Journal of Experimental Psychology:
Human Perception and Performance
2011, Vol. 37, No. 3, 921–934

© 2011 American Psychological Association
0096-1523/11/$12.00 DOI: 10.1037/a0021922

# Perception of Emotional Expression in Musical Performance

Anjali Bhatara, Anna K. Tirovolas, Lilu Marie Duan, Bianca Levy, and Daniel J. Levitin
McGill University

Expression in musical performance is largely communicated by the *manner* in which a piece is played; interpretive aspects that supplement the written score. In piano performance, timing and amplitude are the principal parameters the performer can vary. We examined the way in which such variation serves to communicate emotion by manipulating timing and amplitude in performances of classical piano pieces. Over three experiments, listeners rated the emotional expressivity of performances and their manipulated versions. In Experiments 1 and 2, timing and amplitude information were covaried; judgments were monotonically decreasing with performance variability, demonstrating that the rank ordering of acoustical manipulations was captured by participants' responses. Further, participants' judgments formed an S-shaped (sigmoidal) function in which greater sensitivity was seen for musical manipulations in the middle of the range than at the extremes. In Experiment 3, timing and amplitude were manipulated independently; timing variation was found to provide more expressive information than did amplitude. Across all three experiments, listeners demonstrated sensitivity to the expressive cues we manipulated, with sensitivity increasing as a function of musical experience.

*Keywords:* music, music performance, emotion, musical expression

Timing and amplitude variation are fundamental to auditory communication in animals and humans, and are found in calls, speech, and music (Handel, 1993; Moore, 1997). Human music and language use timing and amplitude variation to distinguish different expressive intentions, to communicate emotion, convey particular interpretations, or resolve ambiguities in a written text or score (Gomez & Danuser, 2007; Palmer, 1996; Palmer & Hutchins, 2006). The present study addresses musical expressivity. Here, we use the term "expressivity" to refer to those aspects of a musical performance that are under the control of the performer, and which the performer manipulates for aesthetic and communicative purposes. These may be considered aspects of musical prosody (Bernstein, 1976).

Musical performances are effective at communicating emotion (Juslin & Laukka, 2003; Meyer, 1956), and the unique emotional expressiveness of a particular performance is likely to contribute to people preferring one performance over another. Our starting point is the observation made by the music theorist Leonard Meyer that a musician's deviations from the notated music are critical to delivering an affective aesthetic experience (Meyer, 1956). That is, expressive performance in Western classical music is largely based on systematic variation of duration and intensity (Gabrielsson, 1999; Meyer, 1956; Repp, 1995b), and to some extent, timbre and intonation, depending on the instrument (for a review of music performance research, see Palmer, 1997). Just as actors and orators rarely read a text isochronously and in a monotone, expert musicians rarely play a score as written; instead, they introduce intentional variations in timing, amplitude and timbre (Repp, 1990; Timmers, 2002). Indeed, notes of the same written duration can vary by a factor of two or more during an expressive performance (Levitin & Menon, 2003). Expressivity in musical performance can serve the function of transmitting metrical information to the listener (Drake, Penel, & Bigand, 2000) as well as signaling the presence of a particular emotion such as happiness, sadness, or fearfulness (Juslin & Madison, 1999; Schellenberg, Krysciak, & Campbell, 2000), a judgment listeners are capable of even in music outside their own culture (Balkwill & Thompson, 1999).

The present study explored the relationship between acoustic parameters of a performance (timing and amplitude variation) and psychological parameters (subjective ratings of emotional expressivity). The experiments addressed the following research ques-

This article was published Online First January 24, 2011.

Anjali Bhatara, Anna K. Tirovolas, Lilu Marie Duan, Bianca Levy, and Daniel J. Levitin, Department of Psychology and Center for Interdisciplinary Research in Music and Media Technology (CIRMMT), McGill University.

Anjali Bhatara is now at the Laboratoire Psychologie de la Perception, CNRS, Université Paris Descartes. Anna K. Tirovolas is now at the Department of Educational Counselling Psychology, McGill University. Lilu Marie Duan is now at the Medical School of the University of Nevada. Bianca Levy is now at the Department of Psychology at Boston College.

