# Time-Dependent Characteristics of Performance Evaluation

Sam Thompson
*Royal College of Music, London, United Kingdom*

Aaron Williamon
*Royal College of Music, London, United Kingdom*

Elizabeth Valentine
*Royal Holloway, University of London, Egham, Surrey, United Kingdom*

**Concertgoers, critics, teachers, and performers** are often called upon to cast judgment on the performances they hear. Research to date has typically focused on the judgments themselves, with very few empirical studies of the processes and decisions that lead to these judgments. This paper details an investigation of time-dependent characteristics of performance evaluation. Thirty-three participants were played five recordings of a Bach Prelude and five of a Chopin Prelude. They rated the quality of each performance continuously, by moving a mouse cursor on a 7-point scale displayed on a computer screen, and using written scales. The results suggest that: the time taken to reach an evaluative decision was typically short (around 15-20 s); there was a significant difference between the initial and final ratings, with a tendency for ratings to improve as the performances progressed; and the largest revisions of opinion took place within the first minute of the performance.

*Received May 22, 2006, accepted April 3, 2007.*

**Keywords:** performance, evaluation, assessment, continuous response methodology, decision making

**Irrespective of domain, evaluative decisions** and judgments can be considered along two dimensions: outcome (the actual decision itself) and temporality (the dynamic process of reaching the decision). Because in the majority of cases the most important aspect of an evaluative decision is how good or bad the thing in question is thought to be, it is the former that is more familiar and more often discussed. That the temporal aspect of the decision—i.e., *when* the decision is made—should have received rather little attention in music is perhaps odd given that music, more so than other domains in which evaluation takes place, is intrinsically temporal.

It has been frequently noted that listeners' emotional responses to music vary throughout the course of a performance (e.g., Meyer, 1956; see Schubert, 2001, for a review). It is not the case, for example, that listeners hear a performance in a state of complete neutrality before enjoying a single emotional response only after it has concluded. Rather, they begin with some expectations about what they are to hear, experience a range of fluctuating emotions as the performance progresses, and only afterwards, if required, attempt to summarize the experience in a single statement.

If emotional reactions vary throughout a performance, it seems likely that judgments of quality will do likewise. It is barely conceivable that a performer could play such that the quality of performance remains constant throughout from the listener's perspective. It is more likely that the perceived quality of performance *will* vary somewhat; for instance, some passages may be performed flawlessly, yet others performed inelegantly or with technical slips. In any case, it might take some time for the listener to hear enough of the performance to gauge its quality.

How, then, is this judgment reached? There seem to be three possibilities. First, the final judgment reflects the "mean perceived quality" across the performance. In this case, reaching the final judgment would essentially be a memory-dependent task in which listeners explicitly recall and "average-out" the quality of the performance over time. Second, the final judgment is simply the product of a recency effect, such that the current perceived quality of the performance at the end of the piece is translated directly into the summary judgment because it is foremost in the listener's mind. Third, the final evaluation results from an evolution of judgment over time. As soon as some minimum amount of musical information has been heard, an initial decision is made. This has the status of a "working hypothesis" about the

*Music Perception* volume 25, issue 1, pp. 13–29, issn 0730-7829, electronic issn 1533-8312  DOI:10.1525/mp.2007.25.1.13

quality of the performance, which is subject to modification and adjustment as the performance continues. At some time before the end of the piece a stable point is reached, reflecting "finalization" of the judgment; this is then translated into the summary evaluation. In studies of musical performance evaluation to date there is little evidence that speaks directly to any of these three explanations. However, work from other areas of behavior has addressed similar issues.

### Temporal Aspects of Decision Making

Much work on decision making has been conducted in the context of economics and management, addressing the factors influencing consumer choice. Several interesting findings have emerged, notably in relation to the influence of prior knowledge about a product (Herr, 1989) and order effects (Moore, 1999). However, the kinds of preference decisions being considered bear little resemblance to those facing a listener trying to determine the quality of a performance. For a start, they are often based on discrete binary or tertiary choices ("choose product A, B, or C") and between options of no personal significance to the experimental participants. More importantly, there is little interest from researchers in temporal elements of the decision.

Decision research in occupational and social psychology has taken more notice of temporal issues. In the former area, the best known finding is in relation to hiring interviews, namely the (in)famous claim that interviewers tend to make "snap" decisions within the first two to three minutes of the interview commencing. Unbeknownst to many who expound it, the empirical basis of this claim is actually just a single study (Springbett, 1958), which has proven difficult to replicate (Buckley & Eder, 1988). In particular, it has been shown that time to decision in interviews varies as a function of other variables. If an interviewer knows the length of the interview beforehand, the time taken to reach a judgment will vary roughly in proportion to the total time available (Tullar, Mullins, & Caldwell, 1979). Moreover, decision time varies with perceived quality of the applicant, such that interviewers will be quicker to decide about low quality applicants than those who are perceived to be of higher quality (Tullar et al., 1979).

Whatever its basis, the notion that people make preference decisions quickly, spontaneously, and based on limited information has an undeniable intuitive appeal. The idea of making a "good first impression" is ubiquitous across many fields of endeavor, and has been given particular consideration in the context of social interaction. Research in this area suggests that evaluative decisions about people—e.g., opinions about their character, temperament, trustworthiness, and so on—evolve over time, with initial impressions formed on the basis of even scant information, but subject to revision as new, more comprehensive information is gleaned (Ybarra, 2002). Ybarra (2001) suggests that evaluative impressions about a person are likely to develop differently according to the valence of the initial impression. If a positive first impression is formed but negative information is subsequently learned, this is likely to lead to a downward revision of the evaluative judgment. By contrast, if a negative first impression is followed by positive information, this is unlikely to alter the initial impression. Positive first impressions are thus less "fixed" than negative impressions, and more susceptible to subsequent revision.

The process of forming an evaluative opinion about an individual along some dimension may be somewhat akin to the experience of hearing and evaluating a musical performance: initial impressions are formed in the early stages of the performance, and revised in the light of subsequent information as the performance progresses. Only one study was found that explicitly addressed the time taken to make an evaluative decision. Vasil (1973) asked a group of experienced evaluators to rate audio recordings of a set of performances. These were presented in a variety of different conditions, including the full performance (about six minutes long), half of the performance (i.e., the first three minutes), and a quarter of the performance (the first one and a half minutes). No significant difference was found between evaluations made in any of the three conditions. This could be reasonably taken to suggest—although Vasil does not draw the inference explicitly—that evaluations made on the basis of the whole performance may have already been "finalized" within the first 90 s, such that the remainder of the performance had little impact on the evaluative judgment.

### Temporal Differences in Between-Category Discrimination

One persistent problem in evaluation is that of limited between-category discrimination, whereby evaluators apparently fail to distinguish meaningfully between aspects of performance such as technique and musicality (e.g. Fiske, 1977; Thompson & Williamon, 2003). This is puzzling chiefly because there is a clear belief

amongst experienced music listeners that such discrimination is possible. Presumably, the persistence of this belief in the face of consistent evidence to the contrary is a product of phenomenological experience—listeners *feel* that they are able to distinguish between aspects of a performance as they listen.

Thompson and Williamon (2003) suggested four possible explanations for the discrepancy: 1) evaluators may be careless or hasty in making written evaluations; 2) despite beliefs to the contrary, evaluators are actually unable to distinguish between the performance aspects specified by most evaluation systems (e.g. the generic "technical/musical/communicative" categories); 3) nominally separate performance aspects are themselves causally related; or 4) in practice, few performances show significant disparity between aspects.

