C Score, inc	1.13	10 (3)	1.92	7 (3)
Bi-C Score, dec	1.19	13 (1)	2.3	7 (1)
Bi-C Score, all changes	1.4	19 (0)	2.9	13 (0)
Pearson on full means	0.68	4 (3)	0.83	3 (3)
Pearson, means less 12 s	0.57	4 (2)	0.83	2 (2)
Pearson, differenced means	0.25	5 (2)	0.36	3 (2)
Spearman on means	0.55	4 (0)	0.73	3 (0)
Spearman, means less 12s	0.54	4 (2)	0.78	1 (0)
Spearman, differenced means	0.17	4 (2)	0.25	4 (2)

Note: Columns two and four report the 95th and 99th percentile values of each measure on the 748 possible pairings of the 40 experiment collections unrelated by musical stimulus. Columns three and five report number of stimulus-related experiment collection pairs, out of 27, found to exceed these thresholds per statistic, and thus suggested to be coherent. (Of these 27 pairs, 23 share the musical stimulus but differ in rating scale, and none are guaranteed to be coherent.) The number of stimulus and rating scale related pairs (4) exceeding threshold is reported in brackets. The measures compared are variations on Activity Analysis Coordination Scores between collections (Bi-C Score), and variations on Pearson and Spearman correlations between the average rating time series (means).

Table 4: Rating change coordination scores for the within-collection coordination on increases (inc) and decreases (dec) in the perceived emotional arousal and valence ratings to each stimulus of the Korhonen (2004) data set.

Composer	Stimulus	Duration (s)	Dimension	C score (inc)	C score (dec)
Liszt	Allegro	315	Arousal	16***	16***
			Valence	13.8***	9.8***
Rodrigo	Aranjuez	164	Arousal	14***	3.0***
			Valence	1.9*	0.9
Copland	Fanfare	169	Arousal	2.5**	3.7***
			Valence	5.6***	7.4***
Beethoven	Moonlight	152	Arousal	2.2**	4.6***
			Valence	1.0	1.9*
Grieg	Morning	163	Arousal	15.6***	16***
			Valence	13.6***	6.9***
Strauss	Pizzicato	150	Arousal	9.9***	7.2***
			Valence	10.5***	3.5***

Note: * p<.05, ** p<.01, *** p<.001. More stimulus details shared in Appendix 1, Table A3.

Table 5: Coordination Scores for rating-change activity in felt emotional intensity reported continuously by participants in audiences attending a live performance of the Boston Symphony Orchestra (Live) or watching a reproduction in a concert hall (Recorded).

Stimulus	Audience	C score (inc)	C score (dec)	Bi-C score (inc)	Bi-C score (dec)
Figaro Overture, K492	Live	3.4***	4.0***	5.5***	4.9***
	Recorded	3.7***	1.6*		
First Symphony, K16	Live	2.3**	0.2	0.1	0.4
	Recorded	0.4	0.1		
Clarinet Concerto, K622	Live	2.2**	0.3	2.1**	0.3
	Recorded	2.0**	1.1		
Jupiter Finale, K551	Live	3.9***	5.4***	2.1**	0.9
	Recorded	1.5*	2.0**		

Note: * $p<.05$, ** $p<.01$, *** $p<.001$. The within-collection rating change activity per concert condition (C Score) is reported along side the between-collection rating-change activity coordination (Bi-C Score) between these two groups. The activity events assessed were changes of at least .025 of the rating scale in 2-s windows of synchrony in the direction of increases (inc) and decreases (dec). The music performed were four excerpts of Mozart's Symphonic repertoire, more details in Table A.2.

Table A.1. The continuous rating collections in the data set from *The Angel of Death* project.

Piece **Performer** **Composer**	**Rating dimension** **Context** **Rating device**	**No. participants** **Sampling rate** **Duration (s)**
The Angel of Death, D-S version	Emotional force	41
G. Cheng, SONOR Ensemble	Live concert	2 Hz
R. Reynolds	1D slider	2054.5
The Angel of Death, S-D version	Emotional force	51
G. Cheng, SONOR Ensemble	Live concert	2 Hz
R. Reynolds	1D slider	2033.5
The Angel of Death, D-S version	Force émotionnelle	54
JM Cottet, Court Circuit	Live concert	2 Hz
R. Reynolds	1D slider	2067.5
The Angel of Death, S-D version	Force émotionnelle	41
JM Cottet, Court Circuit	Live concert	2 Hz
R. Reynolds	1D slider	2067.5
The Angel of Death, D-S version	Resemblance	34
G. Cheng, SONOR Ensemble	Live concert	2 Hz
R. Reynolds	1D slider	2054.5
The Angel of Death, S-D version	Resemblance	43
G. Cheng, SONOR Ensemble	Live concert	2 Hz
R. Reynolds	1D slider	2033.5
The Angel of Death, D-S version	Familiarité	36
JM Cottet, Court Circuit	Live concert	2 Hz
R. Reynolds	1D slider	2067.5
The Angel of Death, S-D version	Familiarité	40
JM Cottet, Court Circuit	Live concert	2 Hz
R. Reynolds	1D sliders	2067.5

Table A.2. The continuous rating collections in the Boston Symphony Orchestra data set.

Piece **Performers** **Composer**	**Rating dimension** **Context** **Rating device**	**No. participants** **Sampling rate** **Duration (s)**
Overture, *The Marriage of Figaro*, K492	Emotional intensity	30
Boston Symphony Orchestra	Live concert	2 Hz
W. A. Mozart	1D slider	239.5
Overture, *The Marriage of Figaro*, K492	Emotional intensity	23
Boston Symphony Orchestra	Recorded concert	2 Hz
W. A. Mozart	1D slider	239.5
Rondo, Symphony No. 1, K16	Emotional intensity	30
Boston Symphony Orchestra	Live concert	2 Hz
W. A. Mozart	1D slider	128
Rondo, Symphony No. 1, K16	Emotional intensity	22
Boston Symphony Orchestra	Recorded concert	2 Hz
W. A. Mozart	1D slider	128
Adagio, Clarinet Concerto in A, K622	Emotional intensity	31
Boston Symphony Orchestra	Live concert	2 Hz
W. A. Mozart	1D slider	401.5
Adagio, Clarinet Concerto in A, K622	Emotional intensity	22
Boston Symphony Orchestra	Recorded concert	2 Hz
W. A. Mozart	1D slider	401.5
Finale, Symphony No. 41 (*Jupiter*), K551	Emotional intensity	31
Boston Symphony Orchestra	Live concert	2 Hz
W. A. Mozart	1D slider	350.5
Finale, Symphony No. 41 (*Jupiter*), K551	Emotional intensity	22
Boston Symphony Orchestra	Recorded concert	2 Hz
W. A. Mozart	1D slider	350.5

Table A.3. The continuous rating collections in the data set collected by Mark Korhonen (2004).

Piece **Performers** **Composer**	**Rating dimension** **Context** **Rating device**	**No. participants** **Sampling rate** **Duration (s)**
Allegro, Piano Concerto No. 1	Emotional arousal	35
J. Banowetz w/ Slovak Radio Symph. Orch.	Recording, alone	1 Hz
F. Liszt	2D emotion space	315
Allegro, Piano Concerto No. 1	Emotional valence	35
J. Banowetz w/ Slovak Radio Symph. Orch.	Recording, alone	1 Hz
F. Liszt	2D emotion space	315
Adagio, *Concierto de Aranjuez*	Emotional arousal	35
Norbert Kraft w/ Northern Chamber Orch.	Recording, alone	1 Hz
J. Rodrigo	2D emotion space	165
Adagio, *Concierto de Aranjuez*	Emotional valence	35
Norbert Kraft w/ Northern Chamber Orch.	Recording, alone	1 Hz
J. Rodrigo	2D emotion space	165
Fanfare for the Common Man	Emotional arousal	35
Slovak Radio Symph. Orch.	Recording, alone	1 Hz
A. Copland	2D emotion space	170
Fanfare for the Common Man	Emotional valence	35
Slovak Radio Symph. Orch.	Recording, alone	1 Hz
A. Copland	2D emotion space	170
Adagio, *Moonlight Sonata*	Emotional arousal	35
J. Jando	Recording, alone	1 Hz
L. van Beethoven	2D emotion space	153
Adagio, *Moonlight Sonata*	Emotional valence	35
J. Jando	Recording, alone	1 Hz
L. van Beethoven	2D emotion space	153
Morning, *Peer Gynt Suite*, No. 1	Emotional arousal	35
BBC Scottish Symph. Orch.	Recording, alone	1 Hz
E. Grieg	2D emotion space	164
Morning, *Peer Gynt Suite*, No. 1	Emotional valence	35
BBC Scottish Symph. Orch.	Recording, alone	1 Hz
E. Grieg	2D emotion space	164
Pizzicato Polka	Emotional arousal	35
Slovak Radio Symph. Orch.	Recording, alone	1 Hz
Johann Strauss II, Josef Strauss	2D emotion space	164
Pizzicato Polka	Emotional valence	35
Slovak Radio Symph. Orch.	Recording, alone	1 Hz
Johann Strauss II, Josef Strauss	2D emotion space	164

Table A.4. The continuous rating collections in the CARS data set, collected in March and October 2009.