The research reported herein was submitted in partial fulfillment of the requirements for the Ph.D. in Psychology at McGill University by the first author. The research was funded by grants to D. J. L. from the Social Sciences and Humanities Research Council of Canada, the Natural Sciences and Engineering Research Council of Canada, the National Alliance for Autism Research (now incorporated into Autism Speaks), and the Canadian Foundation for Innovation. We are grateful to Tom Plaunt for his expressive performances of the stimuli used in this experiment; to Bennett Smith, Karle-Philip Zamor, Armen Bazarian, Mansoor Siddiqui, Susan Rogers, Summer Rankin, and Carla Himmelman for technical assistance, to Athena Vouloumanos, Ursula Hess, and Carol Krumhansl for valuable comments, and to Theo Koulis, Ji-Hyun Lee, and Andrew Mattar for statistical help. We also thank Bruno Repp for the idea for Experiment 2 and for comments on previous versions of the manuscript.

Correspondence concerning this article should be addressed to Anjali Bhatara, or Daniel J. Levitin, Department of Psychology, McGill University, 1205 Avenue Dr. Penfield, Montreal, QC H3A 1B1 Canada. E-mail: anjali.bhatara@mail.mcgill.ca or daniel.levitin@mcgill.ca

tions: (1) To what extent do variations in timing and amplitude affect the perception of a performance? (2) What is the nature of the psychophysical function that relates changes in these acoustic parameters (timing and amplitude) to the perception of those changes? (3) Are musicians more sensitive than nonmusicians to such changes? and (4) What are the relative contributions of timing versus amplitude variation? We investigated these questions by creating a set of specially prepared versions of several musical pieces, wherein timing and amplitude information (and hence expressivity) were manipulated. Listeners heard them in random order and rated how emotional they found them to be.

A reasonable null hypothesis for Question 1 (above) is that listeners would be unable to tell the difference between parametric changes in timing and amplitude information. Our principal alternative hypothesis is that listeners will indeed differentiate the different versions. Moreover, we hypothesize that we will be able to recover the rank orderings in *acoustic* variability of the prepared versions from the listeners' *ratings* of expressivity. Support for this comes partly from research showing that listeners can reliably discriminate among performances of a single piano work by different expert pianists (Sloboda & Lehmann, 2001), and that even nonmusicians are adept at recognizing familiar performances among multiple performances of the same piece (Palmer, Jungers, & Jusczyk, 2001). Young infants already show this ability, looking longer at a loudspeaker playing a familiar performance as compared with a loudspeaker playing a novel performance (Palmer et al., 2001). However, previous studies did not systematically control the amount of expressive timing and amplitude variation or collect ratings of expressivity as we have done here.

Understanding the nature of the psychophysical function that connects acoustic differences to psychological ones (Question 2) can advance our understanding of aesthetic preferences for music and why one performance might be preferred to another. In particular, a sigmoidal relationship would suggest that there exist thresholds for detectability of changes.

There are two alternate hypotheses for Question 3, regarding the sensitivity of musicians vs. nonmusicians. One is that musicians will show more sensitivity to expressive cues in performance, while the other is that musicians and nonmusicians will show equal sensitivity. Musicians are better able to perform certain musical tasks such as differentiating between very similar performances (Sundberg, Friberg, & Fryden, 1988) or classifying pitch with interfering timbre changes (and the reverse timbre-classification task; Pitt, 1994). Musicians are also more sensitive to norms of musical expression, some of which include timing and amplitude variation (Sundberg, Friberg, & Fryden, 1991). Additionally, music training provides enhanced perception of the acoustic cues present in vocal emotion expression (Musacchia, Strait, & Kraus, 2008; Strait, Kraus, Skoe, & Ashley, 2009; Thompson, Schellenberg, & Husain, 2004), superior preattentive auditory processing of chords (Koelsch, Schröger, & Tervaniemi, 1999), and electrophysiological experiments show that musicians process the emotional content of music, as indexed by its mode, differently from nonmusicians (Halpern, Martin, & Reed, 2008). Musicians also show structural differences in visual, auditory, and motor brain structures (Gaser & Schlaug, 2003; Pantev et al. 1998). However, even nonmusicians acquire knowledge of musical structure through passive exposure (Krumhansl & Kessler, 1982; Levitin & Tirovolas, 2009; Palmer et al. 2001; Sridharan, Levitin, Chase, Berger, & Menon, 2007), and this extends specifically to their detection of emotions intended by performers (Juslin, 1997), so nonmusicians may show equal sensitivity to this study's expressive cues. Thus, although one could predict either that musicians are more sensitive or equally sensitive to the expressive manipulations in the current experiment, we hope to contribute to the emerging literature on the relation between music perception and musical background or expertise.