However, the discussion of temporal features of decision making, above, raises a fifth alternative. Reflection suggests that performance aspects might vary differentially over the course of a piece. Imagine, for example, a performance in which a player makes a bold, energetic and interpretatively original start to the work. In doing so, however, she also makes a number of small technical slips, none of which significantly interfere with the flow of the music but all of which would be noticeable to experienced listeners. In an effort to reduce the number of technical errors, the player begins to play more conservatively. The performance is technically immaculate thereafter, but at the expense of some interpretative originality as the performer elects to take fewer risks. In this case, summary ratings of the quality of technical and musical aspects of the performance—i.e., single marks given at the end—might stand at around the same moderate level, reflecting both that there were some initial technical problems and that the musically interesting interpretation was not sustained. However, in practice the two aspects of performance would have varied differently over time; the technical quality increased, and the musical quality decreased.

One plausible hypothesis, then, is that judgments about different facets of performance tend to develop differently over the course of a typical performance but somehow come together in the summary evaluation. In other words, it could be that there are temporal differences in the way that different aspects are perceived and evaluated, which are not reflected in the single summary judgments that—as argued above—represent an artificially static view of the evaluation process.

This leads to two questions. First, do ratings for different performance aspects vary differently over time? Second, if so, when do they converge? In addition to uncovering basic temporal processes of performance evaluation, this study aimed to examine these questions by asking different groups of listeners to focus on different aspects of the performance as they gave continuous ratings. This would allow the pattern of ratings of different performance aspects to be compared as a function of time.

## Aims and Objectives

The present study was designed to investigate temporal aspects of musical performance evaluation. Three groups of musically experienced participants were asked to listen to a number of performances and, respectively, give continuous evaluations of quality along three dimensions: Overall Quality, Technical Proficiency and Assurance, and Musicality. The study addressed the following research questions:

1. How long do listeners take to reach an initial evaluative judgment?
2. How, and how frequently, does this judgment vary during the course of a performance?
3. How do written summary evaluations of the performance relate to the continuous evaluation?
4. Do perceptions of the quality of different aspects of a performance show different patterns of variation over time?

In recent years, continuous response methods have been chiefly employed to investigate emotional responses to music, with participants required to track their perceptions of the music's emotional intensity whilst listening (Madsen, 1997, 1998; Schubert, 1999, 2001, 2004; Schubert & Dunsmuir, 1999; Sloboda & Lehmann, 2001). These studies have usually attempted to identify structural features in the music responsible for eliciting observed patterns of emotional variability. In the present study, continuous response methodology was used differently—to investigate dynamic aspects of the performance evaluation process.

## Method

### *Participants*

Thirty-three participants were recruited for the study (15 men, 18 women). All were actively involved with classical music, as players (amateur or professional), teachers, or researchers. Twenty-four held either a first degree in music and/or a recognized performance diploma.

All were regular concertgoers, reporting an average of 28 classical concerts attended in a typical year.

### *Recordings*

The intention of the study was to identify general patterns in the evaluation process rather than link specific features of music to different patterns of response. Hence, it was necessary to produce a set of performances that were distinct from one another yet broadly similar in terms of overall quality. To achieve this, two pieces of music were chosen, Chopin's Prelude in B minor, Op. 29 No. 6, and the Prelude in G minor from Bach's Well Tempered Clavier Book 1. These were both in minor keys, with approximately the same length in performance (c. 2 minutes) and a similar basic tempo. Two pianists were asked to learn the pieces to performance standard. One was a second year postgraduate performance student at the Royal College of Music (Pianist 1); the other was a junior research fellow at the College and a semiprofessional pianist (Pianist 2). When they had fully rehearsed the two pieces, they recorded polished performances on a Yamaha Disklavier upright piano. To produce a number of additional, similar performances, the pianists recorded two further versions of each piece; for the first, they were asked to play at a deliberately faster tempo than they would ordinarily adopt and to limit the use of rubato, and for the second to play at a deliberately slower tempo with exaggerated rubato. All three performances from each pianist were recorded as MIDI files. Piloting suggested that Pianist 2's two faster performances were unconvincing as realistic interpretations. Consequently, only five performances of each work were employed.

MIDI files for each performance were edited using MIDI Maestro 2 (Pletzer, 2002) to remove any obvious wrong notes and, in the case of the Bach, to ensure that trills were evenly spaced. This was necessary because of slight limitations in the playback mechanism of the Disklavier, such that notes with low velocity (the MIDI control parameter specifying loudness) did not sound evenly (or at all) when played back.

Subsequent to editing, each performance was replayed on the Disklavier and acoustically recorded onto minidisc using a studio-quality microphone. Recording took place in a sizeable room so as to provide a realistic acoustic. These digital recordings were subsequently sampled to PC and edited using Audacity (Mazzoni, 2004). Exactly 4 s of silence was inserted before the start of each performance to give participants a brief pause before the start of each trial. The resulting files were saved in MP3 format.

### *Performance Evaluations*

Performance evaluations were given in two ways: (1) in real time during each performance, using a computer-based continuous response scale; and (2) after each performance, on a number of written 7-point scales.

#### CONTINUOUS EVALUATION

New computer software was used to conduct the study and record the continuous response data, written to a specification devised by the authors (Fenech, 2003). At the start of each experimental trial an MP3 file was triggered. A response area in the form of a single coloured band appeared near the bottom of the computer screen. When the mouse cursor was moved onto the area, its horizontal position was sampled at regular time intervals—thus, the participant was able to give quasicontinuous responses by moving the cursor horizontally along the band. Data output from the program was given in the form of a spreadsheet with two columns, representing: (1) the time (in ms) of each sample after the start of the trial; and (2) the position of the cursor.

Sample rate and scale were both user-defined, and the same software settings were used for all trials and for all participants. A sampling rate of 500 ms was chosen, following Schubert (2001). The horizontal position of the cursor was recorded on a scale of 1 to 70—this ensured that continuous measurement data could be mapped straightforwardly onto the 7-point scales used in the written evaluation task (see below), while providing high "resolution" data output based on the small incremental movements of the cursor. As a guide for participants, numbers from 1 to 7 were displayed on screen above the coloured rating area.

#### WRITTEN EVALUATIONS

After each performance, participants gave written evaluations of the performance along three dimensions: Overall Quality, Technical Proficiency and Assurance, and Musicality. After hearing the five performances of each piece, they also gave three summary ratings indicating: (1) how *difficult* they thought the piece was for a pianist of conservatoire level to play well; (2) how much they *liked* the piece; and (3) how *familiar* they were with the piece before the performance. All ratings were given on 7-point scales.

### *Procedure*

Participants were randomly assigned to three experimental groups of $N = 11$. (Twelve participants were first-study pianists, and were distributed equally amongst

the three groups.) Within each group, participants were assigned numbers from 1 through 11; hereafter, they are referred to by these numbers and the first letter of their group, e.g., Q1, T6. Participants in Group Q were asked to rate the overall quality of each performance, taking into account all aspects of the performance. Those in Group T were asked to rate *just* the technical proficiency and assurance of each performance, consciously detaching it from their opinion of the musical or interpretative aspects. Participants in Group M were asked to rate *just* the musical and interpretative aspects of each performance, again consciously detaching it from their opinion of the technical proficiency and assurance.

The study took place in a laboratory at the Royal College of Music. Participants sat in front of a computer monitor and listened to the performances through headphones. Prior to commencement of the study, participants were given oral instructions explaining the procedure. Two details were emphasized: 1) that they should not move the mouse cursor onto the colored rating area until they had heard enough to reach an initial evaluation; and 2) that they should feel free to adjust this rating as little or as often as they wished throughout the performance.

Each participant heard all ten performances. The order of presentation of the two pieces was counterbalanced across participants and the order of the five performances of each piece was randomized. To provide context, participants were told that all the performances were recorded at a competition recently held at a UK music college. After each trial there was a short pause while participants removed the headphones. They were then asked to complete the written evaluations.