Piece **Performers** **Composer**	**Rating dimension** **Context** **Rating device**	**No. participants** **Sampling rate** **Duration (s)**
Il bianco e dolce cigno	Emotional arousal	30
King's Singers	Recorded, 1D or 2D	10 Hz
J. Arcadelt	iPod GUI	119.7
Il bianco e dolce cigno	Emotional valence	30
King's Singers	Recorded, 1D or 2D	10 Hz
J. Arcadelt	iPod GUI	119.7
Andante-Allegro, String Qt No. 3, Op. 48	Emotional valence	30
St. Laurence Quartet	Recorded, 1D or 2D	10 Hz
R. Schumann	iPod GUI	488.6
Andante-Allegro, String Qt No. 3, Op. 48	Emotional valence	30
St. Laurence Quartet	Recorded	10 Hz
R. Schumann	1D or 2D iPod GUI	488.6
Everybody to the Power of One (V. 1)	Emotional arousal	30
d. andrew stewart	Recorded	10 Hz
d. andrew stewart	1D or 2D iPod GUI	390.4
Everybody to the Power of One (V. 1)	Emotional valence	30
d. andrew stewart	Recorded	10 Hz
d. andrew stewart	1D or 2D iPod GUI	390.4
Il bianco e dolce cigno	Emotional arousal	17
Live (Orpheus Singers)	Live concert	10 Hz
J. Arcadelt	2D iPod GUI	128.4
Il bianco e dolce cigno	Emotional valence	17
Live (Orpheus Singers)	Live concert	10 Hz
J. Arcadelt	2D iPod GUI	128.4
Andante-Allegro, String Qt No. 3, Op. 48	Emotional arousal	8
Live (Student Quartet)	Live concert	10 Hz
R. Schumann	2D iPod GUI	130
Andante-Allegro, String Qt No. 3, Op. 48	Emotional valence	8
Live (Student Quartet)	Live concert	10 Hz
R. Schumann	2D iPod GUI	130
Everybody to the Power of One (V. 2)	Emotional arousal	30
d. andrew stewart	Live concert	10 Hz
d. andrew stewart	2D iPod GUI	446.1
E*verybody to the Power of One* (V. 2)	Emotional valence	30
d. andrew stewart	Live concert	10 Hz
d. andrew stewart	2D iPod GUI	446.1

Table A.5. The continuous rating collections to different scrambled versions of a Brahms piano concert excerpt.

Piece **Performers** **Composer**	**Rating dimension** **Context** **Rating device**	**No. participants** **Sampling rate** **Duration (s)**
Concerto Excerpt (Original)	Liking	22
Soloist w/ Orchestra (conductor)	Recording	10 Hz
J. Brahms	1D GUI	256
Concerto Excerpt (Section Scrambled)	Liking	22
Soloist w/ Orchestra (conductor)	Recording	10 Hz
J. Brahms	1D GUI	256

Table A.6. The impact of continuous rating collection size on within-collection coherence measures, both the number of responses and the response duration. Directions of effect are reported in relation to increases in these factors.

Coherence measure	Size Variance	Size Mean	Duration Variance	Duration Mean
Cronbach's α	Decreases***	Increases***	No Trend	Decreases***
InterCorr	Decrease***	No Trend	Decreases***	Decreases***
Mean Corr (r)	Decreases***	Decreases***	Decrease***	Decreases***
Var Ratio	Decreases***	Decreases***	Decrease***	Decreases***
C score, inc	No Trend	No Trend	No Trend	No Trend
C score, dec	No Trend	No Trend	No Trend	No Trend
NPC score, inc	Increases***	Increases***	Increases**	Increases**

Figure captions

FIGURE 1. An example of a continuous rating response and activity events it contains. A) A single listener's rating time series of perceived emotional valence to the Allegro movement of F. Liszt's Piano Concerto No. 1, performed by J. Banowetz with the Slovak Radio Symphony Orchestra, collected as part of the Korhonen data set (See Appendix 1). In the following panels, different kinds of activity events in this single response as point processes are marked. B) Increases (inc) of at least 2.5% of the rating scale in a sequence of 2-s frames. C) Rating decreases (dec) in the same sequence of frames. D) Zero crossings (0cross), moments when the response moves between the regions of positive and negative emotional valence.

FIGURE 2. Summaries of the collection of perceived emotional valence ratings to the Liszt excerpt from the Korhonen data set (see Figure 1). A) The 35 response time series (Rsp) and their average (Avg). B) Four activity-level time series for rating increases in this collection with minimum rating change thresholds of 0.5%, 2.5%, 10%, and 20% of the valence rating scale in overlapping time frames (2-s window of synchrony). C) Four activity-level time series for rating decreases, with the same rating change thresholds and time frames as B. D) The activity-level time series of increases (inc) and decreases (dec) of at least 2.5% in nonoverlapping time frames, with decreases shown below the x-axis.

FIGURE 3. The assessment of activity coordination in the felt emotional intensity ratings by an audience watching and listening to a recording of a concert performance of the Overture to Mozart's *Marriage of Figaro* (K492). A) Individual response time series and the average time series in black. B) Activity-level time series for increases (inc, above zero) and decreases (dec, below zero) in rating changes of at least 2.5%, over 2-s time frames. Activity-level distributions (act lvls) for increases (C) and decreases (D) in rating values along with the parametric model of random independent activity used to assess the activity coherence. The contraction of the distributions of

activity into four bins for the goodness-of-fit (GoF) test, increases (E) and decreases (F), which finds both to be significantly different from random uncoordinated activity.

FIGURE 4. Joint activity of emotional intensity rating increases in two collections of responses to the same performance of Mozart's K492 by the Boston Symphony Orchestra. A) Two activity-level time series, that of the live audience above (Live (inc)), and the audience watching a recording below (Rec (inc)). The large middle graph (C) shows the actual joint distribution of time frames per combination of activity levels for rating increases in each collection, supported by their respective activity-level distributions of increases for Collection 1 (B) to the live performance and Collection 2 (E) to the recording. D) The expected alternative joint distribution of independent activity levels. Dotted lines on both distribution graphs (B and E) frame the bins used for contingency table sums: independent random model (F) and actual joint-distribution (G). The results of the contingency table test are shown at the bottom right, rejecting the null hypothesis of independent rating increases activity between these collections.

FIGURE 5 Evaluation of moments with exceptional coherence in perceived arousal ratings to the Grieg excerpt (Morning Mood) from the Korhonen data set. A) The collection's responses and average (thick black line). B) Moments of locally extreme high activity levels (open circles) and low activity levels (filled circles) on the activity level time series of rating increases (max NPC score, 3.3) for extremes of $p < .025$. C) Moments (stars) selected by the modified Wilcoxon test with 90th percentile selection criteria on the median first-order difference time series (1 Hz) of the collection's ratings, grey circles marking test positives on 80th percentile selection criteria. D) The collection's standard deviation time series with moments highlighted by Schubert's (2007) second-order SD; black dots mark moments identified using the full stimulus duration and grey circles indicate those selected on the same relative threshold when excluding the first 12 s for thresholds of the times series mean plus one standard deviation; black asterisks mark threshold of time series

mean minus one standard deviation, also without the first 12 s.

FIGURE 6. Continuous ratings of perceived emotion to an excerpt of the Adagio movement from *Concierto de Aranjuez* by J. Rodrigo and performed by Norbert Kraft with the Northern Chamber Orchestra. These ratings from the Korhonen data set were collected continuously along two dimensions, emotional valence and arousal, here treated independently. A) Continuous ratings of arousal (Rsp, responses) and their average (Avg). B) Activity levels for arousal rating increases (Inc) and decreases (Dec) of at least 2.5% in non-overlapping 2-s time frames. C) Continuous ratings of valence (Rsp) and their average (Avg). D) Activity levels for valence rating increases (Inc) and decreases (Dec).

