Making an Impression: Error Location and Repertoire Features Affect Performance Quality Rating Processes

George Waddell, Rosie Perkins, & Aaron Williamon
Royal College of Music, London, United Kingdom

This article examines the effects of composition length, familiarity, and likeability—as well as the location of performance errors—on the process of forming performance quality ratings. Five piano works by Chopin and a twentieth-century composer were chosen to vary by length and familiarity. Three of these pieces were then manipulated to contain performance errors in the opening material, and two of those the same error at the recapitulation. Forty-two musicians provided continuous quality evaluations and final quality ratings of the performances, hearing one version of each piece. The results showed that familiarity had no effect within works of a well-known composer, but times to first and final decision were significantly extended for an unfamiliar work of an unfamiliar composer. A shorter piece led to a shorter time to first decision. An error at the beginning of a performance caused a shorter time to first decision and lower initial and final ratings, where the same error at the recapitulation did not have a significant effect on the final judgment, despite causing a temporary negative drop. These findings demonstrate how evaluators' knowledge of a work can affect their rating process and the importance of making a strong first impression in performance.

Received June 29, 2017, accepted March 12, 2018.

Key words: performance, evaluation, continuous measures, errors, familiarity

A large body of research has examined the effects of musical and extra-musical factors on performance quality judgments (for reviews, see McPherson & Schubert, 2004, and Waddell & Williamon, 2017a). Many of these investigations have focused on aspects of the performer (Platz & Kopiez, 2012; Tsay, 2013, 2014) including their dress (Griffiths, 2008, 2010, 2011), race (Davidson & Edgar, 2003; Elliott, 1995; VanWeelden, 2004), and attractiveness (Wapnick, Darrow, Kovacs, & Dalrymple, 1997; Wapnick, Mazza, & Darrow, 1998, 2000; Ryan & Costa-Giomi, 2004; Ryan, Wapnick, Lacaille, & Darrow, 2006). The role of evaluators has also been scrutinized in terms of their reliability (Wapnick et al., 2005; Wesolowski, Wind, & Engelhard, 2015), their knowledge of the performer (Duerksen, 1972; Kroger & Margulis, 2017), and the validity of the rubrics used (Thompson & Williamon, 2003; Wesolowski, 2017). Recent research is beginning to examine the temporal processes by which these assessments are formed (Himonides, 2011; Thompson, Williamon, & Valentine, 2007; Waddell & Williamon, 2017b) and the points in time when decisions are made, changed, and concretized. The present study examined this process by experimentally manipulating four representative factors hypothesized to interact with the temporal process of quality judgement: the general length, familiarity, and likeability of presented works, and the location of performance errors.

Length is a fundamental feature of any composition and provides an easily quantifiable differentiator—a work may be said to be two, five, or one hundred times longer than another—while mode, expressivity, or difficulty, for example, are not so easily quantified. Of course, the tempo of an interpretation may alter the length of a performance, but the repertoire itself determines the baseline. The length of an excerpt used in research settings is often not varied or questioned, although studies by Wapnick and colleagues found that excerpts of differing lengths were rated differently. In a first study (2005), two groups of participants rated recordings of 19 classical music excerpts 20 or 60 s long. Length condition was randomized within each group and counterbalanced between groups. Participants were not informed in advance of the length of each recording. The results showed that the longer excerpts received significantly higher and more consistent ratings, measured as differences in group standard deviations. In a subsequent study by Wapnick, Campbell, Siddell-Strebel, and Darrow (2009), participants were given 25-, 55-, or 115-s excerpts of performances, again rating the longer two excerpts significantly higher than the 25-s excerpt. The researchers varied certain visual characteristics, finding that attractiveness, sex, dress, and stage behavior produced conflicting effects for different

Music Perception, volume 36, issue 1, pp. 60–76, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2018.36.1.60

lengths of excerpt, such as dress affecting men's ratings of the 25- and 115-s but not the 55-s excerpts. Overall, these studies highlight variation in the evaluation process depending on the length of the task, although they did not examine these effects with full-length compositions or in situations where participants knew in advance the length of time they had to make their decisions.

Outside the musical domain, research has suggested that the total time to determine an applicant's suitability in interviews is mediated by the predetermined length of the interview (Buckley & Eder, 1988). One such study demonstrated that participants viewing video-recorded interviews of approximately 15 minutes took longer to reach a final decision if they were first informed that the session would take 30 minutes (Tullar, Mullins, & Caldwell, 1979). Crucial to the method was the participant's knowledge (though faulty) of the interview length prior to its beginning. Thus, in a musical context, the length of the excerpt would need to be explicitly stated before its presentation for an accurate comparison; in a range of settings, from listening live in concert and examinations to listening to recordings, it is not uncommon for timing information to be available to the listener.

Familiarity with the work takes into consideration the knowledge of the evaluator. Indeed, the very definition of a musical expert in evaluation settings usually includes knowledge of common repertoire or experiences of engaging with new repertoire. Such a connection makes sense: evaluators who are aware of the framework on which the interpretation is to be made are, in theory, primed for the information that is to be presented to them and have a standard to which they can compare variances in individual interpretations. In line with this, Kinney (2009) found that evaluators' familiarity with a work improved their internal consistency when forming quality judgments of performances of that work. In terms of the temporal aspects of decision making, one could hypothesize that familiarity with a work would decrease the time to the first and final judgment, as less effort would be needed to understand and process the nature of the work being presented and thus the attention could be shifted to the quality of the performance itself. However, another advantage of familiarity could be increased awareness of the structure of the work, including perhaps the points at which the most technically challenging and musically defining moments will take place. One could then hypothesize that familiarly would *increase* the time to a final judgment, as evaluators would delay their decision until the expected points of interest arrived. This would be specific to the work and the points which the particular evaluator considered of interest. Such hypotheses have not yet been investigated, but a continuous measures methodology would allow for the relationship to be examined directly.

Related to familiarity is the concept of the *likeability* of a work—that is, does the evaluator have an inherent preference for the composition itself? While every listener (and evaluator) is entitled to such preferences, it would be problematic if they were to interfere with the evaluative process if it were taking place in educational or competition contexts. Research specifically examining the relationship between performance quality rating and preference for the work is lacking. It is generally assumed that one's preference for a work is tied closely to one's familiarity with it, although Thompson (2007) found that the two concepts could be separated to some degree in that likeability, but not familiarity, of a work was predictive of enjoyment. The same study also found that performance quality could be separated from affective response, suggesting that the evaluative process may be unchanged despite differences in preference for a work, but such assumptions have not been experimentally tested.

Regarding *performance errors*, previous studies have examined the ability of musicians of varying experience to detect manipulated "mistakes" in recordings. Byo (1993) asked participants to detect errors in recorded excerpts of polyphonic wind band repertoire, manipulated to contain performance errors. Analyses found that listeners were better able to identify rhythmic than pitch errors and improved in identifying both when the instrument timbres were similar across voices. A later study (Byo, 1997) supported these findings, also demonstrating that experience and rating monophonic (versus polyphonic) textures increased error detection rates. Repp (1996) found that listeners detected only 38% of pianists' pitch-based errors, including missing or unnecessarily repeated notes. The nature of performers' errors has also been examined, with research demonstrating that: 1) errors are more likely to be made in the middle of phrases away from structural boundaries (Mishra, 2010); 2) the majority of pitch errors in a corpus of Chopin recordings were note omissions, with a significant proportion of errors systematically repeated (Flossmann & Widmer, 2011); 3) performers can detect that they are about to perform an error immediately before the motion is executed via electroencephalographic (EEG) event-related potentials (Maidhof, Pitkaniemi, & Tervaniemi, 2013; Maidhof, Rieger, Prinz, & Koelsch, 2009; Ruiz, Jabusch, & Altenmüller, 2009); and 4) EEG negative potentials immediately following the perception of an error are more pronounced when performing than when listening (Maidhof, Vavatzanidis, Prinz, Rieger, & Koelsch, 2010).

Only one study has, to date, examined the effects of errors on temporal quality judgements (Waddell & Williamon, 2017b). In that project we manipulated a video-recorded performance of a Chopin Etude to include a significant performance error and/or a negative facial reaction to that error. Participants were grouped as musicians and nonmusicians and were asked to rate one of four permutations: no error, aural error only, aural error and facial reaction, and facial reaction only. When the error was accompanied by a facial reaction, participants gave a significantly lower overall rating regardless of their musical experience. However, when the performer did not visibly react to the error, only the musicians showed an immediate reaction in their continuous quality judgments, and this drop recovered to the height of the error-free performance by the end. Our study demonstrated the capacity for a continuous measures methodology to isolate and examine real-time reactions to a performance error and how they are mediated by extraneous variables. A question remains as to whether a mistake at the beginning of a piece is more harmful to one's evaluation than one at the end. Research in interpersonal impression formation would suggest so, as negative first impressions have been found harder to alter than positive ones (Ybarra, 2001), yet this is still to be examined in the context of music performance evaluation.

Previous studies have demonstrated that participants form their initial quality judgements within an average of 15 s when rating audio (Thompson et al., 2007) or audiovisual (Waddell & Williamon, 2017b) recordings of standard repertoire, with no correlation found between time to first decision and overall quality rating. Thus, hypothesized increases or decreases in decision time resulting from differences in features of the works themselves were posited based on the existing literature:

(1) Works of lesser familiarity would result in a longer time to first and final decision. This effect would be increased in the case of a work of unfamiliar tonal framework and composer. The direction of the effect of likeability was not hypothesized.
(2) A work of shorter length (when work length is known beforehand) would result in a shorter time to first decision.