After each session, the participant was debriefed about the purpose of the study. They were also asked to what extent they felt the continuous measurement methodology enabled them to reflect accurately their judgements. Experimental sessions were conducted individually and took approximately 45 minutes. Participants were unpaid.

## Results & Discussion

### *Data Handling*

Raw data from the study comprised over 80,000 individual data points, each consisting of two components: a time value (i.e., the time from the beginning of the trial to the time at which the sample was taken) and a rating (from 1 to 70). Initially, it was necessary to perform several data handling and screening procedures. Some of these applied to all analyses, and these are described in this section.

Consideration of the data revealed a problem regarding accuracy of the first mouse movement. Many of the data sets began with a short flurry of movement, suggesting that participants took 1 or 2 s to settle the mouse pointer in the intended position. However, they had been explicitly requested *not* to move the pointer onto the rating area until they had decided on an initial rating. Some experimentation revealed that it was difficult to move the mouse pointer onto the rating scale and settle at the intended position without crossing several data sections and, in doing so, register a series of short movements in the first 1 or 2 s of data recording; this was an artifact of the width of the rating area on the screen, and the physical difficulty of moving the mouse cursor precisely downwards onto it. To correct for this methodological error, the following criterion was used: If the recorded value moved more than four units within the first four samples (i.e., within the first 2 s of moving onto the rating area), these values were adjusted to the first stable rating. A stable rating was defined as one that remained constant for more than four samples, i.e., 2 s. In the large majority (93.3%) of cases, this procedure yielded an initial value that remained constant for at least the first 2 s.

Note that in all of the repeated measures analyses that follow, $F$ values are reported with the Greenhouse-Geisser correction where the assumption of sphericity was violated. In the interest of preserving space, normality statistics have not been given; however, the majority of dependent variables used in the analyses were found to be normally distributed. Repeated measures ANOVA is in any case considered to be relatively robust to violations of normality where group sizes are equal, and above 10.

### *Preliminary Analysis*

Figures 1a and b show graphs of the mean written ratings for Overall Quality, awarded to the five Bach and Chopin performances. These can be thought of as the evaluations that the performances might have received in a "typical" evaluation, in which just a single final mark was awarded. As can be seen from the graphs, the performances did differ somewhat in their overall rating, but by comparatively little; aside from slight fluctuations, the levels of rating were essentially the same across performances. For the Bach, the difference between the highest and lowest rated performances was just 0.86, less than one whole scale point. At 1.88, the difference between highest and lowest Chopin performances was slightly larger, but still

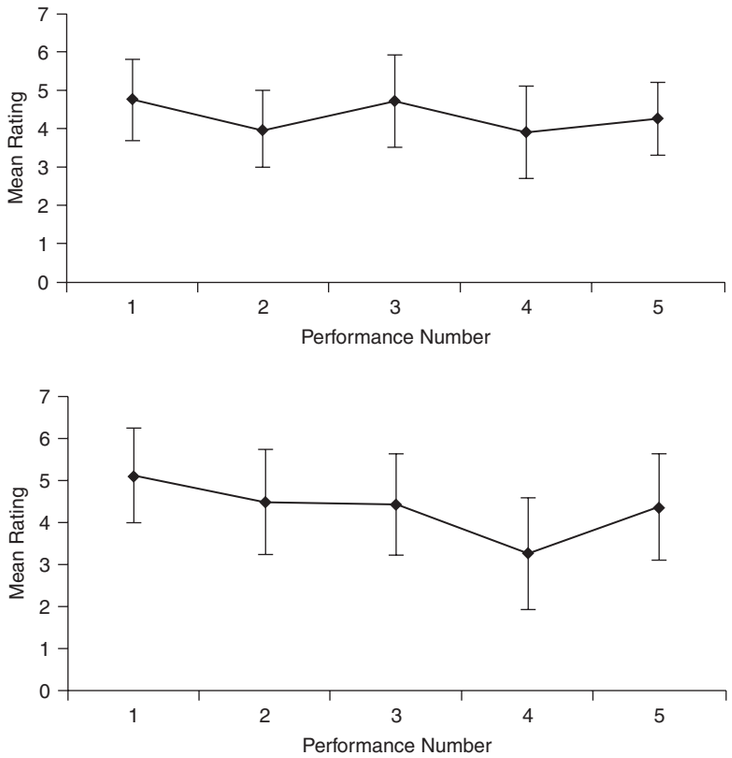


**FIGURE 1. Mean written ratings of overall quality of (a) Bach and (b) Chopin performances, across sample. Error bars show +/– one standard deviation.**

small in the context of a 7-point scale. None of these differences were statistically significant.

Given that no one performance of either piece was rated as being of particularly high or low quality, mean values were taken across all five performances for all of the subsequent between-group analyses.[1]

## Analysis 1: Characteristics of "Process" Variables

### *Start Time*

The distribution of Start Time—the period of time, to the nearest 0.5 s, elapsed between the beginning of the performance and the first reported judgment—across the complete data set (i.e., all trials, for both pieces) was approximately normal but somewhat positively skewed. Several outliers occurred outside the curve suggested by observed distribution. Discussions with participants directly following the experiment suggested that the few extremely late start times were a result of participants "forgetting" to move the mouse cursor, as they had been engrossed in the music. Cases in which the start time was greater than 2/3 of the total length of the performance were therefore eliminated. This removal criterion led to four apparently outlying data points being rejected.

To test for a possible effect of familiarity on the experimental task, a repeated measures ANOVA was calculated for each piece, with Performance Order as a within-subjects variable with five levels, and Experimental Group as the between-subjects variable. In neither case was there a significant main effect either of order of performance or of group, or any significant interaction between the two. An order effect was thus ruled out.

Table 1 gives descriptive data for Start Time. Mean Start Time was found to be 19.27 s. However, given the positive skew of the distribution, it is arguable that the

[1]Note in any case that, in a repeated measures ANOVA such as those used below, between-subjects effects of group would return the same critical value of *F*, irrespective of whether the model incorporated five within-subjects levels (i.e., for each performance) or was calculated on the mean value taken across the five performances of each piece. Since differences between performances were not subject to investigation in the present context, all ANOVAs for group differences used mean values.

**TABLE 1. Descriptive Data for Start Time (in Seconds; Outliers Removed) for each Group and Overall.**

| | Overall | Group *Q* | Group *T* | Group *M* |
|---|---|---|---|---|
| *Across Performances* | | | | |
| Mean | 19.27 | 18.09 | 21.11 | 18.61 |
| Standard Deviation | 14.90 | 11.31 | 17.86 | 10.28 |
| Median | 14.00 | 15.00 | 13.50 | 13.00 |
| Range | 12.26 | 12.16 | 12.10 | 12.50 |
| *Bach* | | | | |
| Mean | 18.19 | 17.61 | 20.07 | 16.06 |
| Standard Deviation | 14.00 | 10.49 | 17.13 | 12.67 |
| Median | 13.00 | 14.50 | 13.25 | 10.75 |
| Range | 12.55 | 13.04 | 11.85 | 12.45 |
| *Chopin* | | | | |
| Mean | 20.36 | 18.56 | 22.06 | 20.81 |
| Standard Deviation | 15.72 | 12.14 | 18.82 | 15.65 |
| Median | 15.00 | 15.50 | 13.50 | 16.00 |
| Range | 11.78 | 11.17 | 13.15 | 12.03 |

median value—14.00 s—provides a more accurate measure of central tendency. What is clear from either statistic is that the typical time required to make an initial judgment was short—somewhere between 14 and 20 s.

To examine whether Start Time differed as a function of experimental group, a repeated measures ANOVA was conducted with Piece as the within-subjects variable and Group membership as the between-subjects variable. No significant main effects or interactions were found, suggesting both that the groups did not show differences in the average time taken to make an initial decision, and that this value was itself approximately the same for both pieces.