FIGURE 7. Continuous ratings of felt emotional intensity by two participant groups, one to a live concert performance of the Finale movement from W. A. Mozart's *Jupiter Symphony* by the Boston Symphony Orchestra (Collection 1) and the other to a reproduction of the same performance in a recital hall with video and stereo audio (Collection 2). A) Ratings by the audience attending the live concert (Rsp) and their average rating time series (Avg). B) Average ratings of both response collections (Coll 1: live, Coll 2: recording). C) Rating-change activity-level time series for the two collections with increases of at least 2.5% above, (Inc 1 from Collection 1, Inc 2 from Collection 2) and the corresponding decreases below the x axis. D) Collection 2's ratings of emotional intensity to the concert reproduction (Rsp) and their average time series (Avg).

FIGURE 8. Continuous ratings of felt emotional arousal ratings to two different interpretations of the Renaissance madrigal *Il blanco e dolce cigno* by J. Arcadelt. A) 30 ratings (Rsp (Rec)) and their average times series (Avg) to a recording by the King's Singers, plotted in metrical time and C) 17 ratings (Rsp (Rec)) and their average (Avg) to a live performance by the semi-professional choir, the Orpheus Singers, plotted in metrical time. B) The activity levels of rating increases (minimum 2.5% in 2-s time frames) on overlapping time frames, aligned in metrical time,

of ratings to the recording above (Rec (inc)) (max NPC Score, 3.3), and to the live performance below (Live (inc)) (max NPC score, 3.3), with time frames of locally extreme high and low activity levels (X-Act) marked in grey circles and black diamonds.

FIGURE 9. Exploring the impact of two-dimensional rating tasks on continuous ratings of felt and perceived emotion, Arousal x Valence, with Coordination Scores. The horizontal line at C score = 2 in each plot marks the $p < .001$ threshold for these scores. From one session of the CARS experiment, a third of participants rated both dimensions of emotion simultaneously (2D) for all three musical stimuli (S1, S2, S3), the other two thirds rating each of the dimensions independently (1D). A) Activity coordination within and between these participant group collections, per dimension, direction of rating change, and stimulus. B) All combinations of rating-change coordination between these dimensions of emotion: any change in Arousal and Valence (A/V change), increases in both Arousal and Valence (A/V inc), decreases in both Arousal and Valence (A/V dec), increases in Arousal against decreases in Valence (Ainc/Vdec), and decreases in Arousal with increases in Valence (Adec/Vinc). These combinations of between-collection activity scores (Bi-C Scores) are of the dimensions of emotion from participants rating both simultaneously (2D ratings) and those between the two groups rating either one alone (1D ratings). C) Between-collection coordination scores or the dimensions of emotion in the same combinations as plot B, for the 2D perceived emotion ratings to the six stimuli from the Korhonen data sets.

FIGURE A.1. Estimates of false positive rates for Activity Analysis coordination tests across combinations of window of synchrony sizes and event thresholds for rating change activity on continuous ratings to music. These heat maps report the proportion of the incoherent collections or collection pairs exceeding the target Coordination score values of 2 ($\alpha = .01$) for each combination of window of synchrony sizes, from 1 to 5 s, and minimum rating-change thresholds, (from .3% to 20% of the rating scale). A) Proportions of the 2000 unrelated-response collections reporting

Coordination Scores > 2 for increases in ratings, given the activity parameters. B) The same for decreases in ratings. C) Proportions of the 2000 unrelated-response collections reporting Nonparametric Coordination Scores > 2 (1000 iterations) for increases in ratings. D) Proportion of the 748 unrelated-collection pairs with Bi-Coordination Scores > 2 for increases in ratings, given these activity parameters.

FIGURE A.2. The effect of the nonparametric coordination test's shuffling range on the proportion of time frames found to have locally extreme high activity levels and low activity levels of rating increases on experiment collections of continuous ratings (40) and a random subset of unrelated-response collections (88). The shuffling range parameter is varied from 2 s to 64 s, sampled at ticks, and reported in log-scale. Each line reports the percentage of time frames with activity levels ranked below 2.5% of their 2000 stimulus-asynchronous alternatives (A) or above 98.5% (B). Each fainter lines make the percentages of individual collections (experiment collections in thin lines, unrelated-response collection in dotted lines), and the average percentages for two types of collections are reported in thicker grey (experiment) and black (unrelated).

FIGURE A.3. The interaction of response collection dimensions and within-collection coherence measures, demonstrated on unrelated-response collections. Each subplot reports the distribution of measure values on 200 randomly generated unrelated-response collections with the size specified on the x axis. The 5th, 50th, and 95th percentile values per collection dimension are traced over the heat maps in black. The top row reports the distributions of Cronbach's α values, below that the MeanCorr values, and at the bottom, Coordination Scores for rating increases. The left column covers the number of responses in a collection, counted over the coherence measure's full value range. The right column shows the distributions per duration of response on a value range limited to the middle 98% of values taken by these unrelated-response collections.

Figure 1

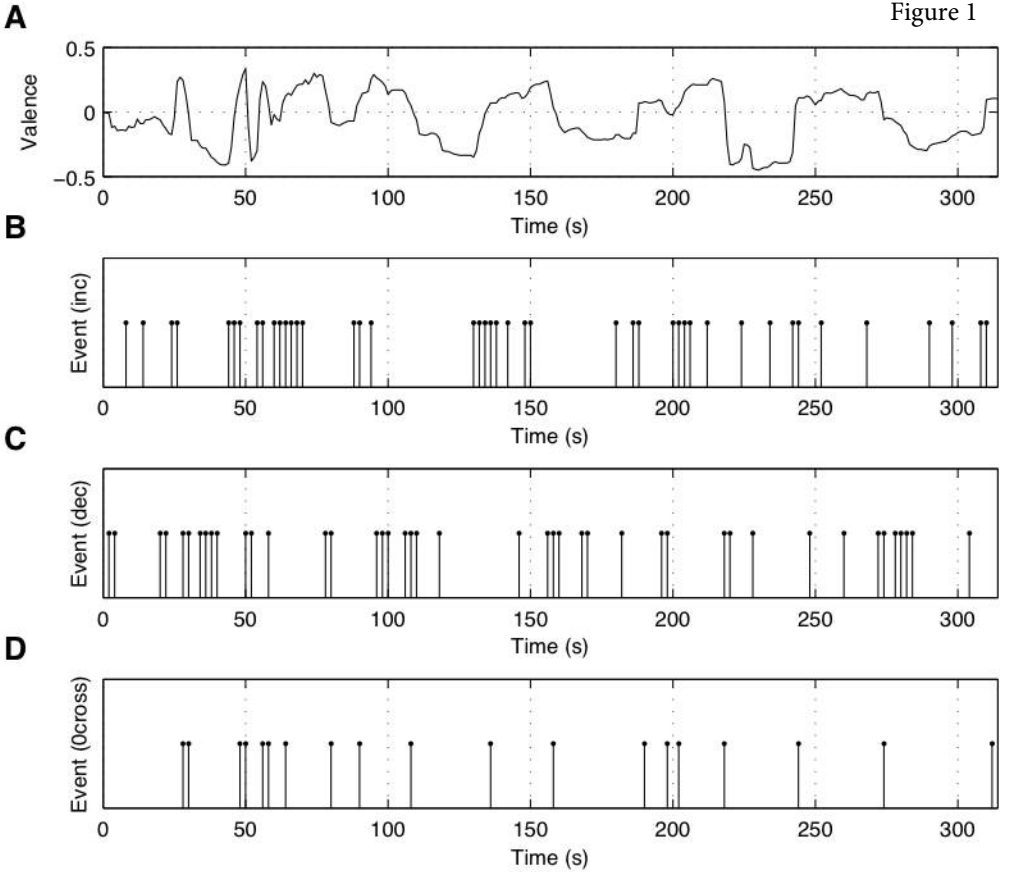

A
Valence
0.5
0
−0.5
0
50
100
150
200
250
300
Time (s)
B
Event (inc)
Time (s)
C
Event (dec)
Time (s)
D
Event (0cross)
Time (s)