The answer to Question 4, relative importance of timing vs. amplitude, is also not immediately obvious; both timing and amplitude variation are important for communication of emotion in music performance (Bernstein, 1976; Clynes, 1983; Juslin & Laukka, 2003). Both of these factors are also important in perception of emotional prosody (Hammerschmidt & Jürgens, 2007; Murray & Arnott, 1993). By varying them separately, we can ascertain the relative contributions of each to the emotional expressiveness of the performance.

Variations in expressive performance are not merely a theoretical construct that is extracted through expertise or mathematical analysis. The composer encodes both structural and expressive ideas in musical notation; performers interpret that notation according to stylistic norms, adding expressive elements (such as varied timing, amplitude, phrasing, etc.) that are handed down through aural tradition (Gabrielsson, 1999; Kendall & Carterette, 1990). Our interest here is that aspect of the model that concerns how listeners interpret those cues jointly arrived at by performers and composers that are intended to communicate emotion to listeners. (Kendall & Carterette, 1990, focused on a similar question, though they examined it from the perspective of the performer: what processes does the performer use to convey his or her ideas to the listener?).

Studying piano performance offers a particularly controlled environment in which to address such questions, because all of the expressive variation in a performance can be characterized by two parameters: duration (timing) and amplitude (or velocity or intensity; Taylor, 1992, p. 127), plus a more subtle parameter involving the position of the foot pedals. "Timing" as referred to in this paper consists of several aspects, including note length, inter-onset intervals, and onset asynchronies. All of these aspects (and the pedaling) will be subject to the same manipulations, described below. "Amplitude" or "intensity" is physically and mechanically equivalent to key velocity in the piano. Note that timbre can also vary in piano performance, but it cannot be reliably manipulated independently of intensity and pedal position (Parncutt & Troup, 2002; Taylor, 1965, p. 175). In the experiments that follow, timing information is encoded as MIDI note onset and MIDI note offset in the computer file, and amplitude information is encoded as MIDI note velocity; each can take on 128 discrete values, and the resolution is considered to be suitably high for judging expert piano performances (Tomassini, 2002).

For the present research, we obtained recordings of expressive performances of standard piano pieces by a concert pianist using a specially equipped recording-reproducing piano (the Yamaha Disklavier). We then systematically reduced the expressive variation by editing the resulting MIDI computer file. We next played the original and modified performances back through the same acoustic piano to obtain expressivity judgments from human listeners.

## Experiment 1

### Method

**Participants.** The participants were 16 adults (9 women, 7 men) between the ages of 19 and 36 (mean age 23.1, *SD* = 4.9), recruited from McGill University and the surrounding community. Participants received either course credit or a $20 gift card to a CD store. (This reimbursement was for 2 hr of their time because they participated in 3 additional studies not related to the present study; the order of all experiments was randomized across subjects.) The background of participants spanned a wide range of musical experience: 6 reported 0 or 1 year of musical training, 4 reported between 3 and 5 years of musical training, and 6 reported 8 or more years of musical training. The mean was 5.9 years of training (*SD* = 7.0).

**Stimuli.** The stimuli were six versions of short (~30 s) excerpts from each of four Chopin nocturnes (Op. 15 No. 1 and Op. 32 No. 1, both in major keys, and Op. 55 No. 1 and KK IVa No. 16, both in minor keys). We selected these pieces after consulting with the head of the piano performance division of the Schulich School of Music at McGill University, Professor Thomas Plaunt. In recommending them, he noted that they are relatively simple melodically and rhythmically (facilitating comparison across versions), known to most professional pianists (hence allowing for additional performances in future studies), and offer ample opportunity for performers to provide expressive interpretation.

A professional pianist played these four pieces using "normal expressivity, as one would in a concert performance," and we recorded his performances of each of these pieces on a Yamaha Disklavier piano (Buena Park, California, Model MPX1Z 5959089, equipped with a DKC500RW MIDI control module), with the output saved as a MIDI file. Then, using a MIDI editor (ProTools 7, Avid, Daly City, CA) we parametrically altered the performances so as to remove some or all temporal and dynamic features associated with expressivity as described below. The expressive versions of the four nocturnes were independently judged by a panel of symphony orchestra conductors and musicians to be of high aesthetic quality. Audio examples of the stimuli as well as graphs detailing the stimulus manipulation can be found at http://ego.psych.mcgill.ca/labs/levitin/expressivity.htm.