### *Movements per Minute*

Participants were instructed to move the mouse cursor only when their evaluative judgment changed. Because Start Time differed for each trial, and the performances were of slightly differing lengths, a scaled variable was defined, Movements per Minute, according to the following formula:

$$\text{Movements per Minute} = 60 \times \frac{\text{Number of movements}}{\text{Length of performance} - \text{Start Time}}$$

This was averaged across performances per person, to yield the mean number of movements made on average per minute.

Identifying "single" movements was not always straightforward. In particular, there was a need to distinguish cases where a participant moved slowly from one stable point to another but apparently within the same gesture, from cases where the participant made a succession of apparently discrete decisions but in comparatively quick succession. A procedure akin to low-pass filtering was adopted (e.g., see Gottman, 1981). If during the course of the performance the mouse cursor was held in the same position for more than 2 s (i.e., if more than 4 consecutive data points were the same) this was regarded as a stable period. Any individual movements that occurred between stable periods (thus defined) were regarded collectively as one single movement.[2]

The distribution of Movements per Minute, as with Start Time, was somewhat positively skewed. Moreover, 28 trials exhibited no movement at all—that is, the participants did not adjust their initial decision as the performance progressed. Table 2 gives average Movements per Minute by group for each piece.

A repeated measures ANOVA with Piece as the within-subjects variable and Group as the between-subjects variables showed no effect of Group. Table 2 suggests that, across groups, participants made slightly fewer changes of judgment per minute during the

[2]This method does not account for "jittery" responses such as a movement, for example, from 49 to 50 and back to 49 over three consecutive samples. However, no such responses were observed in the data.

**TABLE 2. Mean (SD) Movements per Minute for each Group and Overall.**

| | Overall | Group *Q* | Group *T* | Group *M* |
|---|---|---|---|---|
| Across Pieces | 2.62 (2.20) | 2.48 (1.73) | 2.71 (2.70) | 2.66 (2.08) |
| Bach | 2.80 (2.28) | 2.75 (1.96) | 2.78 (2.68) | 2.89 (2.14) |
| Chopin | 2.43 (2.12) | 2.21 (1.42) | 2.64 (2.74) | 2.44 (2.02) |

Chopin (2.80 movements/min) than the Bach (2.43 movements/min) performances, but this was not statistically significant.

To assess whether *density* of movement changed over time, i.e., whether participants tended to make movements at the same rate across a whole performance, for each trial the number of movements made between current rating and final rating was calculated at intervals of 15 s from the beginning of the performance. 5 s was chosen since this was, approximately, the typical (i.e., median) time taken to make an initial decision, ensuring that over half of the trials had registered a rating within the first time section, and some 80% registered a rating within the second time section. Trials in which no movements at all were made were excluded completely.

The number of movements in each 15 s section was calculated as a proportion of the total number of movements in the trial. Mean values were then calculated per section, per person, and used in a repeated measures ANOVA with Piece and Section (eight levels) as within-subjects variables and Group as a between-subjects variable. There was no effect of Piece or Group. However, there was a significant effect of Section, $F(4.16, 99.88) = 12.43$, $p < .001$), with polynomial contrasts suggesting a significant quadratic trend, $F(1, 24) = 163.28$, $p < .001$); this can be clearly seen in Figure 2. Note also that the mean pattern of movements density for Groups T and M tended to peak markedly towards the end of performances.

### *Individual Variability*

#### START TIME

Considered across the whole sample, there was a good deal of variability in Start Time—even with the extreme outliers removed, the standard deviation was 14.90 s, which is high relative to the mean of 19.27 s. One question is whether this can be attributed to particular individuals, such that some were apt to display more variability or consistency than others. Alternatively, it could be a generic feature of the task (and maybe, by

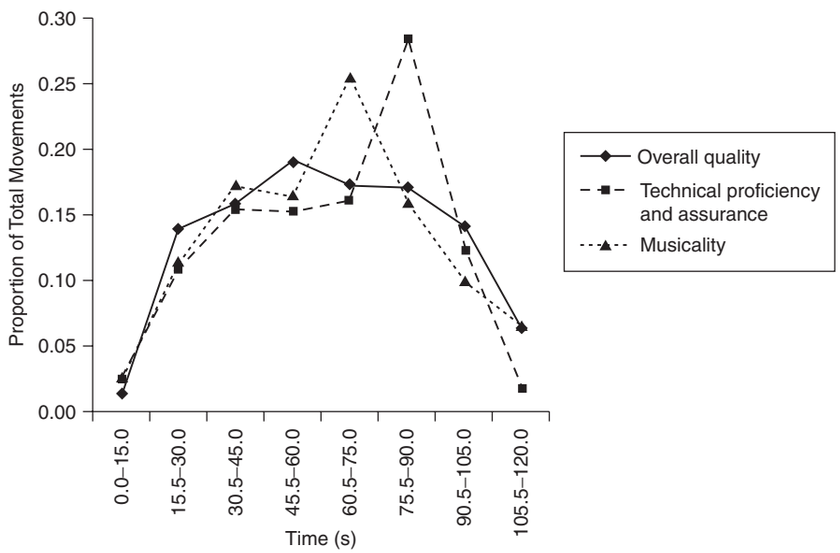


FIGURE 2. Mean proportion of movements in 15-s segments (trials with no movement omitted).

extrapolation, of performance evaluation more generally) that the time to make an initial evaluative decision varies substantially on a case-by-case basis.

Since standard deviation tends to be correlated with the mean (since the range of possible start times increases as Start Time itself increases), a better measure of variability than pure standard deviation, is standard deviation *as a proportion* of the mean start time. Hence, the standard deviation divided by the mean was calculated for each individual. The correlation of this scaled value with the mean for each individual was small and nonsignificant, $\rho(31) = -.14$.

Tendency to show variability in time to initial decision therefore appeared to be unrelated to the actual mean time itself. In other words, people who took longer on average to make their initial evaluative decision appeared to be no more or less likely to show wider variability in the time taken over several consecutive performances than those who made initial decisions more quickly.

#### MOVEMENTS PER MINUTE

A measure of personal variability in Movements per Minute was calculated by dividing the standard deviation per participant, across all performances, by the respective mean value for that participant. A correlation coefficient of $\rho(31) = -.76$, $p < .001$ suggested a negative relationship, such that those participants who made the most movements—i.e., modified their judgment most frequently throughout a performance—also tended to be most consistent in their number of movements.

### *Discussion: Process Variables*

Previous estimates of the time taken to make an evaluative decision about a musical performance have been sketchy and largely based on work conducted in other, notionally comparable domains. However, as discussed, extrapolating from these other fields of study is fraught with difficulty.

Looking across both pieces and rounding to the nearest 0.5 s (as appropriate given the sample rate of 2 Hz), it could be said that the typical time taken to make an initial evaluative judgment of a performance in the present study was 19.5 s. However, given the markedly positive skew of the distribution, the median statistic gives a better indication of the "typical" time to initial decision. This reduces the average time to initial decision by approximately 5 s. Whichever measure is preferred, however, it must be noted that this is a short time-span indeed, even as compared with previous estimates.

Analysis of between-subjects effects on Start Time found no evidence of differences by experimental group. This can be taken to suggest that opinions about the technical and musical merits of each performance, and of the overall quality, took approximately the same time to form. In respect to Start Time at least, then, there is no evidence of between-category discrimination.

Across all groups, the mean number of movements made per minute was 2.62. Again, the distribution was appreciably positively skewed; however, the difference between mean and median was small. Once more, no group differences were observed, therefore providing no evidence of any procedural difference between ratings for overall quality, technical, or musical aspects of the performance. It is notable, as an aside, that some 28 trials (of 330) exhibited absolutely no movement whatsoever: i.e., participants found no reason to adjust their initial rating as the performance progressed. This was permissible, and was in fact explicitly mentioned to participants as an option when the experimental procedure was explained. However, the low number of such trials means that in the large majority of cases participants *did*, in fact, go on to revise their initial impressions on the basis of information subsequently received.