Figure 1

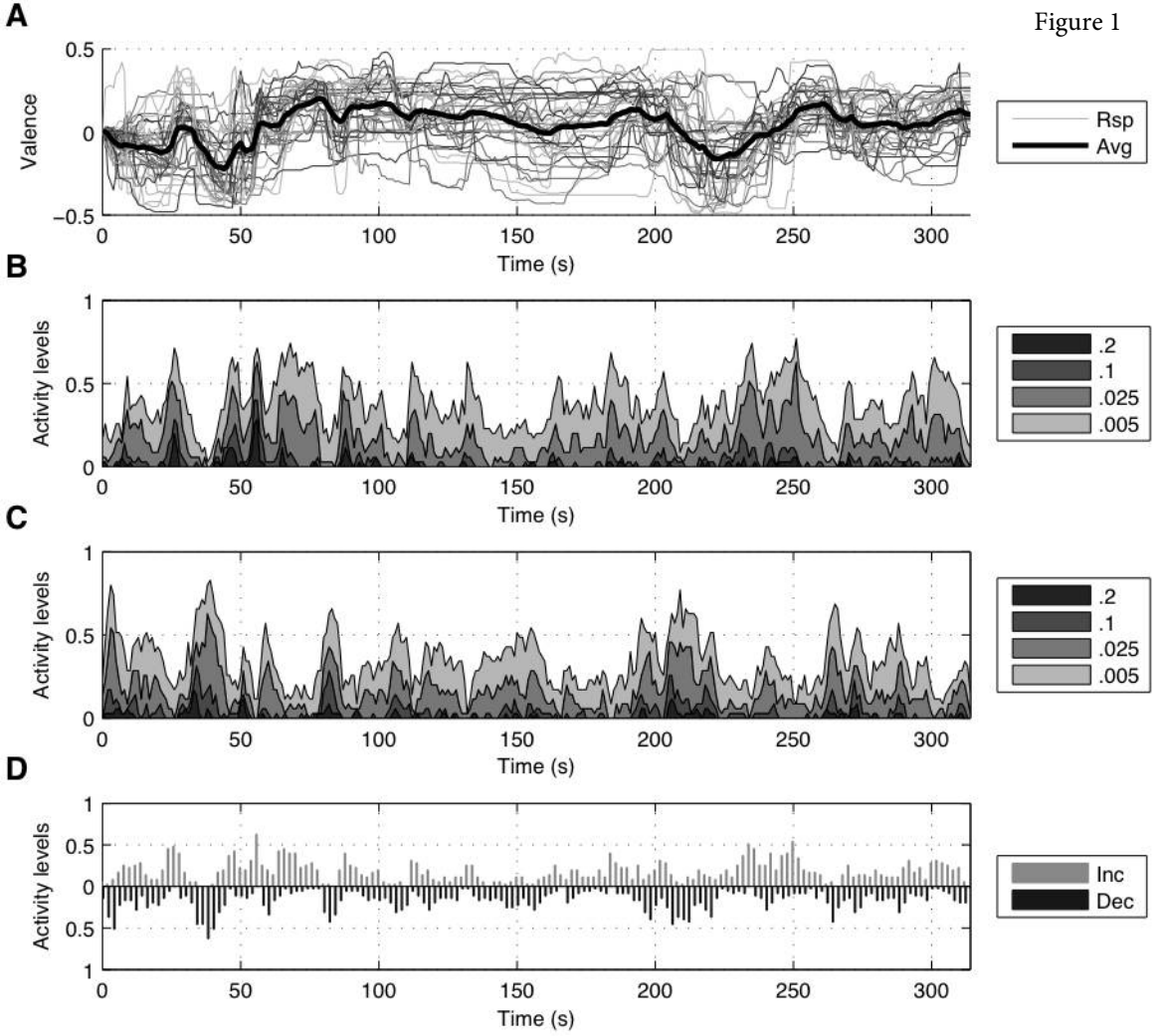

A
Valence
0.5
0
−0.5
0
50
100
150
200
250
300
Time (s)
Rsp
Avg
B
Activity levels
1
0.5
0
Time (s)
.2
.1
.025
.005
C
Activity levels
Time (s)
.2
.1
.025
.005
D
Activity levels
Time (s)
Inc
Dec

Figure 3

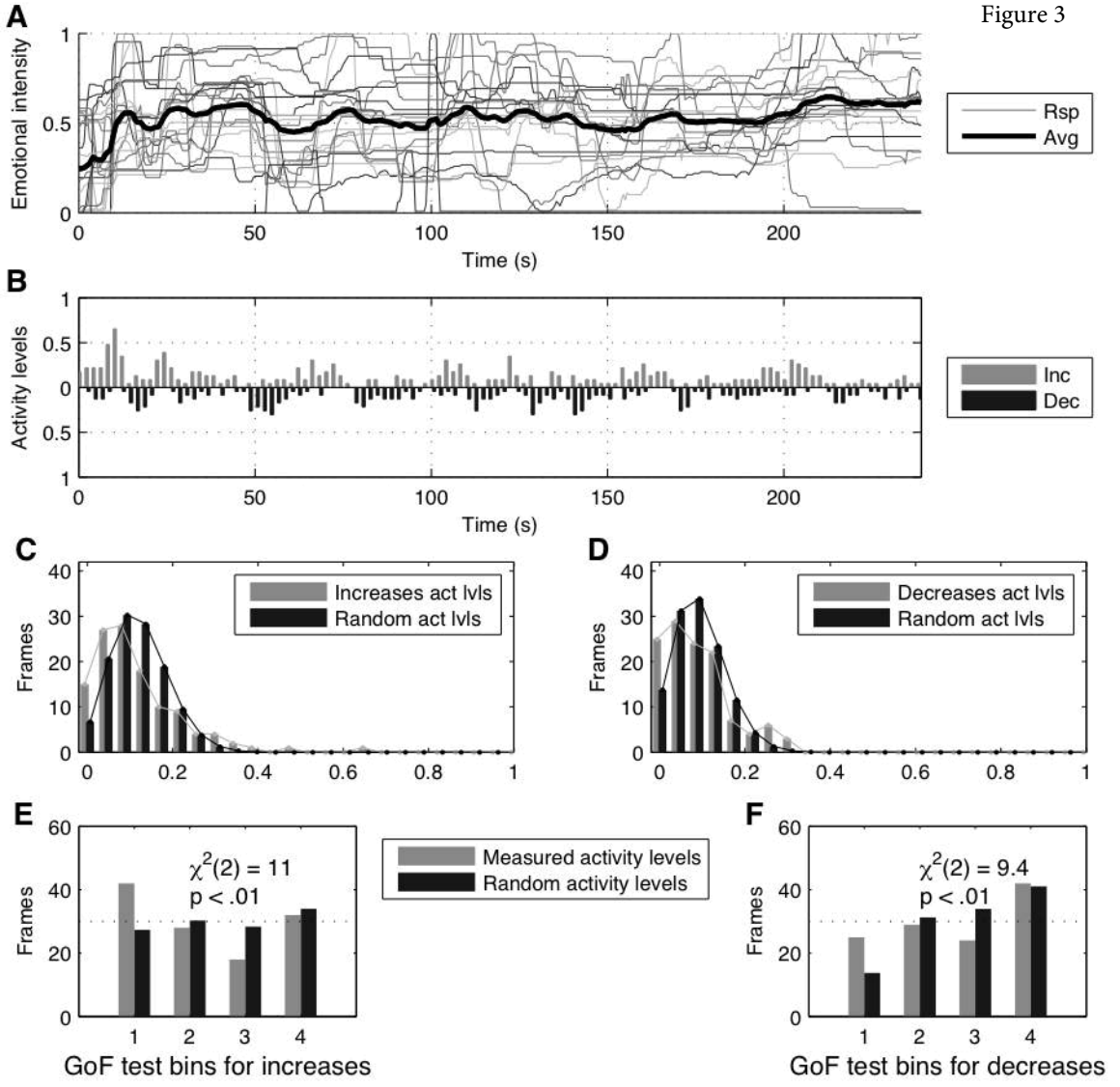

A
Emotional intensity
Time (s)
Rsp
Avg
B
Activity levels
Time (s)
Inc
Dec
C
Frames
Increases act lvls
Random act lvls
D
Frames
Decreases act lvls
Random act lvls
E
Frames
$\chi^2(2) = 11$
$p < .01$
Measured activity levels
Random activity levels
GoF test bins for increases
F
Frames
$\chi^2(2) = 9.4$
$p < .01$
GoF test bins for decreases