Regarding the performance errors, two hypotheses were posited:

(3) A performance error inserted at the beginning of a composition would result in a shorter time to first decision.
(4) A performance error inserted at the beginning of a composition would result in a lower final rating than the same error inserted part way through the performance.

To test these hypotheses, works of varying length and familiarity were chosen. In addition, a difference in genre (i.e., Romantic versus twentieth-century) and popularity of composer (famous versus relatively unknown) was used to emphasize the familiarity contrast in one of the five chosen works. Performance errors were added digitally to several of the performances, with every effort made to create the impression of live, undoctored recordings.

Method

PARTICIPANTS

Forty-two musicians were recruited via email and in person from the Royal College of Music (RCM) and Imperial College London. The cohort comprised 24 women and 18 men with a mean age of 27.2 years ($SD \pm 9.9$, range = 18–55). Musical experience among the group varied, ranging from undergraduate to doctoral students and including 4 professional musicians, with a mean 19.9 years of musical experience ($SD \pm 9.6$, range = 5–51). Fifteen participants reported the piano as their primary instrument, and of the remaining 27 (12 strings, 8 winds, 4 voice, 1 brass, 1 organ, 1 harp), 20 reported the piano as a second study instrument. Informed written consent was obtained from all participants following the ethical guidelines of the British Psychological Society and with internal RCM approval on behalf of the Conservatoires UK Research Ethics Committee. No payment was given in exchange for participation.

STIMULI

Repertoire was chosen to vary in length, familiarity, and genre. The piano works of Frédéric Chopin were selected as they provided a wide range of compositions with a distinct, overarching style by a well-known composer and including compositions of less than one minute in length. Four of Chopin's works were chosen: 1) the "Black Key" Etude in G♭ Major, Op. 10, No. 5; 2) the "Minute" Waltz in D♭ Major, Op. 62, No. 1; 3) the Prelude in D Major, Op. 28, No. 5; and 4) the Tarantelle in A♭ Major, Op. 43. These were selected to match in mode (major key), tempo (fast: 100–150 beats per minute), and texture, with a scalar and arpeggiated right hand over accompanying figures in the left. Of these, the Etude, Waltz, and Tarantelle were chosen as longer pieces (> 100 s) and the Prelude as a short piece (< 30 s). They were also chosen to vary in familiarity, ranging

TABLE 1. *Works Used as Stimuli for the Study*

Composer	*Title*	*Length (s)*	*Tempo (bpm)*
Chopin	Etude in G♭ Major, Op. 10, No. 5 "Black Key" Etude	108	∼110
Chopin	Waltz in D♭ Major, Op. 62, No. 1 "Minute" Waltz	117	∼100
Chopin	Prelude in D Major, Op. 28, No. 5	27	∼100
Chopin	Tarantelle in A♭ Major, Op. 43	156	∼150
Eckhardt-Gramatté	Caprice No. 6 "Klavierstuck" (excerpt)	152	∼130

from very popular with the Etude and Waltz to relatively unknown (as much as is possible with a work of Chopin) with the Tarantelle. To create a stark familiarity contrast, the Caprice No. 6 "Klavierstuck" by twentieth-century composer Sophie Carmen Eckhardt-Gramatté was chosen. The work bears technical similarities to the selected Chopin works in its use of melodic material in the right hand over accompaniment in the left but employs an expanded, less familiar tonal framework. As performances of shorter complete works of this nature were not available, an excerpt taken from the beginning to a point that could be perceived as a functional finale was used to match the length of the Tarantelle, the most unfamiliar Chopin work. The selected compositions, their lengths (in terms of the performance used), and their approximate tempi are shown in Table 1. Piloting was undertaken via informal discussions with undergraduate- and graduate-level pianists to confirm that assumptions made concerning familiarity and the choice of endpoint of the twentieth-century piece were valid.

MIDI recordings of the Chopin works were used to allow for the controlled insertion of performance errors at strategic points. These recordings were taken from *The Classical MIDI Resource*, an online repository of openly submitted MIDI recordings that are editor-reviewed for quality and accuracy before being posted for free download. The Eckhardt-Gramatté Caprice was recorded acoustically by a graduate-level pianist, itself requiring no manipulation as it was not a part of the error examination due to its lack of recapitulating material. To ensure that the artificially inserted errors would be both believable and easily perceived, dissonant errors of pitch in a single voice were chosen as they have been shown to be both the most common in piano performance (Flossmann & Widmer, 2011) and the second most easily perceived, after rhythmic errors (Byo, 1993, 1997). To test the effect of error location, the two familiar Chopin works of the same length (the Etude and Waltz) had also been selected due to the recapitulation of their opening thematic material. Thus, a pitch error in the opening seconds of the performance could be recreated midway through, differing only in temporal location and structural context. To match error type as closely as possible, Logic Pro 9 was used to transpose an arpeggiated figure in the right hand of approximately one bar in length up one semitone in each work, simulating a pianist that had played a brief passage with the hand in the wrong chord position. Three tracks were then created for each of the two works: one with an error at the beginning (error-start), one with an error at the recapitulation (error-recap), and one control condition without an error (no error). An error of the same nature was added to the beginning of the Prelude to test the interaction of opening error and work length. A summary of the variables associated with each work is provided in Figure 1.

Although MIDI files of piano recordings have been successfully employed in previous studies of music performance evaluation (e.g., Kinney, 2009; Sloboda & Lehman, 2001; Thompson et al., 2007; Winter, 1993), digital enhancement was undertaken to add further realism to the files and to match the recording quality of the contemporary excerpt. Specifically, Logic Pro 9 was used both to realize the MIDI data into audio formats and to add three effects: 1) reverb, to emulate the acoustic of a performance space in a live recording; 2) stereo split, to break the mono output of the MIDI realization into slightly varying signals as one would experience in a true stereo recording; and 3) distortion, applied sparingly to approximate the signal loss inherent to audio recording and dull the overly bright and harsh quality often associated with MIDI recordings. Manipulations of the MIDI data also allowed for the removal of overt performance eccentricities (e.g., occasional over accented, jarring notes or the addition of slight tempo fluctuation in overly metronomic passages, a common characteristic of MIDI recordings). The tracks were then converted to .WAV format, and 4 s of silence were added to the beginning of each to allow for the listeners to prepare themselves after commencing each trial. Informal piloting with graduate-level pianists confirmed that the performances could pass for genuine acoustic recordings. Audio recordings of the 10 experimental stimuli are available as Supplementary Material in the online version of this paper.

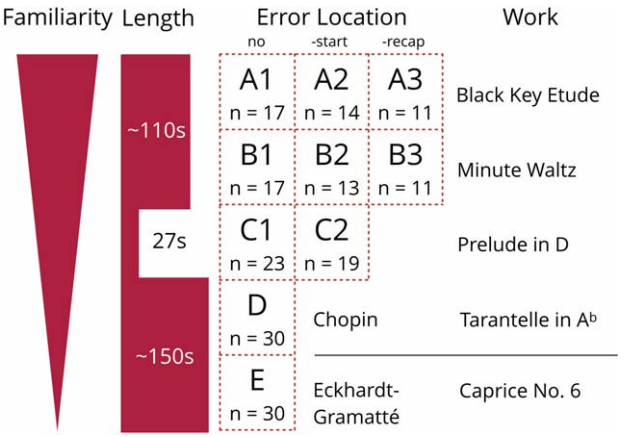

FIGURE 1. The study design, showing the variables of repertoire length, relative familiarity, and error placement in the five works. Following a repeated-measures approach, the number of participants assigned to each condition are shown (total $N = 42$; see also Procedure). Audio recordings of the 10 experimental stimuli are available as Supplementary Material in the online version of this paper.

CONTINUOUS MEASURES

Continuous measures methodologies, where participants provide real-time feedback to a stimulus where temporal location of feedback is valuable, have been used in a wide variety of musical contexts (see Geringer, Madsen, & Gregory, 2004, for a review). The majority of these studies have examined elements of musical expressiveness, arousal, and attention, although several have explicitly examined performance quality ratings (Himonides, 2011; Thompson et al., 2007; Waddell & Williamon, 2017b). For the present study, the continuous measures data were collected using software developed at the Royal College of Music and used by Thompson and colleagues (2007). The software comprised a horizontal blue bar onto which the participant moved their mouse pointer when they were ready to register their first judgment and then along which they could move the pointer to increase or decrease continuously their rating as appropriate. The horizontal area was divided into 70 discrete sections, not visible to the participant, while a 7-point scale (1 = "poor" to 7 = "excellent") was overlaid above the rating area for easy transfer to the written evaluations (see below). Data points were sampled at 2 Hz. The software was presented to each participant on the same Windows-based laptop with USB mouse and Sennheiser HD 380pro headphones.