The rate of movements began low, increased towards the middle of the performance and decreased towards the end (as confirmed by the significant quadratic effect). However, despite the lack of a significant between-subjects effect, the pattern of movement density over time was somewhat different between groups. Groups *T* and *M* both showed appreciable "spikes" towards the end of the two minutes. This is the first evidence of some difference between the groups, although it is not clear what it could be attributed to.

Both Start Time and Movements per Minute exhibited considerable *intra*-rater variability, suggesting inconsistency from trial to trial. Studies of rating in different fields have tended to find broadly the same result, namely that trial-to-trial variability in process is high. Future research is required to establish if such variability is endemic to evaluation per se, an artifact of this particular task, or even a chance anomaly.

## Analysis 2: Characteristics of Rating Variables

### *Written Evaluations*

Table 3 gives descriptive statistics for the written evaluations. To test for differences between the written evaluations by group and segmented category, a repeated measures ANOVA was conducted with Group as the between-subjects variable and both Piece and Category as within-subjects variables. No effect of Group was

**TABLE 3. Mean (SD) Written Evaluations for Each Group and Overall.**

| | Overall | Group *Q* | Group *T* | Group *M* |
|---|---|---|---|---|
| *Across Pieces* | | | | |
| Overall Quality | 4.30 (0.52) | 4.24 (0.62) | 4.41 (0.80) | 4.13 (0.63) |
| Technical Proficiency and Assurance | 4.45 (0.64) | 4.25 (0.21) | 4.24 (0.38) | 4.02 (0.51) |
| Musicality | 4.17 (0.57) | 4.41 (0.63) | 4.70 (0.63) | 4.36 (0.56) |
| *Bach* | | | | |
| Overall Quality | 4.30 (0.47) | 4.34 (0.41) | 4.57 (0.84) | 4.13 (0.46) |
| Technical Proficiency and Assurance | 4.50 (0.69) | 4.33 (0.39) | 4.35 (0.65) | 3.84 (0.70) |
| Musicality | 4.07 (0.58) | 4.24 (0.62) | 4.56 (0.61) | 4.24 (0.54) |
| *Chopin* | | | | |
| Overall Quality | 4.30 (0.78) | 4.15 (0.99) | 4.25 (1.01) | 4.14 (1.00) |
| Technical Proficiency and Assurance | 4.41 (0.80) | 4.18 (0.41) | 4.13 (0.43) | 4.20 (0.54) |
| Musicality | 4.27 (0.79) | 4.58 (0.81) | 4.85 (0.72) | 4.48 (0.78) |

found, and there was no Group by Category interaction. Thus, the task itself did not significantly influence these post-performance judgments.

However, there was a significant effect of Category, $F(1.67, 49.98) = 7.71$, $p < .005$, and the Piece by Category interaction approached significance, $F(2, 60) = 3.00$, $p = .057$. Simple contrasts confirmed that performances were given higher ratings for Technical Proficiency and Assurance than Overall Quality, $F(1, 30) = 4.60$, $p < .05$, or Musicality, $F(1, 30) = 5.24$, $p < .05$. However, this difference was only evident for the Bach (hence the near-significant interaction effect), and the magnitude of difference was very small—less than half a point between Technical Proficiency and Assurance and Musicality.

### *Characteristics of Continuous Ratings*

Table 4 gives the mean Initial Rating, Final Rating, Range (i.e., the difference between minimum and maximum ratings per trial), and Difference (i.e., final rating minus initial rating). To investigate differences between piece and group, repeated measures ANOVAs were calculated for Initial Rating, Final Rating, Range, and Difference. Group was the between-subjects variable and piece was the within-subjects variable. No main effects or interactions were found for Initial Rating, Final Rating, or Range. For Difference, no within-subjects effects or interactions were observed. However, the between-subjects effect of Group approached significance, $F(2, 30) = 3.24$, $p = .053$. Whereas both Groups *Q* and *M* showed a mean positive difference between initial and final rating, the value of the mean difference for Group *T* was only slightly above zero for the Bach, and somewhat below zero for the Chopin. In other words, it seems that opinions about the performers' technical assurance remained at about the same level or even went slightly downwards throughout the course of each performance. By contrast, opinions of the performances' overall quality and musical content tended to improve, albeit moderately, as the music progressed.

### *Relationship Between Continuous and Written Evaluations*

The following analysis considered the extent to which the ratings given via continuous measurement were similar to the final written scores. As a preliminary, to enable the analysis, Initial and Final scores from each individual continuous evaluation were mapped onto the written scale. Mean ratings were then calculated for each participant, for each piece.

Three separate analyses were conducted, with each group's data compared with their respective written evaluation. For all three analyses a repeated measures ANOVA was calculated with three levels of within-subjects measures: Initial rating, Final rating, and Written rating, entered into the model so as to correspond with the chronological order in which they were made.

For Group *Q*, a significant within-subjects effect of rating was found, $F(2, 20) = 3.77$, $p < .05$. Repeated contrasts revealed a significant positive difference between Initial and Final ratings, $F(1, 10) = 6.13$, $p < .05$, but no significant difference between Final and Written ratings. This same pattern of results was found

**TABLE 4. Mean (SD) Initial Rating, Final Rating, Range and Difference for Each Group and Overall.**

| | Overall | Group *Q* | Group *T* | Group *M* |
|---|---|---|---|---|
| *Across Pieces* | | | | |
| Initial Rating | 35.63 (5.29) | 35.17 (6.01) | 37.78 (3.42) | 33.93 (5.76) |
| Final Rating | 37.74 (5.87) | 38.59 (6.95) | 37.39 (3.21) | 37.26 (7.11) |
| Range | 12.26 (6.48) | 12.16 (4.22) | 12.10 (9.99) | 12.50 (4.08) |
| Difference | 2.12 (4.26) | 3.42 (4.54) | −0.38 (2.68) | 3.32 (4.47) |
| *Bach* | | | | |
| Initial Rating | 35.26 (10.51) | 34.89 (8.53) | 37.53 (10.93) | 33.36 (11.58) |
| Final Rating | 37.93 (12.85) | 38.93 (11.97) | 38.24 (13.57) | 36.64 (13.10) |
| Range | 12.48 (9.35) | 12.55 (6.78) | 13.04 (11.94) | 11.85 (5.16) |
| Difference | 2.67 (10.96) | 4.04 (10.35) | 0.71 (11.87) | 3.27 (10.51) |
| *Chopin* | | | | |
| Initial Rating | 35.99 (10.36) | 35.45 (11.58) | 38.03 (8.58) | 34.49 (10.55) |
| Final Rating | 37.55 (14.71) | 38.25 (15.24) | 36.55 (13.90) | 37.85 (15.16) |
| Range | 12.03 (9.04) | 11.78 (8.20) | 11.17 (10.24) | 13.15 (8.59) |
| Difference | 1.56 (11.19) | 2.80 (11.27) | −1.48 (11.45) | 3.36 (10.39) |

for Group *M*, with a significant within-subjects effect of rating, $F(2, 20) = 8.68$, $p < .005$, and repeated contrasts showing a significant positive difference between Initial and Final rating, $F(1, 10) = 7.03$, $p < .05$, but no significant difference between Final and Written.

For Group *T*, however, no significant within-subjects effects or interactions were found. This is unsurprising when considered in relation to the finding of the previous analysis that initial-to-final difference scores for this group did not show an appreciable increase or decrease.