Figure 4

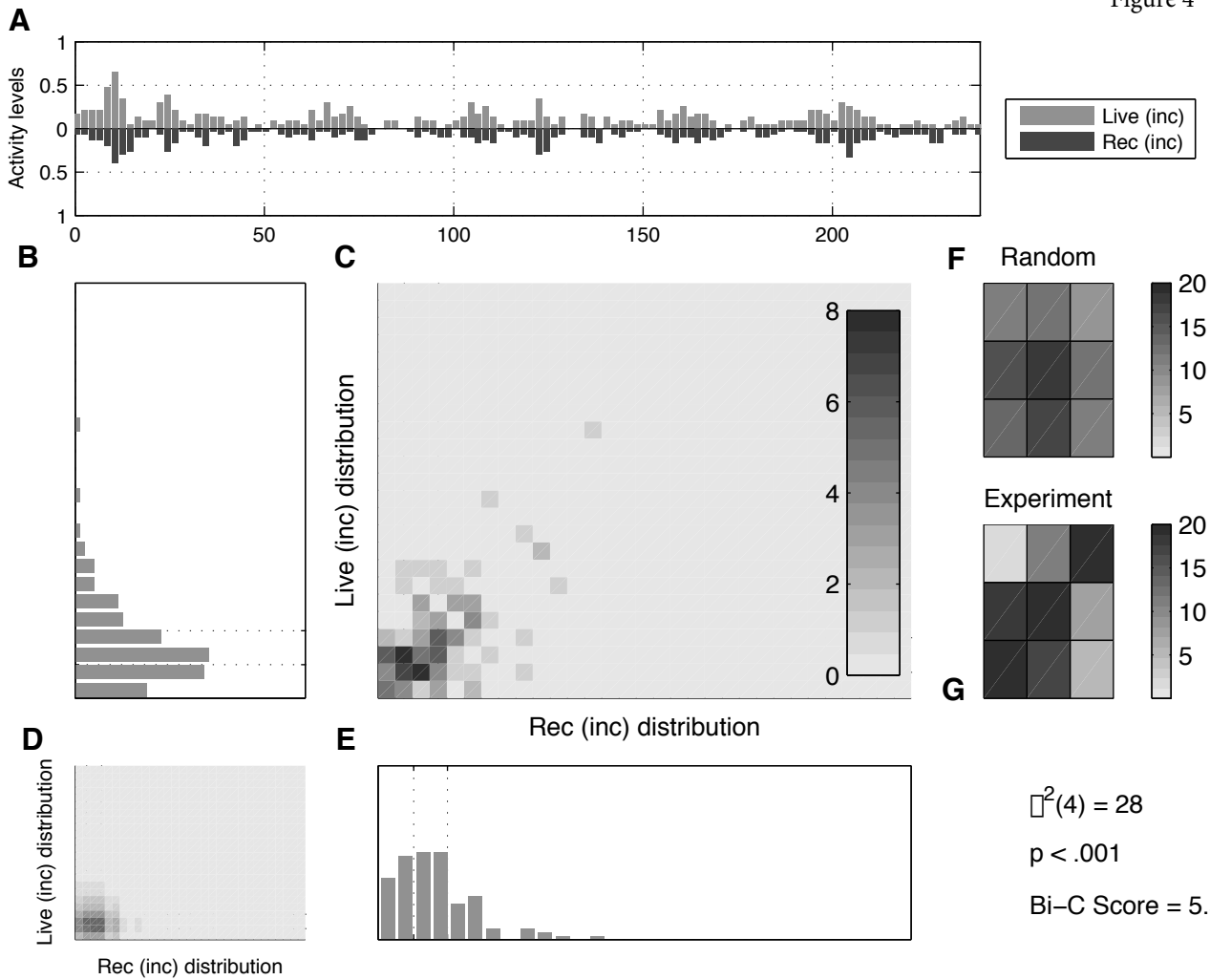
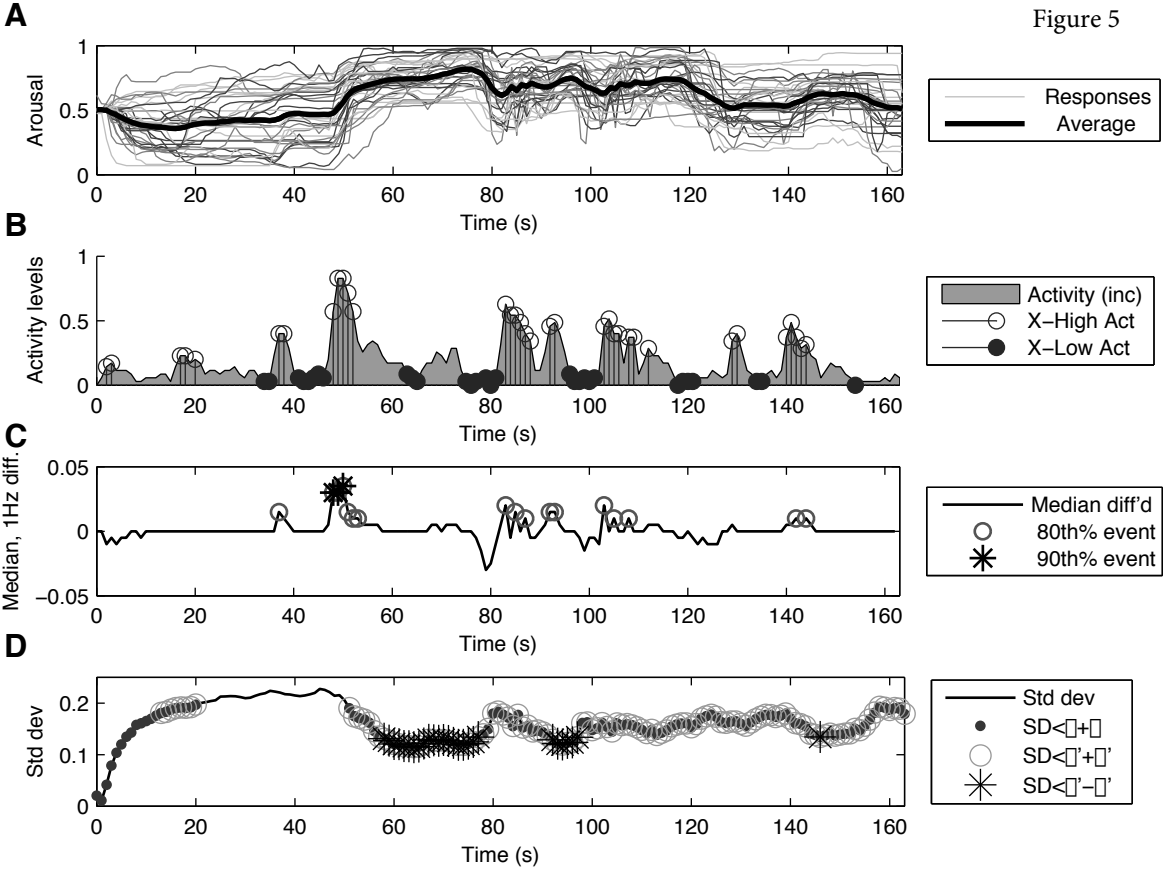
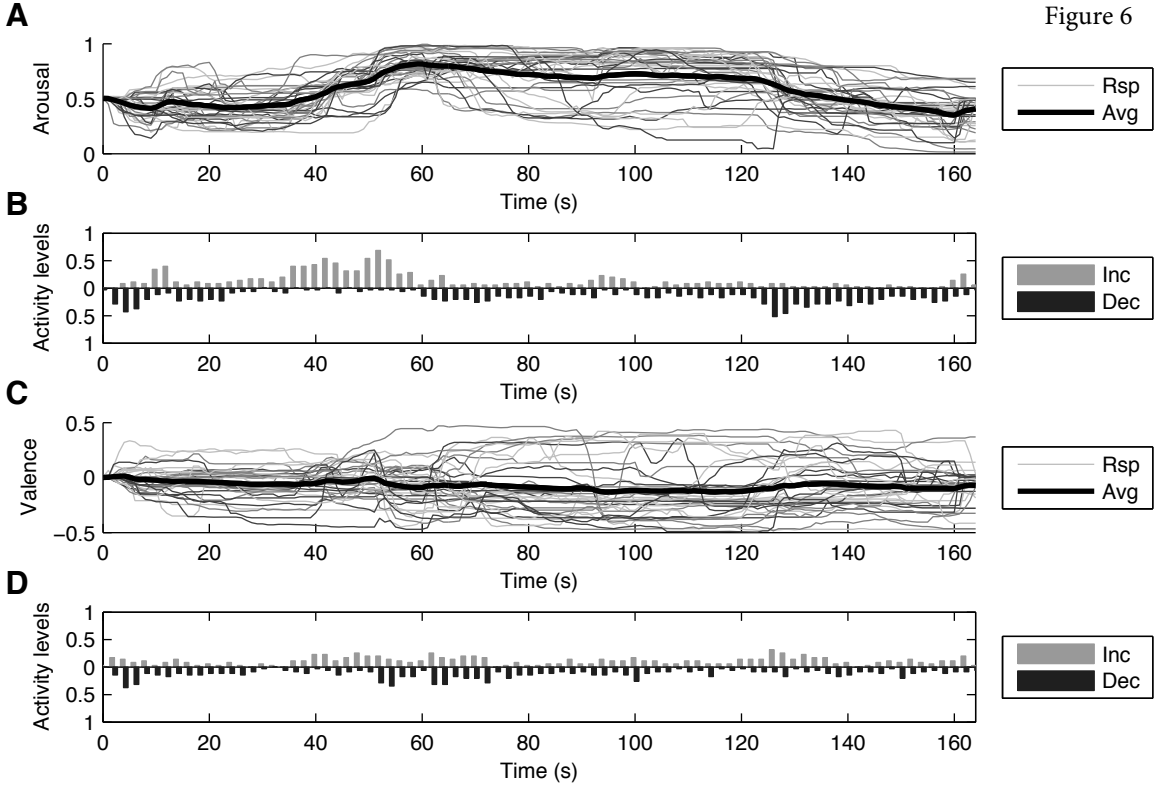