WRITTEN EVALUATIONS

Two bespoke questionnaires were used in the study. The first was completed immediately following each trial and assessed the participants' relation to the work and overall evaluation of the performance on 7-point Likert-type scales along several categories: overall quality of the performance (1 = "poor" to 7 = "excellent"), familiarity with the work (1 = "never heard it" to 7 = "extremely familiar"), and degree to which they like the composition (1 = "not at all" to 7 = "very much"). The typicality of the performance in relation to others they have heard (if applicable) and the perceived difficulty of the work to perform was also measured on 7-point scales. Participants were encouraged to provide comments concerning each performance. The second questionnaire, completed at the end of the study, elicited background information on music training and listening preference by musical genre: Baroque, Classical, Romantic, and twentieth-century, each measured on a 7-point scale. The questionnaires can be downloaded in the Supplementary Material that accompany the online version of this paper.

PROCEDURE

Participants met the researcher in a quiet room at the Royal College of Music or Imperial College London and were presented with an information sheet and consent form. They were then introduced to the continuous measures software and encouraged to make and record their decisions as instinctively and intuitively as possible, emphasizing that their decisions should be made not on the basis of their enjoyment of the performance but rather on the objective quality of the performance "as though they were a competition judge." A brief (< 20 s) excerpt of a Beethoven piano sonata was used as a test piece, which the participants were allowed to repeat as many times as they wished until they felt comfortable with the input method. Following this, participants were told that they were about to hear several live performances by different undergraduate pianists—as opposed to studio, professional recordings so that the obvious performance errors would not seem implausible—and to rate the performance quality. For each trial, the name of the composer, the name of the work, and the length of the recording was presented orally to the participant. They were then able to start the first recording in their own time, and when it finished, they completed the first questionnaire. This procedure was repeated for each work in a randomized order with a questionnaire following each continuous measurement. Concerning the performance errors, participants randomly heard either the no error, error-start, or error-recap condition of the Etude and Waltz and either the no error or error-start condition of the Prelude; separate randomization procedures were used for each work.

The randomization was established to favor conditions with no error to maximize opportunities to compare performances without such manipulations across the five works. Following the final trial the second questionnaire was presented, and participants were invited to give comments concerning the procedure as a whole. Each session lasted 30–40 minutes.

Due to time constraints, 12 of the 42 participants were presented only the three works containing variations in errors (i.e., the Etude, Waltz, and Prelude) following the same randomization procedures described previously. These pieces were emphasized to maximize opportunity for between-groups examination of error placement, as the other 30 participants had rated the Tarantelle and Caprice but only 10 would have rated the no error, error-start, or error-recap versions of the other three works. The final *n* values for each condition are shown in Figure 1. These shorter sessions lasted approximately 20 min.

DATA TREATMENT AND ANALYSIS

Data were treated to several operations, primarily following Thompson et al. (2007) and Waddell and Williamon (2017b), in which three discrete variables were extracted from the full continuous data, along with the quality rating provided in the written comments:

(1) Time to first decision, $\mathbf{T_1}$: As a brief amount of time was necessary to move the mouse to the desired first rating point, the moment the cursor entered the horizontal bar and data collection began was noted as the initial decision time, T_1. The continuous measurement ratings were measured from the moment the trial was started, yet the first note was not played until 4 s; therefore, 4 s were subtracted from each score, giving initial ratings made prior to the first note a negative time value.
(2) First rating, $\mathbf{R_1}$: The first point at which the participant maintained a stable rating of at least 2 s was taken as the first rating.
(3) Final rating, $\mathbf{R_2}$: The final continuous score reported formed the final rating.
(4) Overall rating, $\mathbf{R_3}$: The overall written score provided in the questionnaire on a scale of 1-7. When comparisons were made directly with continuous ratings, R_1 and R_2 were converted from 70-point to 7-point figures as per Thompson et al. (2007).

Three general approaches were taken to the analyses, requiring careful selection of subgroups and tests necessitated by the complex nature of the experimental setup. For analyses of scores that would not be affected by the presence of errors in the performance (i.e., familiarity and likeability), 5 x 2 factorial repeated-measures ANOVAs were conducted among the 30 participants who had rated a version of all 5 trials. For analyses of scores affected by the presence of an error (e.g., T_1, R_1, R_2, R_3), comparisons could only be made between participants who had heard an error-free version or, in measuring time to (T_1) or rating at (R_1) the first decision, between participants who had heard the error-free version or the error-recap where the error took place after first decisions had been recorded. Between-groups analyses of the error conditions were conducted using factorial ANOVAs. Planned repeated contrasts and *t*-tests were used to examine the four hypotheses as appropriate. Where Mauchly's *W* indicated a violation of sphericity ($p < .05$), Greenhouse-Geisser corrections are reported.

Results

The first section of analyses examines familiarity and likeability levels of each of the works to validate assumptions of familiarity made in work selection and to define groupings for between-groups comparisons. This is followed by repeated-measures examinations of the five works to determine effects of familiarity, likeability, and composition length on time to first decision (T_1), final continuous rating (R_2), and overall written rating (R_3). Between-groups analyses are then used to determine effects of the error placement within the Etude, Waltz, and Prelude on the rating profile, and examine differences in the rating profile between the relatively unfamiliar Tarantelle and completely unfamiliar Caprice. The final section examines the influence of participants' perception of the difficulty of the works, musical experience, and listening preferences on the rating process.

PRELIMINARY ANALYSES: ESTABLISHING FAMILIARITY AND LIKEABILITY

The first stage of analyses involved defining reported familiarity and likeability levels of each of the works. As participants rated familiarity and likeability regardless of assigned error condition (and as they were asked to rate opinions of the composition itself, not of how it was performed), analyses could be conducted between all 30 participants who rated the five works. Table 2 shows descriptive values for the two dimensions, including correlations for each piece (using Kendall's tau due to the smaller sample size and large degree of tied ranks). While a matching overall trend from high to low familiarity and likeability can be seen across the compositions (see Figures 1 and 2), correlations

TABLE 2. *Familiarity and Likeability Ratings and Correlations for Each of the Five Works*

Work	Familiarity Mean (SD)	Likeability Mean (SD)	Correlation r_τ (p)
Etude	5.10 (1.90)	5.67 (1.21)	.46 (< .005)
Waltz	5.48 (1.86)	5.63 (1.40)	.09 (ns)
Prelude	2.70 (1.99)	4.73 (1.36)	.46 (< .01)
Tarantelle	2.28 (1.48)	4.72 (1.35)	.01 (ns)
Caprice	1.06 (0.25)	4.45 (1.66)	.06 (ns)

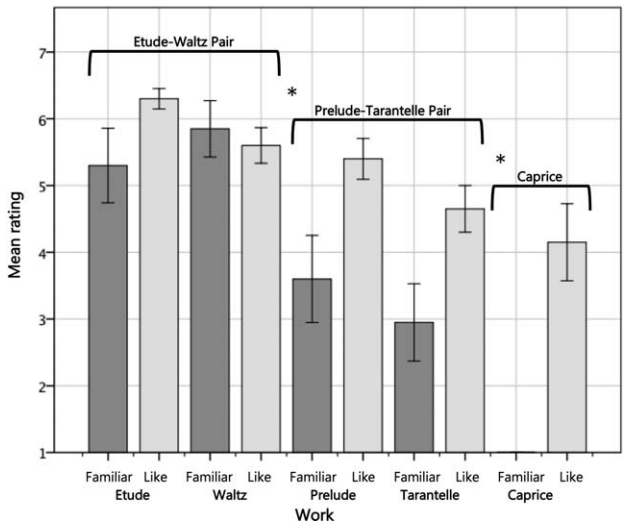

FIGURE 2. Mean familiarity and likeability scores for the five works. Three distinct groupings appeared: the Etude-Waltz pair showed high familiarity with no significant difference in likeability scores; the Prelude-Tarantelle pair showed significantly lower overall scores with significantly lower familiarity ratings than likeability ratings; the Caprice showed the lowest familiarity (approaching the minimum possible) with a significantly higher likeability score. Error bars show +/- 1 *SE*. * = $p <$ .005, as tested using planned repeated contrasts in which the mean of each combined familiarity/likeability score was compared with that of the next.

between each pair varied across the pieces, with only the Etude and Prelude showing significant medium correlations between the two items. For the remaining three works, familiarity with the work was not necessarily indicative of the degree to which participants liked the piece. The low mean familiarity score and standard deviation (where a response of 1 indicated that the participant had never heard the work) for the Caprice resulted from the fact that 28 of 30 participants had indicated that the work was entirely unknown to them.

To examine overall trends, a 5 x 2 factorial repeated-measures ANOVA was conducted with work (Etude, Waltz, etc.) and rating construct (familiarity and likeability) as within-subjects variables. The ANOVA was followed by planned repeated contrasts in which the mean of each score was compared with that of the next (i.e., A versus B, B versus C, C versus D, D versus E). This was done as the works were chosen with a hypothesized pattern towards descending familiarity from the Etude to the Caprice and as likeability was predicted to follow a similar trend across the five works, both of which were confirmed by the descriptive values. A significant main effect of work was found, $F_{3.06,88.61} = 34.53$, $p < .001$, $\eta_2 = .29$, resulting from the descending familiarity and likeability scores moving from the Etude and Waltz to the Caprice (see Table 2). The repeated contrasts showed that the descent was not uniform, however, with significant differences of familiarity and likeability (when combined) between the Waltz and Prelude, $F_{1,29} = 35.42$, $p < .001$, $r = .74$, and between the Tarantelle and Caprice, $F_{1,29} = 10.99$, $p < .005$, $r = .52$, but not between the Etude and Waltz or between the Prelude and Tarantelle (see Figure 2). A significant main effect of rating construct was found, $F_{1,29} = 55.24$, $p < .001$, $\eta_2 = .18$, where likeability scores were generally higher than those of the familiarity scores. These differences between constructs were not uniform, highlighted by the significant interaction between piece and construct, $F_{3.11,90.16} = 22.00$, $p < .01$, $\eta_2 = .08$. Once again, the planned repeated contrasts demonstrated that these interactions were only significant between the Waltz and Prelude, $F_{1,29} = 24.14$, $p < .001$, $r = .67$, and between the Tarantelle and Caprice, $F_{1,29} = 4.38$, $p < .05$, $r = .36$.