### *Discussion: Rating Variables*

The written evaluations showed no effect of experimental group, suggesting that the process of listening deliberately to just technical or just musical aspects of the performances did not have an appreciable impact on listeners' summary evaluations. This implies one of two explanations. Listeners in respective groups may have been successfully able to isolate their evaluations of the technical and musical aspects as each performance progressed, whilst simultaneously absorbing sufficient information about the performance as a whole to make judgments about all aspects of it at the end. Alternatively, listeners in Groups *T* and *M* may have been unable to separate the technical and musical aspects of the performance in a consistent way, and instead gave continuous evaluations that did not differ appreciably from those given by Group *Q*. Since participants as a whole agreed that there were overall differences between aspects of the performances (as written ratings for Technical Proficiency and Assurance were significantly higher than those for Musicality or Overall Quality when considered across all performances), the answer to which explanation should be preferred hangs on whether any group differences in ratings were observed during the continuous evaluation itself.

In fact, analysis of this data reveals the first evidence so far of systematic group differences in the evaluation process. Whilst there were no differences between the absolute levels of mean Initial rating, Final rating, or Range, there was a clear and significant group difference between initial-final difference scores. Whereas the evaluations of participants in Groups *Q* and *M* tended to rise over the course of the performance by 3-4 points on average, participants in Group *T* exhibited very little difference between initial and final ratings. It must be noted that the magnitude of difference being considered here is very small in the context of a 70-point scale; that it should still be statistically significant, however, is perhaps therefore all the more convincing. One explanation could be that opinions about the technical competence of performances were subject to comparatively less overall change. However, this seems strange when coupled with the fact that no group difference was observed on the variable Movements per Minute. In other words, participants in Group *T* did not appear to move any less frequently than those in the other two groups – it was not the case that they simply made their decision and stuck to it.

The same group difference is emphasized by the analysis comparing continuous and written evaluations. For Groups *Q* and *M*, initial ratings were found to be

significantly different from final, but there was no significant difference between final and written evaluations; for Group *T*, by contrast and as would be expected given the lack of initial-to-final Difference already noted, there was no significant difference between any of the three measures.

This result related directly to the issue noted above regarding whether summary judgments are the result of a mean of perceived quality over time, a recency effect, or an evolving judgment. The initial-final difference observed in Groups *Q* and *M* suggests that the first explanation is unlikely. However, the result is consistent with either of the second two explanations.

## Analysis 3: Relationships Between Rating and Process

This final analysis combines the variables defined in parts 1 and 2 by considering how evaluative judgments changed over time.

### *Rating as a Function of Time*

#### INITIAL-TO-FINAL DIFFERENCE SCORE

Mean values of current-to-final difference score were calculated per person at intervals of 15 s. In six cases it was not possible to give a mean current-to-final difference score for the first time-point, since participants gave responses later than 15 s on all five trials for one or other of the pieces. These cases were thus omitted from the analysis. Trials in which no subsequent movement was observed were also omitted from the calculations of individual means.

A repeated measures ANOVA was calculated with Piece and Time-Point (8 levels) as within-subjects variables and Group as the between-subjects variable. No main effect of Piece was found. However, there was a main within-subjects effect of Time-Point, $F(2.29, 54.86) = 9.22$, $p < .001$. To identify approximately the point at which the mean current-to-final difference score ceased to be significantly different between consecutive time points, simple contrasts were calculated in which each of the first seven levels of the within-subjects variable were compared in turn with the eighth level. The contrasts revealed that the difference was significant at $p < 0.05$ for the first three comparisons and close to significance for the fourth ($p = .051$), but did not approach significance thereafter. Thus, the majority of listeners had settled at, or very close to, their final judgment by 60 s into the performance.

However, in addition to the within-subjects variable of time-point, a significant between-subjects effect of Group was also found, $F(2, 24) = 6.14$, $p < .01$. Post hoc Bonferroni comparisons suggested that, as might be expected, Group *T* differed significantly from both of the other groups at the .05 level. Figure 3 shows the mean pattern of evaluations for those trials in which some movement took place subsequent to the initial decision, by group and averaged across both pieces.

All three groups displayed different patterns of rating over time. Group *Q* tended to make initial ratings

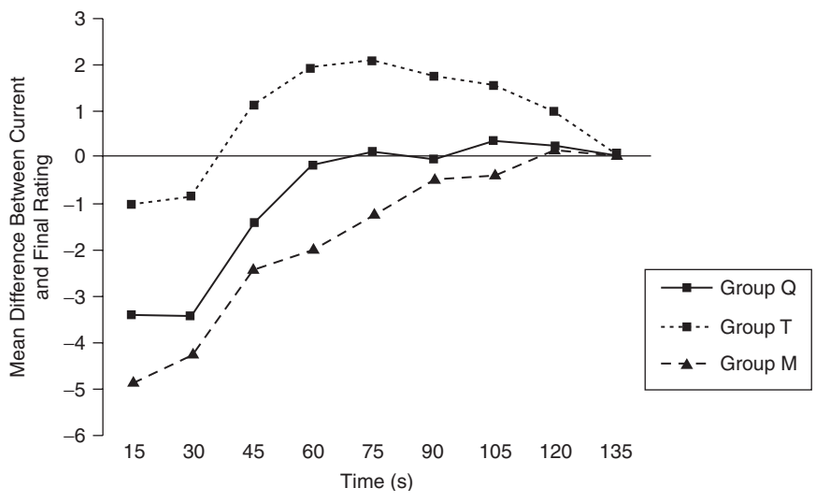


FIGURE 3. Mean initial-final difference score at 15 s intervals, across pieces.

somewhat below the level of their final rating. However, they reached their final opinion comparatively quickly – by approximately 60 s—and made relatively few large adjustments thereafter. Group *M* displayed a similar pattern to Group *Q* in that they began with initial ratings some way below their final rating. However, they took slightly longer to reach a final decision, arriving within 1 unit at around the 90 s mark.

Participants in Group *T* tended to revise their judgments *above* the level of their final rating during the first minute, before correcting them downwards to approximately their initial level just before the end of the piece. This is curious—it might have been expected, on the basis of analyses presented above, that mean ratings from this group would remain largely static over time since they showed no significant difference between initial and final ratings.

#### MAGNITUDE OF MOVEMENTS

It might be expected that if participants began to finalize their ratings some way before the end of the performance, the average magnitude of movements would decrease at about the same point. To examine this, the average magnitude of movements in each 15-s section was calculated for each trial. These values were then averaged across pieces, per participant (note that these values were also corrected to account for variation in Start Time between trials, such that sections of trials in which the initial rating had not been recorded were left blank). Again, a repeated measures ANOVA was calculated with Piece and Section as within-subjects variables and Group as between-subjects. No main effect of Group or Piece was found, and there were no significant interactions. There was a main within-subject effect of Section, $F(4.79, 143.70) = 21.36$, $p < .001$, with repeated contrasts suggesting a significant difference between sections 6 and 7, $F(1, 30) = 12.67$, $p < .005$.

However, Figure 4 shows mean magnitude per movement plotted against time section for each group, across pieces. From the graph, it seems that the point at which magnitude per movement began to decrease was actually different between the groups, with *Q* and *M* showing a marked decline between sections 5 and 6. Since the pattern of magnitude per movement was similar for all three groups up until that point, and then also in the final two 15-s sections, it seems likely that this group difference was not reflected in a main between-subjects effect in the ANOVA.

Repeated measures ANOVAs were thus calculated separately for each group, with repeated contrasts performed between consecutive sections. For Group Q the main effect of Section was significant, $F(7, 70) = 11.09$, $p < .001$, and the contrasts revealed a significant decline that began between sections 5 and 6, $F(1, 10) = 5.58$, $p < .05$. For Group M the pattern was the same, with a significant main effect of Section, $F(7, 70) = 7.46$, $p < .001$, and a significant decline beginning between sections 5 and 6, $F(1, 10) = 7.58$, $p < .05$. For Group T, however, the pattern was different; whilst the main effect of section was significant, $F(7, 70) = 6.26$, $p < .001$, the decrease came later, between sections 6 and 7, $F(1, 10) = 9.95$, $p < .05$.