A
Activity levels
1
0.5
0
0.5
1
0
50
100
150
200
Live (inc)
Rec (inc)
B
C
Live (inc) distribution
Rec (inc) distribution
8
6
4
2
0
F
Random
20
15
10
5
Experiment
G
D
Live (inc) distribution
Rec (inc) distribution
E
$\chi^2(4) = 28$
p < .001
Bi–C Score = 5.5

Figure 5


A
Arousal
Time (s)
Responses
Average
B
Activity levels
Time (s)
Activity (inc)
X–High Act
X–Low Act
C
Median, 1Hz diff.
Time (s)
Median diff'd
80th% event
90th% event
D
Std dev
Time (s)
Std dev
SD<□+□
SD<□'+□'
SD<□'–□'


A
Arousal
1
0.5
0
0
20
40
60
80
100
120
140
160
Time (s)
Rsp
Avg
B
Activity levels
1
0.5
0
0.5
1
0
20
40
60
80
100
120
140
160
Time (s)
Inc
Dec
C
Valence
0.5
0
−0.5
0
20
40
60
80
100
120
140
160
Time (s)
Rsp
Avg
D
Activity levels
1
0.5
0
0.5
1
0
20
40
60
80
100
120
140
160
Time (s)
Inc
Dec

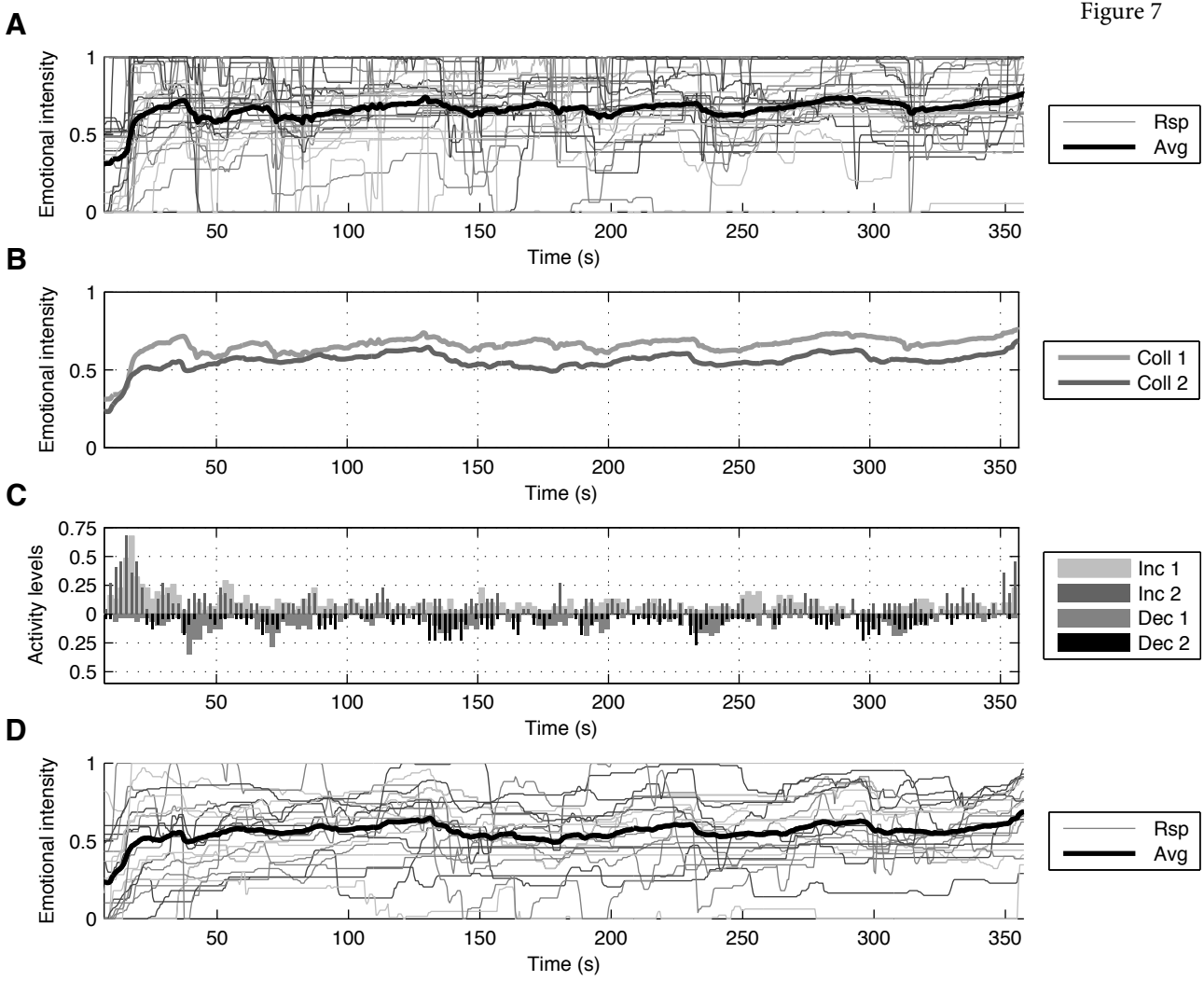

A
Emotional intensity
1
0.5
0
50
100
150
200
250
300
350
Time (s)
Rsp
Avg
B
Emotional intensity
1
0.5
0
50
100
150
200
250
300
350
Time (s)
Coll 1
Coll 2
C
Activity levels
0.75
0.5
0.25
0
0.25
0.5
50
100
150
200
250
300
350
Time (s)
Inc 1
Inc 2
Dec 1
Dec 2
D
Emotional intensity
1
0.5
0
50
100
150
200
250
300
350
Time (s)
Rsp
Avg