**Temporal expressivity.** To manipulate temporal expressivity, we first created a version of the piece in which all the temporal variation (and hence temporal expressivity) was removed: the *mechanical* version. We next used interpolation to create intermediate versions between this mechanical version and the original, unaltered performance (the *expressive* version). This was accomplished in the following steps:

1. To obtain the *mechanical* version we removed all expressive temporal variation from the recorded performance by editing the MIDI file using the program ProTools (Avid) so that this new version conformed to the musical score and composer's rhythmic marking. To do so, we divided the duration of the performance (in seconds) by the number of beats in the written score to obtain an average tempo and thus an average value for each quarter note. We then set every note to its nominal duration (eighth notes were exactly half the length of quarter notes, half notes were exactly twice the length, etc.). The note onset times were adjusted to be immediately after the end of the previous note, creating a "legato" feel appropriate for the piece.

2. Intermediate versions were created using linear interpolation to obtain 75, 50, or 25% of the temporal variance of the expressive version. For example, to create the 50% version, we assigned each event a duration that was halfway between its duration in the expressive (original) version and the mechanical version. The note onset times were altered in the same way, with values calculated from the onset of the previous note, referred to as the inter-onset interval (IOI). We used linear interpolation to create IOIs that were between the original and the mechanical version. We chose linear interpolation rather than a nonlinear or threshold function because it was not known a priori how the psychological construct of expressivity is perceived in relation to the parameters we were altering; this is one of the aims of the present experiment. Therefore, in the absence of a compelling reason to do otherwise, we chose the more straightforward linear interpolation for the sake of parsimony, and rather than risking a complex function that may have clouded our results.

**Dynamic expressivity.** We altered the piece's dynamic expressivity in the same fashion as above. A mechanical version was created by assigning to each note the mean MIDI velocity (the portion of the MIDI signal that determines amplitude) of the expressive version. The expressive version contains virtually full amplitude variation (limited only by the 128 levels available in MIDI). For the three intermediate versions, we assigned 75, 50, or 25% of the amplitude variation contained in the expressive version, again using linear interpolation.

**Pedaling.** We altered the pedaling in the same fashion as the timing and dynamic expressivity, with one exception (*mechanical*) that is discussed below. We assigned 100, 75, 50, and 25% of the pedaling values in their respective conditions. Pedaling values referred to the height of the pedal; "0" signifies a pedal that is at its topmost, resting position while "127" signifies a fully depressed pedal. The exception for the mechanical version came about because during the original performance, the pianist used some pedal nearly all the time, and this served to create de facto note durations that were not captured by the MIDI file; in other words, the performer may have lifted his finger from a key while the note continued to sound because of pedaling. When we created the mechanical version with no pedaling at all, these note durations were altered in a way that noticeably distorted the performance. Moreover, the subjective impression of the experimenters was that the version sounded qualitatively different from the others: lacking legato, it sounded too "staccato" (choppy), and this would have caused it stand out rather than sounding as though it were simply one point along a continuum. We thus assigned 25% of the pedaling value to the mechanical version.

**Additional control.** We also created a *random* condition as an additional control. It is possible that some participants might base their judgments on the overall variability of the performance, or, in information-theoretic terms, the amount of *information* (Pierce, 1961/1980; Shannon, 1948). That is, the expressive version of the piece always contains greater variability in both timing and amplitude when compared with the altered versions. A random version of each piece was therefore created by reassigning all of the note durations of the original performance randomly within note type groups: eighth notes' durations were rearranged only among eighth notes, quarter notes among quarter notes, etc. The silent interval between notes was randomized within groups of consecutive notes of the same type. The MIDI velocities of each note were also randomly reassigned, though not restricted to the same note type group. The pedaling profile was the same for that of the expressive version; introducing random pedaling was deemed to be outside the scope of our study, which focuses principally on amplitude and timing. The result of all of these manipulations was that this version contained the same amount of temporal and amplitude variability as the expressive version but did not make musically expressive sense.[1] The experiment thus used six categories of expressiveness (*expressive*, 75%, 50%, 25%, *mechanical*, and *random*) for each of the four pieces, resulting in a total of 24 stimuli.