Together, these two sets of contrasts demonstrate three distinct groupings between familiarity and likeability scores in which the works were rated similarly: the Etude-Waltz pair, the Prelude-Tarantelle pair, and the Caprice (see Figure 2). The Etude-Waltz pair showed significantly higher scores overall (as demonstrated above) with no significant differences between familiarity and likeability, tested with multivariate simple effects tests using the estimated marginal means. The Prelude-Tarantelle pair showed lower overall familiarity and likeability, although both showed significantly higher familiarity scores than likeability scores with nearly identical effect sizes ($F_{1,29} = 44.43$, $p < .001$, $r = .78$; $F_{1,29} = 44.85$, $p < .001$, $r = .78$). Finally, the Caprice showed the lowest familiarity, with a significantly higher likeability score than its familiarity score ($F_{1,29} = 124.73$, $p < .001$, $r = .90$).

With the Etude-Waltz, Prelude-Tarantelle, and Caprice familiarity/likeability groupings established, these were used for the basis of repeated-measures comparisons to test the relationship between familiarity/likeability and the time to first decision (T_1) as posited in hypothesis 1. Furthermore, the Prelude-Tarantelle

grouping provided an opportunity to compare works of differing lengths while maintaining a consistent familiarity/likeability profile. This allowed for a direct examination of hypothesis 2, which predicted a decrease in time to first rating (T_1) for a work of shorter length. Correlations (Kendall's tau) between time to first decision (T_1) and the first (R_1) and overall (R_3) ratings were conducted for each of the five works to test the assumption that any significant differences in time to first ratings were due to the nature of the works and not simply a result of differences in the perceived quality of the individual performances. Correlations remained very low (< .20) and nonsignificant across the 10 tests, supporting this assumption and in line with previous findings (Waddell & Williamon, 2017b).

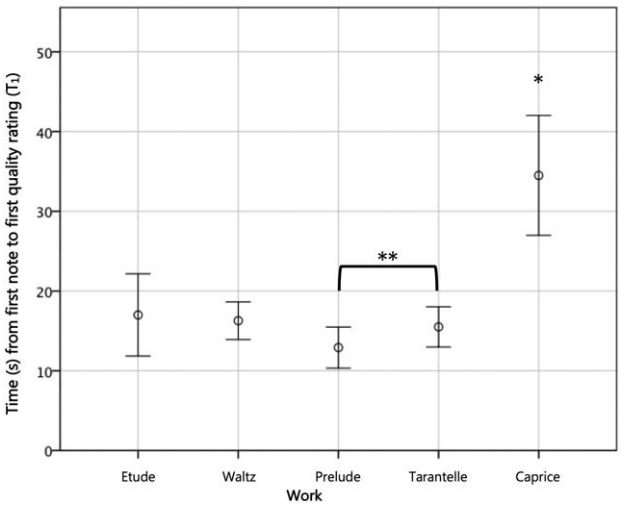

FIGURE 3. Mean time in seconds from the first note to first decision (T_1) for the five works. The Caprice resulted in a significantly longer time to first decision than the four stylistically similar works of Chopin in a repeated-measures comparison of 11 participants. * = $p < .05$ as tested using planned repeated contrasts in which each time was compared with that of the next. A further test between works of equal familiarity but differing length—the Prelude and Tarantelle—with $n = 16$ found a significantly lower time to first decision for the shorter Prelude (27 s in length) versus the Tarantelle (156 s in length). ** < .05 as tested using a Wilcoxon signed-rank test. Error bars show +/- 1 *SE*.

HYPOTHESES 1 AND 2: REPEATED-MEASURES EFFECTS OF FAMILIARITY, LIKEABILITY, AND LENGTH ON TIME TO FIRST DECISION (T_1)

To examine the effect of condition on the time to first decision (T_1), a repeated-measures ANOVA was calculated between the five works among the 11 participants who had rated all five performances without an error at the beginning. Despite the small sample size, a significant main effect of condition was found, $F_{2.16,21.66} = 5.20$, $p < .05$, $\eta_2 = .52$. Again, a planned reverse contrast was used to compare the differences between each condition and the previous condition, as was employed in the likeability/familiarity comparisons. The only significant difference was between the Tarantelle and Caprice where a mean 34.50 s ($SD \pm 24.93$) was taken to first decision versus 15.50 s ($SD \pm 8.35$; $F_{1,10} = 6.78$, $p < .05$, $r = .64$ (see Figure 3). No significant difference was found between the other levels, although medium effect sizes were seen between the Etude and Prelude ($r = .29$) and between the Prelude and Tarantelle ($r = .38$; for reference, the Etude versus Waltz comparison showed $r = .04$) suggesting the descriptively shorter time to first decision for the Prelude ($M = 12.90$, $SD \pm 8.56$ s) versus the Waltz ($M = 16.27$, $SD \pm 7.83$ s) and Tarantelle ($M = 15.50$, $SD \pm 8.35$ s) could represent a significant effect in an analysis with greater power.

While the small sample size afforded by the 5-group ($n = 11$) test had enough power to reveal the relatively large difference between the Caprice and the remaining works, with participants taking on average twice as long to register their first judgment, the nature of the experimental setup allowed for larger sample sizes in focused comparisons. Hypothesis 2 suggested that the shorter Prelude would result in shorter time to first decision than a work of equal familiarity, which was above demonstrated to be the Tarantelle and could be tested with a higher degree of power as 16 participants rated both the Tarantelle and the error-free version of the Prelude. This hypothesis was confirmed with a one-way related-samples Wilcoxon signed-rank test, $Z_{16} = 1.66$, $p < .05$, $r = .28$, with first decisions for the Prelude taking a mean 10.83 s (median = 7.75, $SD \pm 7.70$) and for the Tarantelle a mean 13.38 s (median 10.25, $SD \pm 7.74$). For comparison, a similar test run between groups of similar familiarity (the Etude-Waltz pair) showed no significant difference, despite an even greater availability of matched pairs ($n = 20$) and the corresponding increase in power.

Correlations between each of the familiarity scores for the Etude, Waltz, Prelude, and Tarantelle and their respective times to first decision (T_1) were tested using Kendall's tau; the Caprice could not be tested as only 2 of the 30 participants indicated they had ever heard the work. *R* values were low (< .10) and none approached significance, further suggesting a lack of relationship between familiarity and time to first decision among the stylistically familiar Chopin works. Examination of the likeability scores (which included the Caprice) also showed no significant correlations between how much one liked the work and the speed with which a first rating was made.

Overall, these analyses revealed a significant effect of work on the time taken to form a first decision. Participants rating the Caprice took significantly longer to form their first judgments, due perhaps to the unfamiliarity of the piece and its composer. This relationship between familiarity and decision time was not reflected among the stylistically similar Chopin works, although a significantly faster time to first decision was demonstrated within the shorter Prelude.

COMPARISONS OF THE FINAL RATINGS (R_2 AND R_3)

Direct comparisons of the final ratings in this study are complicated by the experimental setup, in which very few (3) participants heard no error (i.e., uncontaminated) versions of all five works. While such comparisons were not the primary focus of the study due to its focus instead on the decision-making process, two of interest could be made: 1) whether final continuous scores (R_2) were representative of the final written ratings (R_3); and 2) individual correlations between familiarity, likeability, and the final scores within each work.

For the first comparison, the R_2 scores were converted to a 7-point scale as described in "Data treatment and analysis" allowing for direct comparison with R_3. A 5 x 2 factorial repeated-measures ANOVA was then calculated with work and rating condition (converted R_2 versus R_3) as within-subjects factors. A significant main effect of work was found, $F_{4,112} = 6.18$, $p < .001$, $\eta_2 = .16$, where final scores increased from the Etude as the lowest to the Caprice as the highest (see Table 3), unsurprising as these ratings included versions of the Etude, Waltz, and Prelude that contained performance errors. Crucially, no significant main effect of rating condition was found, or any significant interaction between work and rating condition. This suggests that the final continuous ratings (R_2) were reflected in the overall written scores (R_3) across all works, supporting the use of continuous ratings as an adjunct for standard written rating procedures and for using R_2 scores to examine the effects of error placement on final scores.