Figure 8

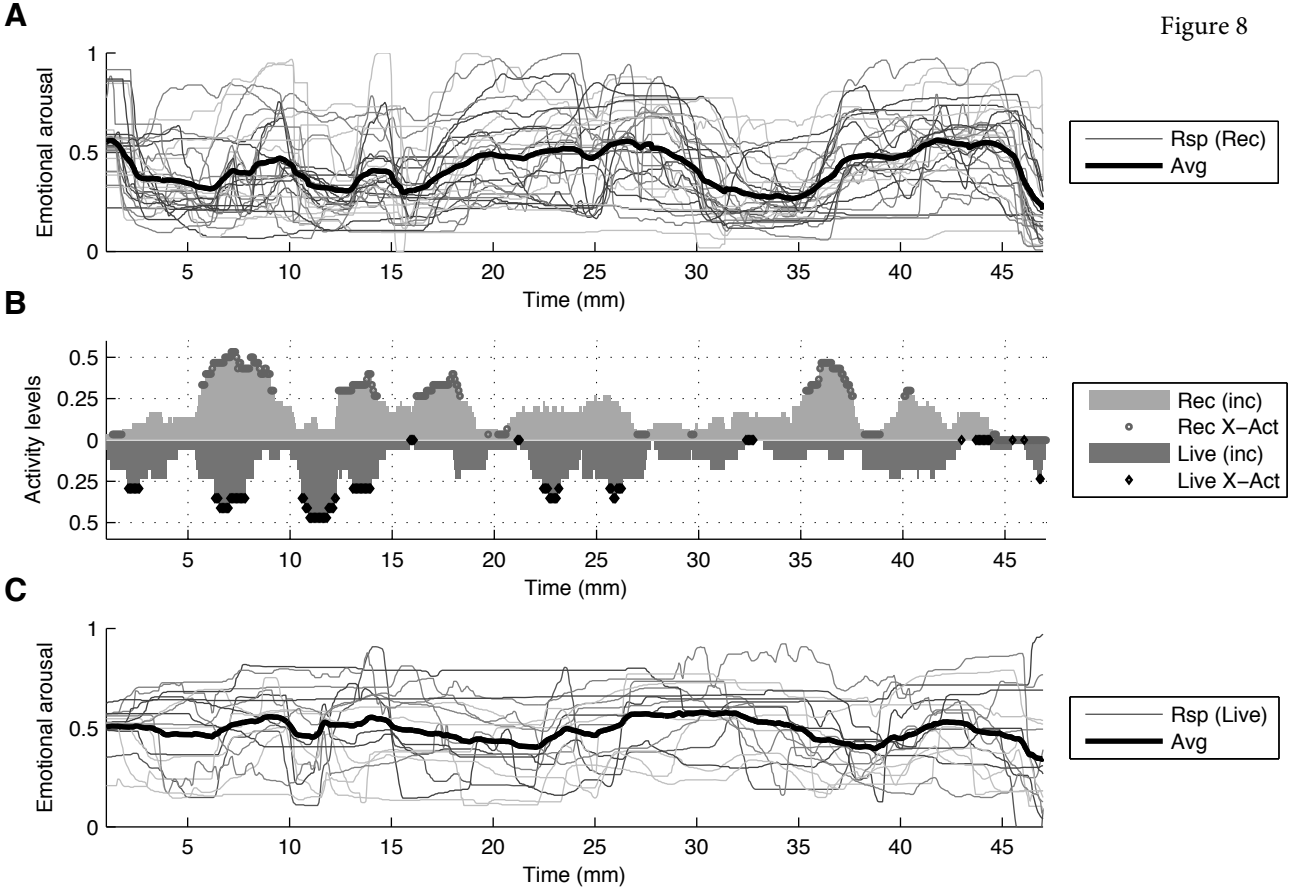

A
Emotional arousal
1
0.5
0
5
10
15
20
25
30
35
40
45
Time (mm)
Rsp (Rec)
Avg
B
Activity levels
0.5
0.25
0
0.25
0.5
5
10
15
20
25
30
35
40
45
Time (mm)
Rec (inc)
Rec X−Act
Live (inc)
Live X−Act
C
Emotional arousal
1
0.5
0
5
10
15
20
25
30
35
40
45
Time (mm)
Rsp (Live)
Avg

Figure 9

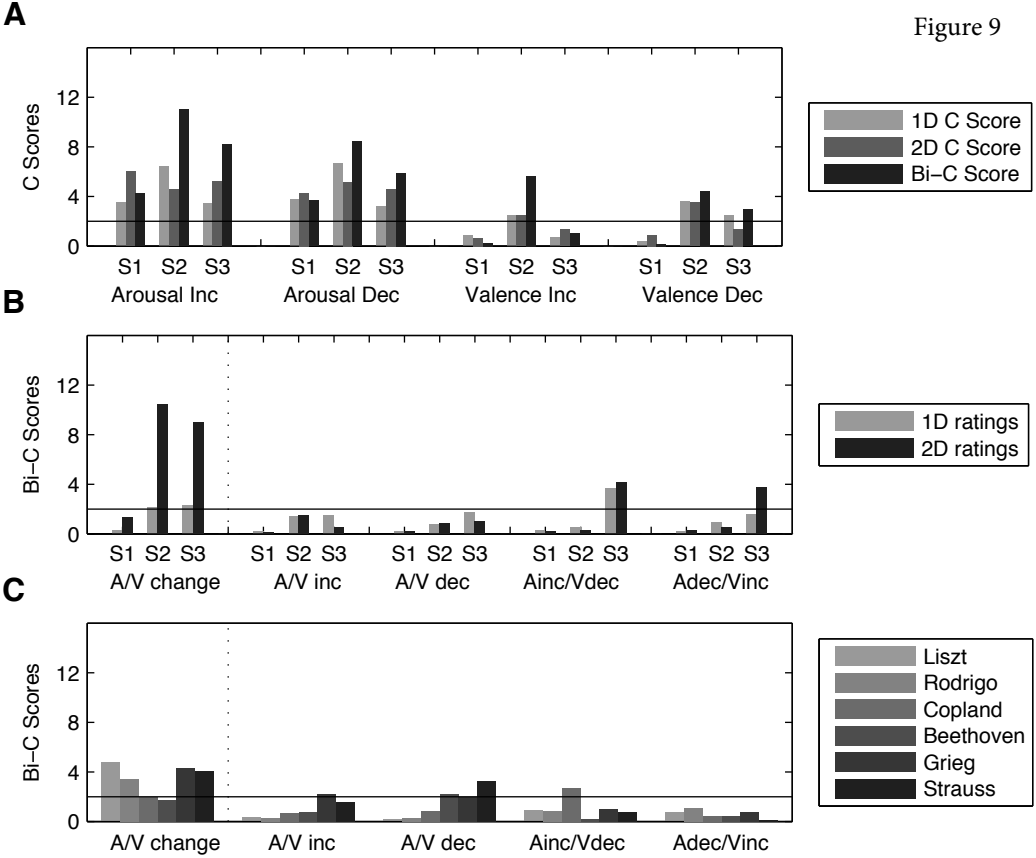

A
C Scores
12
8
4
0
S1 S2 S3
Arousal Inc
Arousal Dec
Valence Inc
Valence Dec
1D C Score
2D C Score
Bi–C Score
B
Bi–C Scores
A/V change
A/V inc
A/V dec
Ainc/Vdec
Adec/Vinc
1D ratings
2D ratings
C
Liszt
Rodrigo
Copland
Beethoven
Grieg
Strauss