Taken in tandem with the analysis of rating change over time, these results suggest that participants in Group *T* not only showed a different pattern of rating change over time from the other groups, but took longer to *finalize* their rating than did those in Groups *Q* and *M*.

### *Discussion: Relationships between Rating and Process*

It seems clear from the pattern of results that listeners in different groups showed somewhat different patterns of rating. Those in Group *Q* tended to revise their judgments upward over time, reaching their "final" evaluation by around 60 s into the performance—about halfway through. Participants in Group *M* also revised their judgments upward, and by approximately the same amount, although taking slightly longer to settle at their final decision. The mean pattern of ratings from Group *T*, however, was both unusual and unexpected. While the typical initial-to-final difference was minimal, the mean pattern of evaluation over time actually went upward, falling down to the final value only towards the end of the performance.

How can this result be explained? First, it should be recalled that all of the performances were technically "solid," with any obvious wrong notes or other slips edited out during the preparation stage. It may be that performances with obvious technical problems would have elicited different responses from listeners in Group *T*. With judicious editing, it would be possible to prepare a set of performances that were identical, but for some including "wrong notes" strategically inserted at points in the performance. A continuous response methodology could then be used to ascertain if perceived technical quality was modified as a direct result of these apparent errors.

In the present study, however, there were no obvious wrong notes or technical slips in any of the performances. Moreover, whilst neither of the works could be described as "easy" to play well, neither of them was especially virtuosic (this was reflected in the ratings of perceived difficulty of the music for the performer to play well: Bach = 3.9, Chopin = 3.0). Given the concomitant

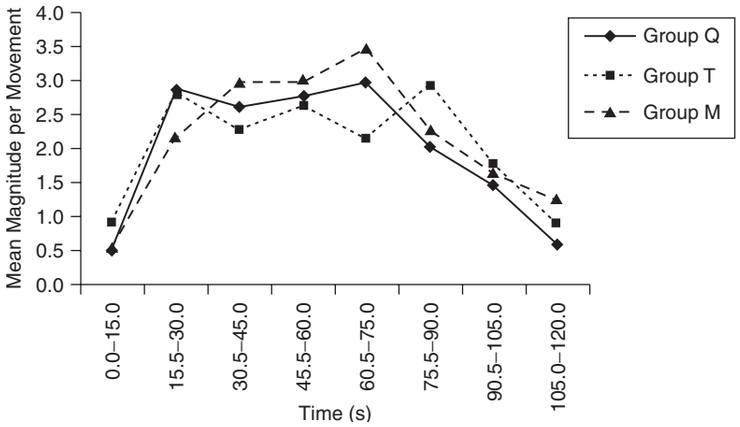


**FIGURE 4.** Mean magnitude per movement plotted against time segment for each group, across pieces.

lack of—as it were—"hard evidence" of technical proficiency and assurance, it might be understandable had participants in Group *T* simply waited to hear as much of each performance as they could before making an initial evaluation, and then not adjusted it thereafter. However, overall they neither made their initial evaluations later nor modified their evaluation less throughout the performance than either of the other groups. Rather, they modified their evaluation differently.

It is not uncommon to hear people describe performances that while not entirely secure technically were otherwise arresting, such that their technical deficiencies seemed unimportant. Ordinarily, it may be that listeners concentrate on the broadly "musical" aspects of a performance unless there is something so significantly awry with the technical aspects that it actively interferes. It could be that in the present study, where there were no such technical problems, the task of evaluating just technical aspects of the performance was so difficult—because listeners had so little information on which to base their evaluation—as to be unfeasible. Participants may have simply been unable to complete the task as required, the lack of information about technical features of performance leading them to be unconsciously influenced by other aspects. However, this would imply that listeners were only sensitive to obvious technical problems rather than more subtle cues. In any case, such an explanation would not account for the pattern of evaluation observed in Group *T*. This is difficult to justify on the basis of present knowledge, and more work is certainly required.

Overall, this analysis does not answer the question, posed earlier, as to whether written summary evaluations evolve over time or are attributable to a recency effect—as long as the Final and Written ratings are shown not to differ significantly, a recency effect cannot be ruled out. However, participants in Groups *Q* and *M* appeared to reach their final evaluation some time before the end of the piece; this was most apparent for the former group, but true of both. In both cases, while the sheer *rate* of movement after this point did not immediately decrease, the average magnitude of individual movements did, suggesting subsequent fine adjustments rather than significant changes of opinion. This may be evidence that listeners "finalized" their evaluations after receiving a certain amount of information about the performance (i.e., hearing 60-90 s).

One interesting implication of this hypothesis is that there may be a temporal point after which a listener's evaluative judgment becomes relatively fixed and inflexible. The general idea that opinions may become fixed such that subsequent information is interpreted in light of them has good face validity in the context of performance evaluation. For example, a performance that is assured and confident for the first few minutes may give the impression to the listener that the player has a sound technique and good musical understanding. Subsequent information, such as technical slips or instances of poor phrasing, may thus be put down to other factors—perhaps tiring towards the end of the piece, or being distracted by noise from the audience—without changing the basic belief that the player is good and competent.

An important question for future research is whether the time required to finalize an evaluative decision

varies as a function of the length of the piece. There seems to be no reason why time to make an *initial* evaluation should increase for longer pieces, but it may be that the finalization of judgments takes longer (at least if listeners are aware of the length of the piece). Data from Vasil (1973) suggests that the timeframe for final judgments may remain within 90 s for pieces that are up to six minutes in length. However, no other studies have explored this issue in the context of musical performance evaluation. Needless to say, a continuous measurement procedure would provide a suitable method for doing so.

## General Discussion

Returning to the four main research questions posed at the beginning of the paper, results of the study can be summarized as follows:

1. *How long do listeners take to reach an initial evaluative judgment?*
   On the basis of the present data, not long—of the order of just 15 s. Moreover, by around 60 s into the performance, listeners (at least those listening more "holistically," i.e., in Group Q) had typically reached their "final" evaluation, only making small adjustments thereafter.
2. *How, and how frequently does this judgment vary during the course of a performance?*
   Listeners' evaluations of performance were subject to change at an average of approximately 2.6 times per minute. Moreover, the rate of movements tails off only towards the end of the piece. The mean *magnitude* of each movement, however, tended to decrease appreciably from around 60 s into the performance, again suggesting small modifications rather than large changes of opinion.
3. *How do written summary evaluations of the performance relate to the continuous evaluation?*
   Written summary evaluations were not significantly different from final judgments, but were somewhat different from initial judgments. Coupled with the evidence that evaluations became relatively fixed at around 60-90 s, this probably reflects an *evolving* process of preference formation rather than a recency effect or a "mean over time" of evaluative opinions. However, more research is required to verify this absolutely.
4. *Do perceptions of the quality of different aspects of a performance show different patterns of variation over time?*
   To some extent, yes. While there was no evidence that the time taken to reach an initial evaluation, or the number of subsequent revisions of that evaluation, differed between groups, the pattern of ratings over time did show differences. Groups *Q* and *M* tended to revise their opinions upwards within the first half of the performance before settling at their final evaluation. Group *T*, on the other hand, began close to their final evaluation but tended to revise their opinions upwards and then down again. This pattern of change over time is difficult to account for on the basis of existing knowledge. Crucially, however, this result invites the conclusion that listeners *are*, after all, able to distinguish between aspects of performance. The phenomenon of limited between-category discrimination thus appears to result from the actual process of *summarizing*, rather than being caused simply by a deficit in listeners' powers of discrimination.