TABLE 3. *Mean Final Continuous Scores (R_2), Final Scores Converted to a 7-point Scale, and Overall Written Scores (R_3) for the Five Works*

Work	*R_2 (SD)*	*Converted R_2 (SD)*	*R_3 (SD)*
Etude	38.31 (15.52)	4.31 (1.58)	4.31 (1.55)
Waltz	39.93 (14.98)	4.48 (1.45)	4.41 (1.23)
Prelude	46.31 (9.51)	5.10 (1.01)	4.93 (0.80)
Tarantelle	44.66 (13.67)	4.93 (1.31)	4.84 (1.25)
Caprice	51.72 (10.03)	5.62 (0.98)	5.41 (0.81)

Correlations were tested between each of the familiarity and likeability scores for each of the works (again, correlations could not be checked with familiarity for the Caprice) and their respective final continuous ratings (R_2) using Kendall's tau. The strongest correlation, and the only one to reach significance following a Bonferroni correction for multiple comparisons, was a medium correlation between likeability and final continuous score for the Caprice ($r_\tau = .46$, $p < .01$). A linear regression between the two variables produced a significant model, $F_{1,27} = 9.76$, $p < .005$, $R^2 = .27$, $b = 3.10$, wherein an increase of one point on the 7-point likeability scale predicted a 3.1-point increase on the 70-point final continuous rating (see Figure 4).

HYPOTHESIS 3: BETWEEN-GROUPS EFFECTS OF THE ERROR ON TIME TO FIRST DECISION (T_1)

In the cases of the Etude, Waltz, and Prelude, listeners were randomly assigned to a condition with no performance error (no error), a performance error in the opening seconds (error-start), or in the case of the Etude and Waltz, that same performance error at the recapitulation of the opening material (error-recap). This randomization was not consistent for each work; a participant hearing a no error version of the Etude, for example, may have heard a start-error version of the Waltz. Thus, direct repeated-measures comparisons were not possible. Instead, the data offered the opportunity for an effective replication of the test with the same sample but a new stimulus and different randomization.

Hypothesis 3 predicted that the time to first decision (T_1) for a performance would be lower in conditions

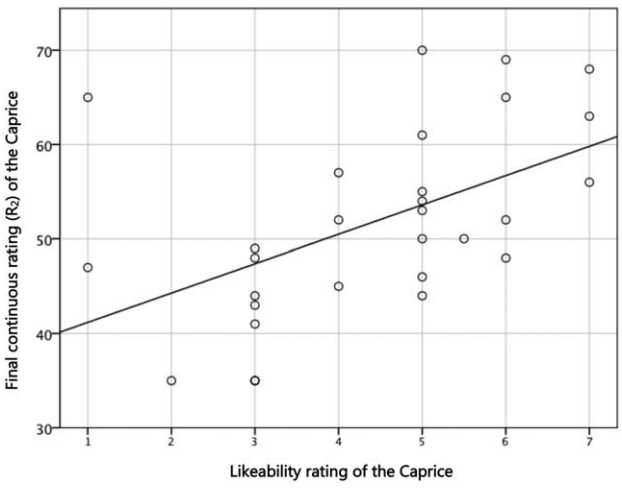

FIGURE 4. Scatter plot showing likeability score and final continuous rating (R_2) for the Caprice, wherein greater liking of the composition predicted a higher quality rating for the performance ($R^2 = .27$).

with an error at the beginning when compared with those without. Thus, one-way ANOVAs were conducted for the Etude and Waltz with error condition (no error, error-start, and error-recap) as a between-subjects factor and T_1 as the dependent variable. These were followed by planned simple contrasts where each error condition was compared with the no error control. For the Etude, while no main effect of error condition was found, the contrast showed that the mean 6.36 s ($SD \pm 3.43$) to the first decision in the error-start condition was significantly shorter, $t_{39} = -8.85$, $p < .05$, $d = .72$, than the mean 15.21 s ($SD \pm 17.11$) in the no error control condition (see Figure 5A). This finding was replicated in examining the Waltz; the main effect of error condition was nonsignificant, but the contrasts again showed the mean 7.69 s ($SD \pm 4.94$) to an error-start first decision was significantly shorter, $t_{39} = -13.51$, $p < .05$, $d = .63$, than the 21.21 s ($SD \pm 29.78$) in the no error control condition (see Figure 5B). No significant differences were found between the error-recap conditions and the no error control in either work. As the Prelude was the shorter work, only two conditions (no error and error-start) existed and required testing. However, to maintain consistency in alpha inflation, ANOVA was also used to examine differences between the conditions. No significant main effect was found (see Figure 5C), influenced perhaps by the fact that the shorter length of the work already reduced times to first decision in the Prelude condition. Overall, these results support hypothesis 3; participants made their first decisions more quickly when an error was present in the opening seconds.

HYPOTHESIS 4: THE EFFECTS OF THE ERRORS ON FIRST AND FINAL RATINGS AND CONTINUOUS RATING PROFILE

The same approach as above could be taken for analyses of the rating profile, treating the Etude and Waltz as replications of the same study with different randomization procedures. In this case, tests examined differences between first (R_1) and final (R_2) ratings—as the analyses above demonstrated that R_2 scores were representative of the final R_3 written scores—and how they were affected by the presence of errors. For the Etude and Waltz, differences in the overall rating profile were tested with a mixed 2 x 3 ANOVA in which the first and final ratings (R_1 and R_2) served as the within-subjects variable and the 3 error conditions (no error, error-start, and error-recap) the between-groups. A planned simple contrast was used to determine group differences between the error conditions in which the error-start and error-recap conditions were compared with the no error control.

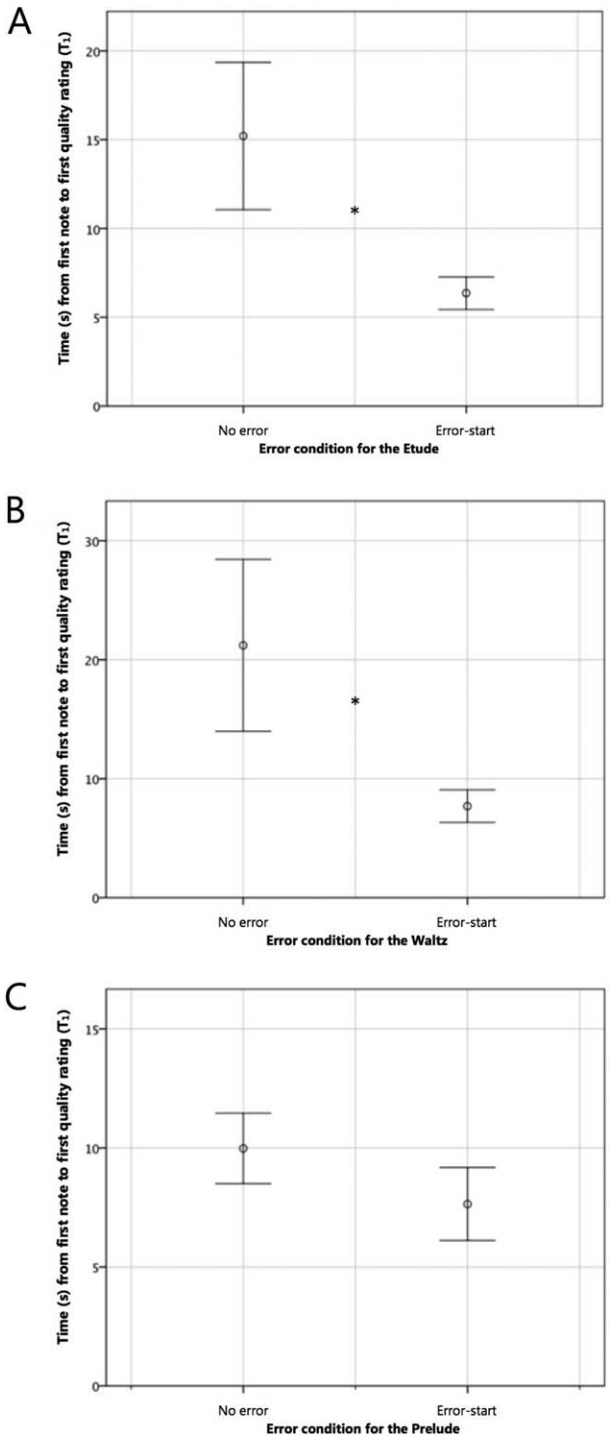

FIGURE 5. Mean time to first decision between the no error control and error-start conditions for the (A) Etude, (B) Waltz, and (C) Prelude. The Etude and Waltz showed a significantly shorter time to first decision when an error was inserted into the opening seconds of the performance; *$p < .05$ as tested with planned simple contrasts where each error condition was compared with the no error control. The Prelude did not show a significant difference. Error bars show +/− 1 *SE*.