Figure A.1

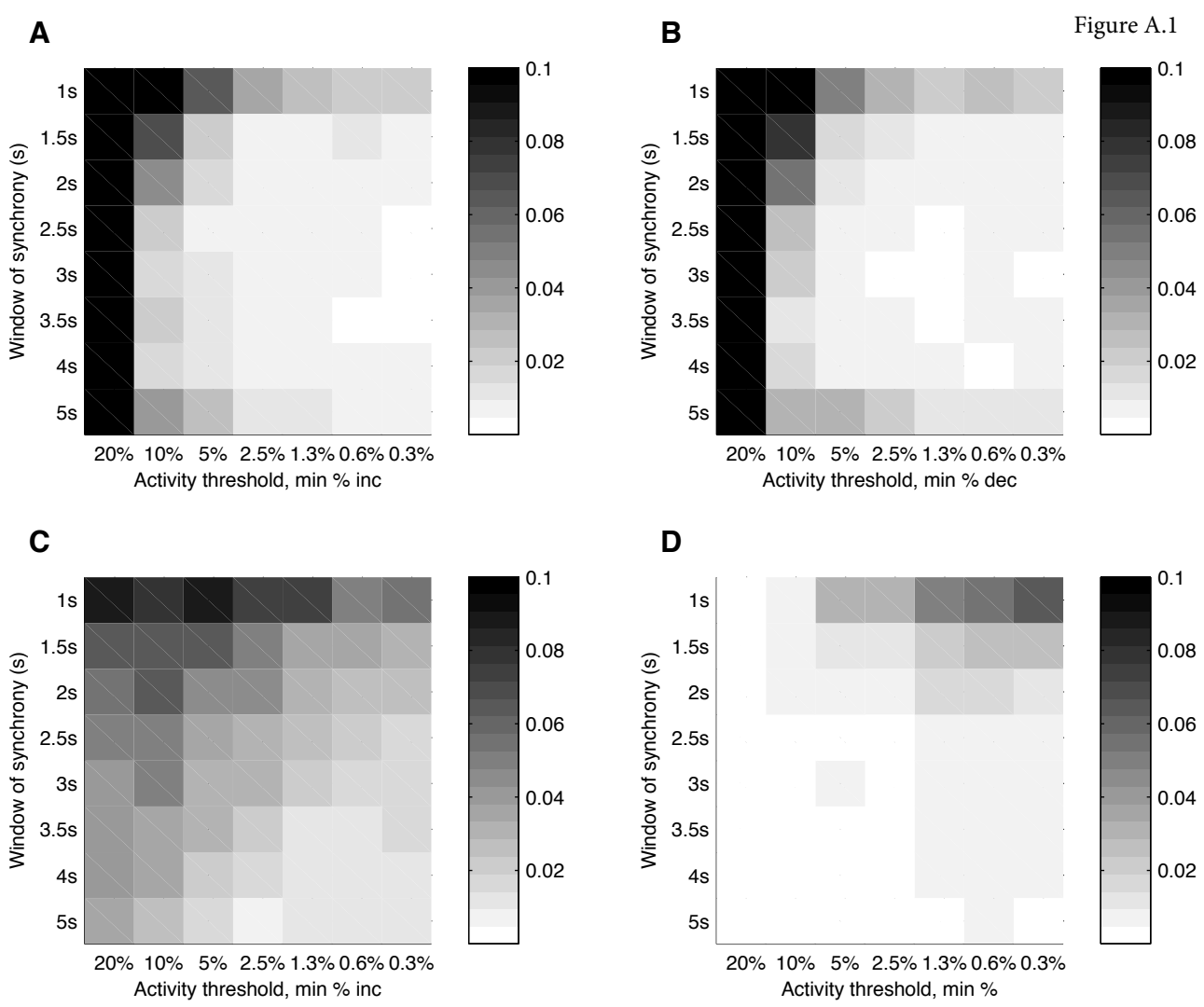
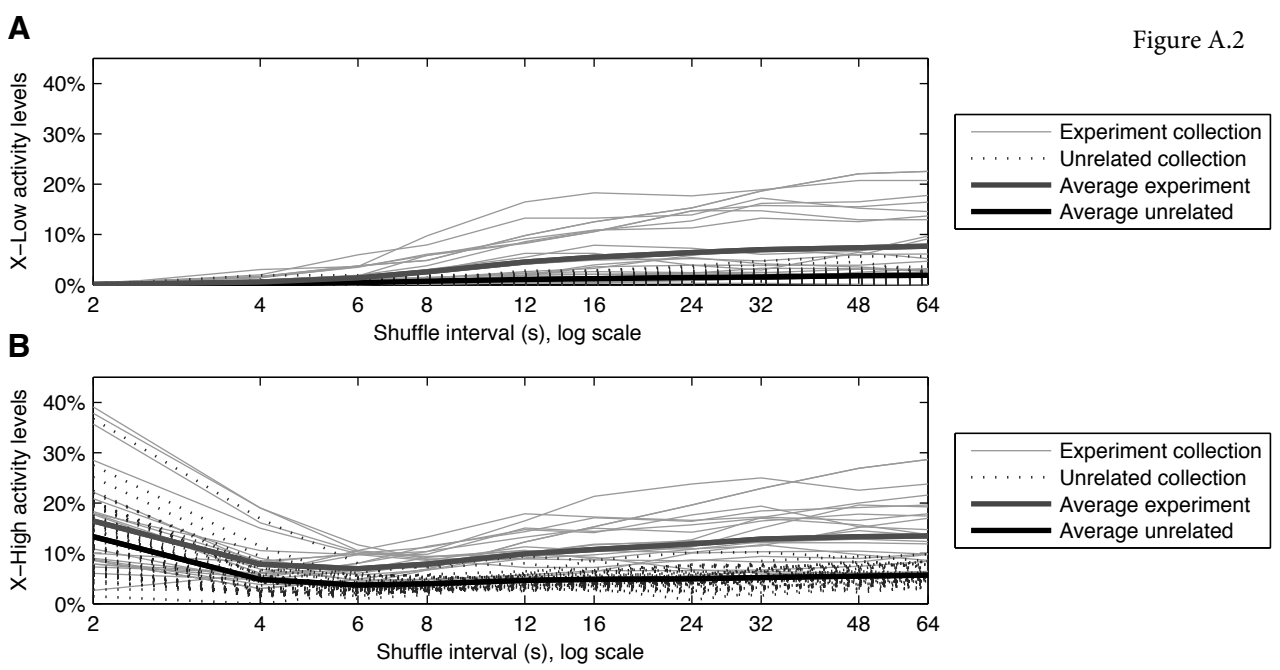
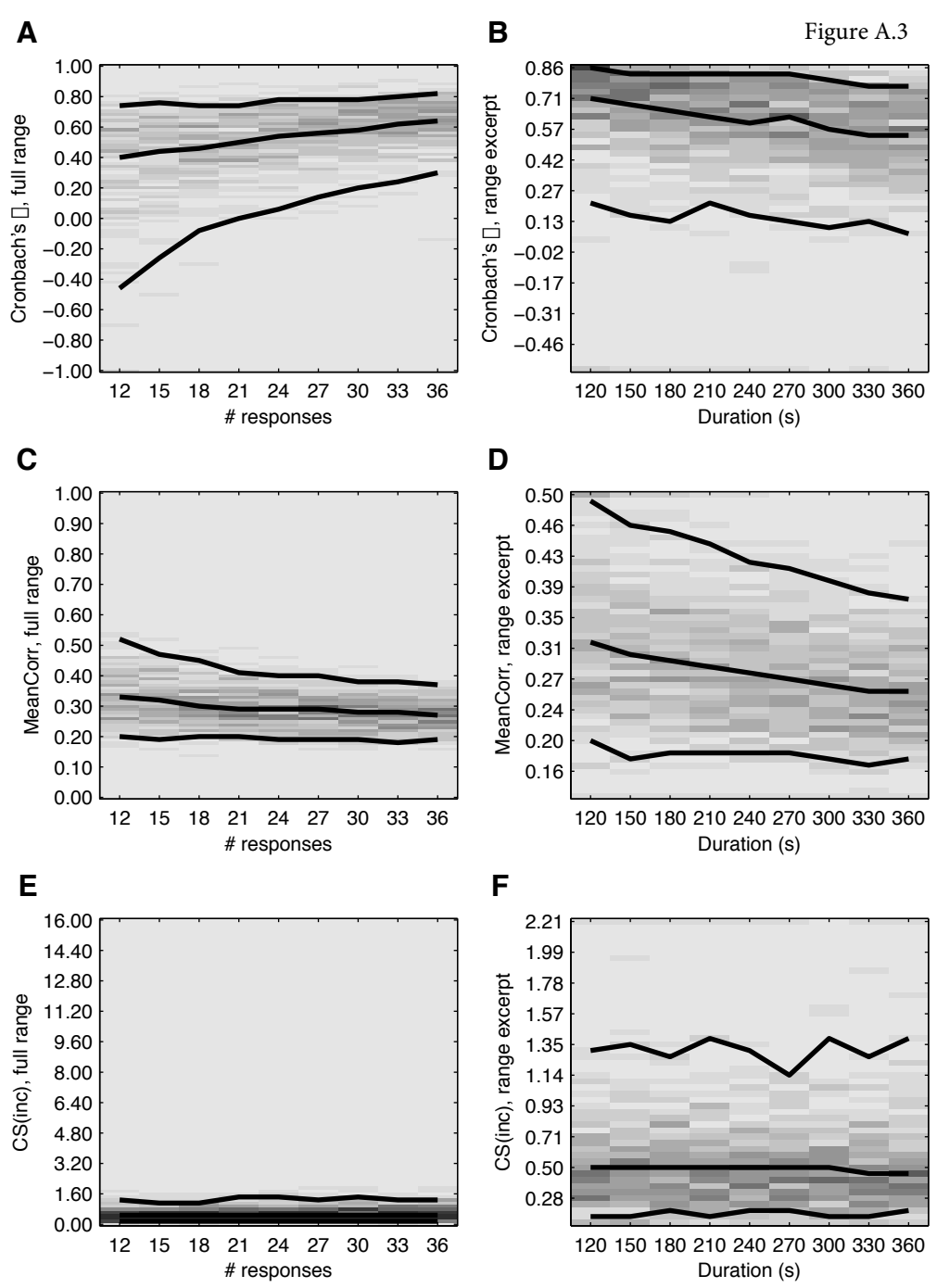

A
B
C
D
Window of synchrony (s)
1s
1.5s
2s
2.5s
3s
3.5s
4s
5s
20% 10% 5% 2.5% 1.3% 0.6% 0.3%
Activity threshold, min % inc
Activity threshold, min % dec
Activity threshold, min %
0.1
0.08
0.06
0.04
0.02

Figure A.2


A
X–Low activity levels
40%
30%
20%
10%
0%
2
4
6
8
12
16
24
32
48
64
Shuffle interval (s), log scale
Experiment collection
Unrelated collection
Average experiment
Average unrelated
B
X–High activity levels
40%
30%
20%
10%
0%
2
4
6
8
12
16
24
32
48
64
Shuffle interval (s), log scale
Experiment collection
Unrelated collection
Average experiment
Average unrelated


A
Cronbach's α, full range
1.00
0.80
0.60
0.40
0.20
0.00
−0.20
−0.40
−0.60
−0.80
−1.00
12 15 18 21 24 27 30 33 36
responses
B
Cronbach's α, range excerpt
0.86
0.71
0.57
0.42
0.27
0.13
−0.02
−0.17
−0.31
−0.46
120 150 180 210 240 270 300 330 360
Duration (s)
C
MeanCorr, full range
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
12 15 18 21 24 27 30 33 36
responses
D
MeanCorr, range excerpt
0.50
0.46
0.43
0.39
0.35
0.31
0.27
0.24
0.20
0.16
120 150 180 210 240 270 300 330 360
Duration (s)
E
CS(inc), full range
16.00
14.40
12.80
11.20
9.60
8.00
6.40
4.80
3.20
1.60
0.00
12 15 18 21 24 27 30 33 36
responses
F
CS(inc), range excerpt
2.21
1.99
1.78
1.57
1.35
1.14
0.93
0.71
0.50
0.28
120 150 180 210 240 270 300 330 360
Duration (s)