In addition:

5. Individual consistency in dynamic characteristics (i.e., the listeners' *modus operandi* in making evaluative judgments) was surprisingly low.

Is this large variability in the evaluation process itself really that unexpected? Other than the fact that it is widely *assumed* that experienced listeners can make evaluative judgments consistently and reliably, there is little evidence to suggest that this is the case; indeed, previous research has suggested that evaluations of performances may be significantly less reliable and consistent than generally hoped. The present study adds weight to this argument by suggesting that amongst a sample of experienced musicians and music listeners there is considerable variability not just in the absolute ratings awarded, but in the actual mechanisms through which these ratings are produced. Moreover, this variability is evident both inter- and intra-rater.

It could be that such variability is a function of the continuous measurement task itself, which was unfamiliar and, in some respects, artificial. However, in the debriefing that took place after each experimental session, participants were asked how they had found the task, and specifically whether they had felt able to give a true reflection of their evaluative responses to the performances using the continuous measurement interface. To these inquiries, no participants noted any difficulty with the task and the large majority reported that they had been able to give an accurate picture of their judgment.

In considering the results of the study, certain caveats and limitations must be acknowledged. First, it could be claimed that the experimental stimuli were not highly ecologically valid. Being complete performances of

whole pieces, they were arguably more realistic than in some previous studies such as those of Fiske (1975, 1977), in which only short fragments of performances were used. At the same time, however, they were also chosen to be moderately, rather than highly contrasting. This broad homogeneity of general attributes was intentional, since the aim of the study was to identify overall characteristics of the evaluation process unhindered by wide variations in the actual performances themselves. Had the performances been very different, it would have been impossible to establish whether any consistent pattern emerged in the way in which participants tackled the evaluation task. Clearly, it will be important for future work to establish whether time-dependent aspects of the evaluation procedure vary with different types of pieces. It could be that, for example, evaluative judgments will be finalized more quickly for faster tempo pieces than slower pieces, since a greater amount of performance information is received per unit of time. However, much more research is required before anything more than speculative hypotheses can be proposed.

A second issue may be that the requirement made of participants in Groups *T* and *M*—i.e., to listen to the performances whilst explicitly concentrating on just one aspect—was too artificial. While isolating aspects of performance is required when making evaluations using segmented schemes, it is unusual to do so for an entire performance. It is perhaps feasible that the process of attempting this led participants to give responses that were not genuinely representative of their ordinary evaluation processes. This kind of concern is endemic to experimental designs that attempt to separate out variables normally enmeshed within complex behaviors and as such is perhaps unavoidable.

Some basic procedural concerns might be raised about the rather sterile location of the experimental sessions (a laboratory space) and participants' motivation for listening to the music and making the evaluations. These are no more than the familiar problems that beset a great deal of work in experimental psychology. Interestingly, much the same observation has been noted in the literature on consumer decision making (Moore, 1999).

Perhaps the main limitation of the study is the extent to which the results are generalizable. It would be tempting to try and extrapolate from them that initial evaluative judgments are always made in the region of 15-20 s into any performance, or that perceptions of technical quality will always exhibit a different pattern of change over time from perceptions of musical quality. However, in the present study, the performances were audio-only and participants had no extra-musical information on which to make their judgments. In a real performance setting, by contrast, a host of other factors would probably play a part in determining a listener's evaluative responses. It has been shown, for instance, that the performer's physical appearance (Wapnick, Mazza, & Darrow, 1998, 2000) race and gender (Davidson & Edgar, 2003; Elliott 1995/6), and self-efficacy (McCormick & McPherson, 2003) can all affect the final rating awarded in an evaluation. Outside a laboratory, this information would be available to listeners. Crucially, moreover, it may be available some time *before* the player begins performing. If anything, then, it could be hypothesized that in a real concert situation an evaluative judgment of some kind might be made even earlier than the typical 14-20 s suggested by the present data. It may even be made, perhaps unconsciously and on a provisional basis, before the music starts.

## Author Note

This work was supported in part by a grant from The Leverhulme Trust.

*Correspondence concerning this article should be addressed* to Dr. Aaron Williamon, Centre for Performance Science, Royal College of Music, Prince Consort Road, London SW7 2BS, UK; E-MAIL: awilliamon@ rcm.ac.uk

## References

BUCKLEY, M. R., & EDER, R. W. (1988). B.M. Springbett and the notion of "snap decision" in the interview. *Journal of Management, 14*, 59-67.

DAVIDSON, J. W., & EDGAR, R. (2003). Gender and race bias in the judgment of Western art music performance. *Music Education Research, 5*, 169-181.

ELLIOTT, C. A. (1995/6). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education, 127*, 50-56.

FENECH, M. (2003). Continuous Measurement Software (Version 1.0). Unpublished computer software, Royal Holloway, University of London.

FISKE, H. E. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education, 23*, 186-196.

FISKE, H. E. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education, 25*, 256-263.

GOTTMAN, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge: Cambridge University Press.

HERR, P. M. (1989). Priming price: Prior knowledge and context effects. *Journal of Consumer Research, 16*, 67-75.

MCCORMICK, J., & MCPHERSON, G. (2003). The role of self-efficacy in a musical performance: An exploratory structural equation analysis. *Psychology of Music, 31*, 37-51.

MADSEN, C. K. (1997). Emotional response to music as measured by the two-dimensional CRDI. *Journal of Music Therapy, 34*, 187-199.

MADSEN, C. K. (1998). Emotion versus tension in Haydn's Symphony No. 104 as measured by the two-dimensional continuous response digital interface. *Journal of Research in Music Education, 46*, 546-554.

MAZZONI, D. (2004). Audacity (Version 1.2.2) [Computer software]. Retrieved from http://audacity.sourceforge.net

MEYER, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.

MOORE, D. A. (1999). Order effects and preference judgments: Evidence for context dependence in the generation of preferences. *Organizational Behavior and Human Decision Processes, 78*, 146-165.

PLETZER, K. (2002). MIDI Maestro 2 (Version 2.32) [Computer software]. Retrieved from http://www.midimaestro.com.

SCHUBERT, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion space. *Australian Journal of Psychology, 51*, 154-165.

SCHUBERT, E. (2001). Continuous measurement of self-reported emotional response to music. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research*. Oxford: Oxford University Press.

SCHUBERT, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception, 21*, 561-585.

SCHUBERT, E., & DUNSMUIR, W. (1999). Regression modeling continuous data in music psychology. In S. W. Yi (Ed.), *Music, mind, and science*. Seoul, Korea: Seoul National University.

SLOBODA, J. A., & LEHMANN, A. C. (2001). Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude. *Music Perception, 19*, 87-120.

SPRINGBETT, B. M. (1958). Factors affecting the final decision in the employment interview. *Canadian Journal of Psychology, 12*, 13-22.

THOMPSON, S., & WILLIAMON, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception, 21*, 21-41.

TULLAR, W. L., MULLINS, T. W., & CALDWELL, S. A. (1979). Effects of interview length and applicant quality on interview decision time. *Journal of Applied Psychology, 64*, 669-674.

VASIL, T. (1973). The effects of systematically varying selected factors on music performance adjudication. Unpublished doctoral dissertation, University of Connecticut.

WAPNICK, J., MAZZA, J. K., & DARROW, A. A. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance evaluation. *Journal of Research in Music Education, 46*, 510-521.

WAPNICK, J., MAZZA, J. K., & DARROW, A. A. (2000). Effects of performer attractiveness, stage behavior, and dress on evaluation of children's piano performances. *Journal of Research in Music Education, 48*, 323-335.

YBARRA, O. (2001). When first impressions don't last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition, 19*, 491-520.

YBARRA, O. (2002). Naive causal understanding of valenced behaviors and its implication for social information processing. *Psychological Bulletin, 128*, 421-441.