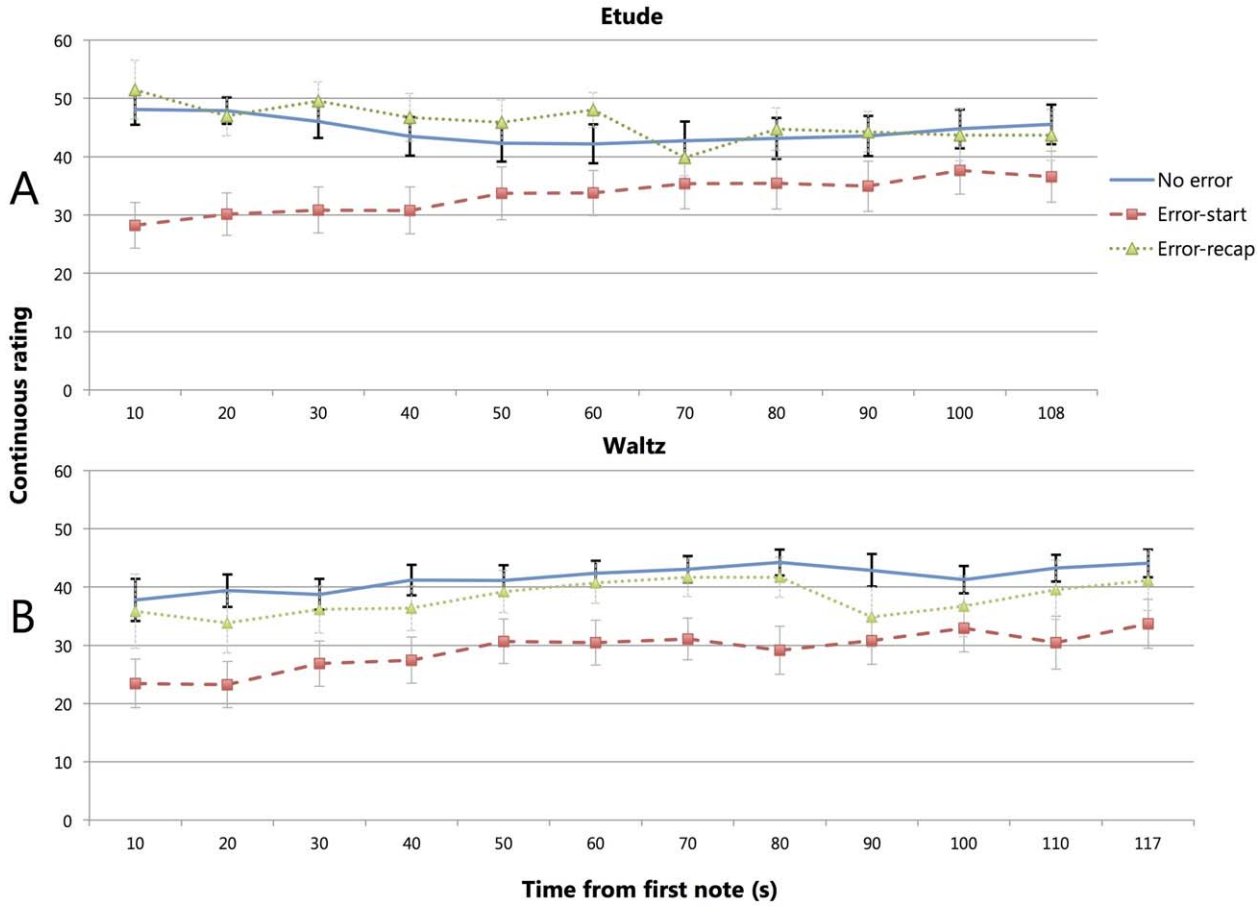

FIGURE 6. Continuous rating profiles of the no error, error-start, and error-recap conditions for the Etude and the Waltz, showing mean ratings at 10-s intervals. In both cases, the error-start condition resulted in a significantly lower first (R_1) and final (R_2) rating than the no error control. The error-recap condition resulted in a noticeable drop at the point of the error—between 60 and 70 s in the Prelude and 80 and 90 s in the Waltz—that recovered by the end of the performance, resulting in a final score (R_2) not significantly different from the no error control. Error bars show +/- 1 *SE*.

In the case of the Etude, no significant repeated-measures effect was found, although a significant main between-groups effect of error condition was demonstrated, $F_{2,39} = 4.78$, $p < .05$, $\eta_2 = .20$, as was a significant interaction between rating and error condition, $F_{2,39} = 3.38$, $p < .05$, $\eta_2 = .14$. As can be seen in Figure 6A, this was due to the general downward trend of the no error and error-recap conditions and the upward trend of the error-start condition. The simple contrast confirmed that, while the error-recap performance did not differ significantly from the no-error performance in terms of first and final ratings, the performance with an error at the beginning did, $t_{39} = -11.83$, $p < .05$, $r = .88$, prompting first ratings ($M = 28.00$, $SD \pm 13.81$) well below those of the standard performance ($M = 42.71$, $SD \pm 11.56$) and concluding with a narrower but still significant gap ($M = 36.57$, $SD \pm 16.40$ versus $M = 45.53$, $SD \pm 13.98$). Thus, when the error was placed at the beginning of the work, the evaluators penalized the performer with a significantly lower rating that did not recover to no error levels by the end of the performance. In the case of the performance with an error part way through, the continuous measures data revealed a sharp drop in ratings immediately following the missed notes, but interestingly, this deficit was "forgiven" by the end of the work (see Figure 6A), with no significant difference in the final score. An observer seeing only the final ratings would have no indication that an error had been made.

An analysis of the Waltz replicated the overall finding but did so under different circumstances. While the between-group analyses again showed a significant difference of error condition, $F_{2,39} = 4.35$, $p < .05$, $\eta_2 = .18$, there was in this case an additional main

repeated-measures effect of first-to-final rating, $F_{1,39} = 10.45$, $p < .005$, $\eta_2 = .20$, and no significant interaction. The reason for this reverse of significant main and interaction effects can be seen in Figure 6B where all three conditions show a similar upward trend for the Waltz in contrast to the converging lines of the Etude (see Figure 6A) and thus show a significant overall increase in rating across the performances of the Waltz. However, the error-start condition once again lay significantly lower than the standard performance, confirmed by a significant difference from the standard condition shown by the simple contrast, $t_{39} = -12.10$, $p < .01$, $r = .89$, with lower first ratings ($M = 24.92$, $SD \pm 12.10$ versus $M = 38.77$, $SD \pm 8.67$) and final ratings ($M = 33.69$, $SD \pm 15.25$ versus $M = 44.06$, $SD \pm 9.87$). As with the Etude, the version of the Waltz with an error midway through, despite again causing an immediate drop in rating at the point of the mistake, did not differ significantly from the standard performance in terms of first or final ratings (see Figure 6B).

Regarding the Prelude, no significant main effects of rating or condition, or interactions between them, were found as a result of the error at the start. This mirrors the previous section, where the error also failed to affect time to first decision in the Prelude despite a significant effect within the Etude and Waltz. This suggests that the error itself may not have been dramatic enough to cause a reaction in the Prelude. For the Etude and Waltz the results are clear: an error in the opening material caused a shorter time to first decision and a lower initial rating that never fully recovered, where an error midway through caused a temporary drop that was not significantly reflected in the final ratings.

HYPOTHESIS 1 REVISITED: THE EFFECTS OF FAMILIARITY ON CONTINUOUS COMPARISONS OF THE TARANTELLE AND CAPRICE RATING PROFILES

As 30 participants provided continuous ratings of both the Tarantelle and Caprice, and as analyses of time to first decision (T_1) demonstrated a different rating process between the two works in the greater amount of time taken to form a first decision, similar continuous analyses could be conducted to further test hypothesis 1, which predicted that familiarity would affect the time to form a final decision. To determine the point at which the cohort reached a final consensus on the two works, scores at 10-s intervals from the beginning of the performance were extracted and analyzed to determine the point at which raters' responses did not differ significantly from their final scores. Repeated-measures ANOVAs were calculated for each work followed by reverse simple contrasts comparing each 10-s mean score with the final, beginning with the interval at which at least 50% of the participants had first reported (thus providing full datasets for analysis): this was the 10-s mark for the Tarantelle (with 15 respondents) and the 20-s mark for the Caprice (with 20 respondents). For the Tarantelle, the overall effect was not significant, although the contrasts showed a significant difference between the final score ($M = 49.07$, $SD \pm 10.54$) and both the 10-s point ($M = 43.93$; $SD \pm 10.56$; $F_{1,14} = 10.33$, $p < .001$, $r = .65$) and 20-s point ($M = 43.87$, $SD \pm 11.11$; $F_{1,14} = 6.43$, $p < .05$, $r = .56$), with no significant difference from the end from the 30-s point onward. In the case of the Caprice, a significant main effect of the ANOVA was found, $F_{3.51,66.78} = 9.48$, $p < .001$, $\eta_2 = .33$, and the contrast revealed significant differences between the 20–80-s points and the final score, with no significant results following. As can be seen in Table 4, effect sizes at the cutoff are still moderately strong, but using the significance value as a conservative cutoff, these results suggest a time to final group decision at least 3 times longer in the Caprice than the Tarantelle (see Figure 7).

TABLE 4. *Mean Performance Ratings for the Caprice at 10-s Increments from the Beginning of the Recording*

Time (s)	*Mean*	*SD*	*F*	*p*	*r*
20	43.95	8.94	20.97	.000	.72
30	45.65	10.04	16.75	.001	.68
40	46.20	10.56	14.90	.001	.66
50	47.70	10.19	9.00	.007	.57
60	48.20	10.56	11.52	.003	.61
70	48.80	10.46	8.76	.008	.56
80	49.55	11.10	5.81	.026	.48
90	50.00	11.26	3.71	.069	.40
...					
Final (152)	52.50	10.10			

Note: Results were derived from a repeated-measures ANOVA comparing each score with the final continuous rating.

CORRELATIONS WITH EXPERIENCE, DIFFICULTY, AND LISTENING PREFERENCES

Further tests were conducted to determine whether years of musical experience, perceived difficulty of the work, typicality of the performance, and listening preference (Romantic when examined against the Chopin works, twentieth-century when examined against the Caprice) correlated with time to first decision (T_1) or final continuous ratings (R_2). The only significant correlations (after correcting for multiple comparisons across the five works) were between perceived difficulty

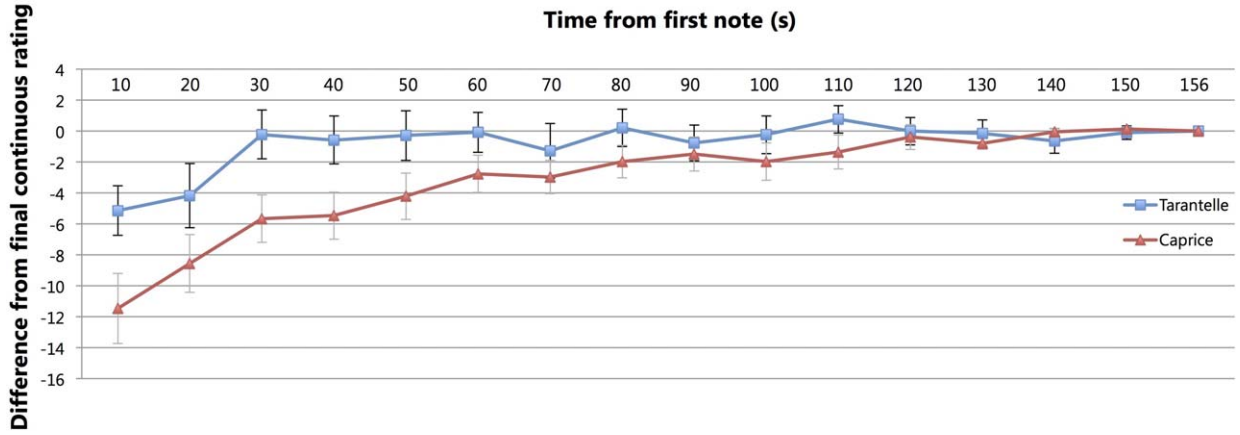

FIGURE 7. Continuous rating profiles of the Tarantelle and the Caprice. Data are normalized to show mean difference from the final score at 10-s intervals. Using a reverse simple contrast the Tarantelle showed no significant difference from the final score from 30 s onward, whereas the Caprice showed no significant difference from 90 s onward. Final time for the Caprice was at 152 s. Error bars show +/- 1 *SE*.

of performance and final continuous score (R_2) for the Etude, $r_\tau = .38$, $p < .005$, and the Caprice, $r_\tau = .40$, $p < .01$, where higher difficulty scores correlated with higher performance ratings. This relationship showed small but nonsignificant correlations across the other three works.

Discussion

The purpose of most music performance quality assessments, whether conducted as part of an audition, recital, competition, or examination, is to determine the quality of the performance and performer. They are not intended to be an assessment of the quality of the work being performed, at least not in most Western classical contexts where the composer and performer of the work are separate entities. Otherwise, music competitions intended to identify a top performer would become repeating debates over the relative merits of Mozart and Haydn or of Beethoven's Op. 110 and 111. We questioned this assumption, examining how qualities related to the repertoire—such as its length, familiarity, and likeability—affected the process by which assessments are formed. We also examined the nature of performance errors, and whether an error placed at the beginning of the performance had the same effect as the same error placed midway through the piece. To achieve this, trained musicians evaluated recordings of five works, selected to vary in familiarity and length, using a continuous measures methodology and standard written questionnaires. Furthermore, we manipulated three of the works to create conditions with performance errors at the beginning of the performance, and two of those manipulated again to have errors midway through the performance. The continuous measures approach revealed effects of these variables that could not have been seen in the standard written evaluations which followed, allowing for direct examination of each of four hypotheses.

Hypothesis 1 predicted that works of lesser familiarity would result in a longer time to first decision and that this would be exaggerated for a work of unfamiliar tonal structure and composer. This hypothesis was partially confirmed: within performances by a familiar composer (Chopin), relative familiarity and likeability had no effect on or correlation with time to first decision (T_1). However, for the unknown work by the unfamiliar composer, the first decision took significantly longer. Furthermore, the rating profile for the Caprice showed that the group took three times longer to settle on their final decisions than they did for Chopin's Tarantelle of equal length and that the likeability of the Caprice showed a medium correlation with the final continuous score (R_2).

Hypothesis 2 predicted that a work of shorter length would also result in a shorter time to first decision. This was confirmed, wherein the 27-s long Prelude resulted in a significantly shorter T_1 score compared with the 156-s Tarantelle, which matched in familiarity and likeability ratings. Hypotheses 3 and 4 concerned the placement of performance errors, predicting that an error in the opening seconds of a performance would both reduce the time to first decision and result in a significantly lower final quality rating when compared with a performance with no error or an error in the middle. The continuous ratings confirmed both. Time to first

decision was shorter for both the Etude and Waltz when the initial error was present, and while a decreasing trend was seen for the Prelude, it was not significant. For both the Etude and the Waltz, the error-start condition caused a significantly lower first and final rating than the no error control, and the continuous measurement profile demonstrated that the error-recap condition, while not differing from the control in terms of first or final decision, caused an immediate negative reaction to the error that recovered by the end of the performance. No effect of the error on ratings was seen for the short Prelude.

Overall, these results support previous findings that time to first decision takes place within an average window of 15 s when rating audio (Thompson et al., 2007) or audiovisual (Waddell & Williamon, 2017b) recordings of performances of high musical quality. The present results go on to demonstrate that the time to first decision can vary, as Waddell and Williamon (2017b) found when stage entrances judged as inappropriate triggered shorter times to first decision. Here, the unfamiliar nature of the Caprice led to a twofold increase. This could suggest that the listeners needed more time to orient themselves to the work and determine their criteria for assigning performance quality. Alternatively, the unfamiliar nature of the work could have taken attentional focus away from the task at hand. Moreover, a shorter work resulted in a decrease in time to first decision. This supports the findings of Tullar et al. (1979), who found that the decision-making process took longer when assessors were informed that job interviews would be longer. This suggests that assessors accelerate the decision-making process when they are aware that they will have less time to conduct it. Anecdotally, the participants in this study often expressed visible and/or verbal surprise when informed that the work they were about to assess was less than 30 s in length; many seemed aware that this was a relatively rare situation in rating full performances of standard repertoire and perhaps prepared themselves accordingly. Finally, the insertion of an error at the beginning of the performance caused significantly lower times to first decision for the Etude and the Waltz. Thus, participants were more temporally reactive to negative than positive (or at least neutral) information in the opening moments of the performance. That this effect was not replicated within the Prelude could be explained by the corresponding lack of significant effects on the first and final quality ratings; it could be that the error itself was not as easily perceived or considered as serious as the error in the other two works.

The effects of the error on the start and middle of the Etude and Waltz were dramatic, demonstrating that the temporal location of an otherwise identical error matters. This provides strong support of Ybarra's (2001) findings that it is difficult to reverse judges' negative first impressions. In this study, the significantly lower first ratings did recover over time, but never reached the height of the final score in the no error conditions. There are at least two possible explanations for these findings. It may be that the low quality of the opening seconds caused an anchoring effect in the listener, whereby the remainder of the performance was perceived as being of lower quality and was rated as such, with the perceptual effect of the error gradually fading. Alternatively, the listeners may have perceived the quality of the rest of the performance as high as those rating the no error condition, but their moment-by-moment continuous rating represented an overall decision reflecting both the current material and that which has come before it. The lack of a significant difference between final continuous (R_2) and overall written (R_3) scores in this study supports the latter explanation, as it suggests that an extract of a moment-by-moment continuous rating emulates the same performance-averaging result as provided when a judge is asked to give an overall quality score. This continual comparison is also supported by research examining evaluations of affective experience that show global evaluations can be best predicted by an averaging of extreme peaks in rating and the material in recent memory (e.g., Fredrickson & Kahneman, 1993; Varey & Kahneman, 1992). Retrospective ratings of pain, for example, have been found to correlate most strongly with the point of highest pain intensity and the intensity during the final stage of the treatment, not reflecting the duration of treatment or accumulated pain ratings (Redelmeier & Kahneman, 1996). The question remains as to whether, given enough time, ratings of performances with an initial error could eventually recover regardless of their severity. Future studies could examine the effect in pieces of significantly longer length; the classical repertoire offers examples of works that are hours long. They could also examine how the presence of errors in one performance affects the ratings of subsequent performances by the same performer, as participants in this study were informed that they were rating different pianists. The role of musical structure may also be important. Perhaps those hearing the mistake at the recapitulation were more forgiving because they had already heard an example of the performer navigating that exact passage correctly at the beginning of the piece. On the other hand, those hearing the mistake in the introduction did

not obviously reward the performer for avoiding the error later on.

The positive correlation between likeability and final quality ratings in the unfamiliar Caprice raises interesting questions about reactions to a completely unfamiliar performance, as the finding was not replicated in the other works where familiarity scores were higher and rating processes (represented by the time to first decision) were unchanged. As this is a correlational finding, the direction of causality can only be speculated upon, although the fact that the finding was not replicated among the more familiar works suggests that it was not the case of participants being unable to separate the constructs of likeability and performance quality or having a third variable (e.g., tendency to provide generally higher responses on the rating scales) influencing both. It may be that, when orienting oneself to an unfamiliar work in an unfamiliar style, one's enjoyment of the work itself influences the interpretation of performance quality. Alternatively, those that felt the work was performed better may have developed a stronger liking for the composition itself. Further work is required with other unfamiliar compositions to determine whether this is a generalizable effect, as well as the direction of causality.

The examination of the errors in the present study focused only on the works of higher familiarity in a recognizable tonal style, in which an error could be easily perceived as a harsh dissonance. This raises questions about the nature and perception of performance errors within a contemporary work that lacks the familiar tonal frameworks of standard Romantic repertoire. It is interesting to note the similarities between the continuous data for the Caprice and those for the Etude and Waltz when the error was inserted at the beginning of the performance. It stands to reason that participants took significantly longer to come to their first decision when rating the Caprice; as discussed above they had to acclimatize to an unfamiliar style. However, when that first decision was eventually made it was established, on average, at a point significantly lower than the final rating, gradually increasing across the length of the performance to reach the highest mean final rating of the five works (see Table 3 and Figure 7). This was mirrored in the error-start conditions. While the comparison is cursory in this case, one could hypothesize that participants initially could not determine whether errors were being made and, hearing the repeated extra-tonal dissonances, rated it as though they were. It could also be that the performance was of genuinely lower quality at the beginning, although this would contradict the performer's reported intention and perception of a polished performance throughout and one that was true to the notated score. Interpretation is limited by the fact that only one such composition was investigated in this study. As discussed above, future work should examine whether this effect is generalizable. It should also compare ratings of unknown works with a specialized cohort familiar with its structure, language, and style, either through prior experience or an experimental intervention. A growing body of research has demonstrated the ability of listeners to remember and perceive errors in non-tonal contexts when first given an accurate reference (e.g., Dienes & Longuet-Higgins, 2004; Kuusi, 2015; Ockelford & Sergeant, 2013; Samplaski, 2004).

While the present study was conducted in laboratory settings with digitally manipulated stimuli, every effort was made to replicate the experience of rating audio recordings of genuine performances as a juror might be asked to do in an audition or competition setting. As such, there are several points of which musicians can take note. The nature of their repertoire, whether its length or its familiarity, can affect the process by which their performances are judged. In particular, unfamiliar works may cause their audiences to take longer to orient themselves to the performance and be more critical in their initial judgments of quality. In addition, the adage that "first impressions count" appears to hold true. Performers are well advised to ensure that, if nothing else, the opening seconds of their performances are as prepared and polished as possible. Otherwise, a few misplaced notes could tarnish judgements of the thousands that follow.

Author Note

The authors would like to thank Elisabeth Cook, Richard Dickins, the pianists who provided the recordings, and the study participants for their help throughout the project.

Correspondence concerning this article should be addressed to Aaron Williamon, Centre for Performance Science, Royal College of Music, Prince Consort Road, London, SW7 2BS, United Kingdom. E-mail: aaron.williamon@rcm.ac.uk

References

Buckley, M. R., & Eder, R. W. (1988). BM. Springbett and the notion of the "snap decision" in the interview. *Journal of Management*, *14*, 59–67.

Byo, J. L. (1993). The influence of textural and timbral factors on the ability of music majors to detect performance errors. *Journal of Research in Music Education*, *41*, 156–167.

Byo, J. L. (1997). The effects of texture and number of parts on the ability of music majors to detect performance errors. *Journal of Research in Music Education*, *45*, 51–66.

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

Dienes, Z., & Longuet-Higgins, C. (2004). Can musical transformations be implicitly learned? *Cognitive Science*, *28*, 531–558.

Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education*, *20*, 268–272.

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

Flossmann, S., & Widmer, G. (2011). Toward a model of performance errors: A qualitative review of Magaloff's Chopin. In A. Williamon, D. Edwards, & L. Bartel (Eds.), *Proceedings of the International Symposium on Performance Science 2011* (pp. 63–68), Utrecht, The Netherlands: European Association of Conservatoires.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*, 45–55.

Geringer, J. M., Madsen, C. K., & Gregory, D. (2004). A fifteen-year history of the Continuous Response Digital Interface: Issues relating to validity and reliability. *Bulletin of the Council for Research in Music Education*, *160*, 1–15.

Griffiths, N. K. (2008). The effects of concert dress and physical appearance on perceptions of female solo performers. *Musicae Scientiae*, *12*, 273–290.

Griffiths, N. K. (2010). 'Posh music should equal posh dress': An investigation into the concert dress and physical appearance of female soloists. *Psychology of Music*, *38*, 159–177.

Griffiths, N. K. (2011). The fabric of performance: Values and social practices of classical music expressed through concert dress choice. *Music Performance Research*, *4*, 30–48.

Himonides, E. (2011). Mapping a beautiful voice: The continuous response measurement apparatus (CReMA). *Journal of Music, Technology and Education*, *4*, 5–25.

Kinney, D. W. (2009). Internal consistency of performance evaluations as a function of music expertise and excerpt familiarity. *Journal of Research in Music Education*, *56*, 322–337.

Kroger, C., & Margulis, E. H. (2017). "But they told me it was professional": Extrinsic factors in the evaluation of musical performance. *Psychology of Music*, *45*, 49–64.

Kuusi, T. (2015). Musical training and musical ability: Effects on chord discrimination. *Psychology of Music*, *43*, 291–301.

Maidhof, C., Pitkaniemi, A., & Tervaniemi, M. (2013). Predictive error detection in pianists: A combined ERP and motion capture study. *Frontiers in Human Neuroscience*, *7*:587, 1–14.

Maidhof, C., Rieger, M., Prinz, W., & Koelsch, S. (2009). Nobody is perfect: ERP effects prior to performance errors in musicians indicate fast monitoring processes. *PLoS One*, *4*(4), e5032.

Maidhof, C., Vavatzanidis, N., Prinz, W., Rieger, M., & Koelsch, S. (2010). Processing expectancy violations during music performance and perception: An ERP study. *Journal of Cognitive Neuroscience*, *22*, 2401–2413.

McPherson, G. E., & Schubert, E. (2004). Measuring performance enhancement in music. In A. Williamon (Ed.), *Musical excellence: Strategies and techniques to enhance performance* (pp. 61–82). Oxford, UK: Oxford University Press.

Mishra, J. (2010). Effects of structure and serial position on memory errors in musical performance. *Psychology of Music*, *38*, 447–461.

Ockelford, A., & Sergeant, D. (2013). Musical expectancy in atonal contexts: Musicians' perception of "antistructure." *Psychology of Music*, *41*, 139–174.

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception*, *30*, 71–83.

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*, 3–8.

Repp, B. H. (1996). The art of inaccuracy: Why pianists' errors are difficult to hear. *Music Perception*, 161–183.

Ruiz, M. H., Jabusch, H.-C., & Altenmüller, E. (2009). Fast feedforward error-detection mechanisms in highly skilled music performance. In A. Williamon, S. Pretty, & R. Buck (Eds.), *Proceedings of the International Symposium on Performance Science 2009* (pp. 187–197). Utrecht, The Netherlands: European Association of Conservatoires.

Ryan, C., & Costa-Giomi, E. (2004). Attractiveness bias in the evaluation of young pianists' performances. *Journal of Research in Music Education*, *52*, 141–154.

Ryan, C., Wapnick, J., Lacaille, N., & Darrow, A. A. (2006). The effects of various physical characteristics of high-level performers on adjudicators' performance ratings. *Psychology of Music*, *34*, 559–572.

Samplaski, A. (2004). The relative perceptual salience of Tn and TnI. *Music Perception*, *21*, 545–559.

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.

Thompson, S. (2007). Determinants of listeners' enjoyment of a performance. *Psychology of Music*, *35*, 20–36.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception*, *21*, 21–41.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Perception*, *25*, 13–29.

Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 14580–14585.

Tsay, C.-J. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes*, *124*, 24–33.

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.

VanWeelden, K. (2004). Racially stereotyped music and conductor race: Perceptions of performance. *Bulletin of the Council for Research in Music Education*, *160*, 38–48.

Varey, C., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making*, *5*, 169–185.

Waddell, G., & Williamon, A. (2017a) Measuring the audience. In S. Lee (Ed.), *Scholarly research for musicians* (pp. 148–155), Abingdon: Routledge.

Waddell, G., & Williamon, A. (2017b). Eye of the beholder: Stage entrance behavior and facial expression affect continuous quality ratings in music performance. *Frontiers in Psychology, 8*(513), 1–14.

Wapnick, J., Campbell, L., Siddell-Strebel, J., & Darrow, A.-A. (2009). Effects of non-musical attributes and excerpt duration on ratings of high-level piano performances. *Musicae Scientiae*, *13*, 35–54.

Wapnick, J., Darrow, A. A., Kovacs, J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education*, *45*, 470–479.

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.

Wapnick, J., Ryan, C., Campbell, L., Deek, P., Lemire, R., & Darrow, A.-A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performances. *Journal of Research in Music Education*, *53*, 162–176.

Wesolowski, B. C. (2017). A facet-factorial approach towards the development and validation of a jazz rhythm section performance rating scale. *International Journal of Music Education*, 17–30.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*, 147–170.

Winter, N. (1993). Music performance assessment: A study of the effects of training and experience on the criteria used by music examiners. *International Journal of Music Education*, *22*, 34–39.

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.