For presentation to the participants, we made high quality digital recordings of the stimuli. To prepare the stimuli, the MIDI data were played back through the Disklavier, and the acoustic output was recorded in stereo (using a matched pair of Neumann U87 microphones, a GML microphone pre-amp, and ProTools 6) and saved as digital .wav files.

**Dependent measure.** Our research question concerned how expressivity is conveyed through musical performance. We conducted a series of pilot tests to determine if participants understood the term "expressivity" without our defining it, and furthermore, to better understand if different ways of asking the question might yield different results. Fifty participants, recruited from the same subject pool we would later use for the actual experiment (but who were not used in the later experiment), were randomly assigned to one of five conditions that varied only by question. In this between-subjects design, participants listened to all of the versions of two of the stimuli used in Experiment 1 (Op. 32, major and Op. 55, minor), in random order. The pilot followed all of the procedures of Experiment 1, except for changes in how the question was worded:

1. How emotional was the performance you just heard?
2. How expressive was the performance you just heard?
3. How much emotion do you think the performer was feeling in this passage?
4. How much emotion did this performance make you feel?
5. How musical was the performance you just heard?

We performed a repeated measures ANOVA with question type as the between-subjects factor, and it was not significant, $F(4, 45) = 1.56$, $p = .20$. We also performed a two-way repeated measures ANOVA with question type as the between-subjects factor and expressivity level as the within-subjects factor to examine the interaction between these two factors. Mauchly's Test of Sphericity failed ($W = .252$, $\chi^2 = 60.065$, $p < .001$), so we performed a Huynh-Feldt correction on epsilon for the subsequent *F* test. The results showed that there was no significant interaction between the question types and expressivity level, $F(3, 11) = 1.89$, $p = .2$. We decided to use Question 1 in the experiments because it best reflected the way we thought of the research question.

**Procedure.** Two blocks of trials were created, with each stimulus appearing in random order in each block; thus, participants heard each stimulus twice, once in each block. The blocks were separated by a 30 s silent rest period. Stimuli were played back through a Macintosh PowerBook G4 laptop (Apple Computer, Cupertino, CA) and controlled by the program Psiexp (for the graphics, Smith, 1995) and MaxMSPRunTime (for the sounds, Cycling 74/IRCAM, 2005). Participants were tested individually and listened through loudspeakers (Acoustic Research, Model 570, Hauppauge, NY) at a comfortable volume level. Stimuli were presented at 73 dB(A) ± 2 dB for all participants.

Participants were instructed to rate how emotional the music was. We emphasized that it did not matter which emotion they perceived in the performance; we wanted them to tell us how much emotion the performance conveyed. After hearing each stimulus, participants saw the question "How emotional was the music you just heard?" displayed on the computer screen along with a graphical slider, and they rated the emotional level by using the computer mouse to move the slider along a continuous scale, of which one end was labeled "not at all" and the other end was labeled "very emotional." (The responses were coded as ranging between 0 and 1 to mirror the 0 to 100% expressivity levels.) Participants were asked to use the whole range of the scale.

## Results

One participant was excluded from the analysis because his ratings showed the opposite pattern from all other participants (*mechanical* was rated as the most emotional and *expressive* was rated as least emotional). The independent variables in this analysis were expressivity level (*expressive*, 75, 50, 25%, *mechanical*, and *random*) and tonality (major or minor key). The dependent variable was the subjects' ratings; means of the ratings by expressivity level are shown in Figure 1.

We first performed a one-way repeated measures ANOVA to test for the effect of repetition (recall that each stimulus was presented twice). Our a priori hypothesis was that repetition should have no effect, and this was confirmed: $F(1, 14) = .12$, $p = .73$. In addition, reliability of the measure was assessed by correlations between the first and second presentations of the stimuli. The correlations from the 50 and 75% expressivity levels were marginally significant, $r(16) = 0.38$, $p = .07$. The correlations for all other expressivity levels were greater than 0.61 (all *p* values $<$ 0.01). Therefore, in subsequent analyses we ignored repetition as an experimental factor.

[1] In expressive performances, the tempo and amplitude variations imparted to the piece by the performer are not random–they are manifestations of a coherent musical plan, and performers can replicate prior performances with great precision (Juslin, 2001). Therefore, randomly permuting the amplitude and timing should not result in an emotionally meaningful performance, but should create a performance that was superficially equivalent, in terms of structural variability.

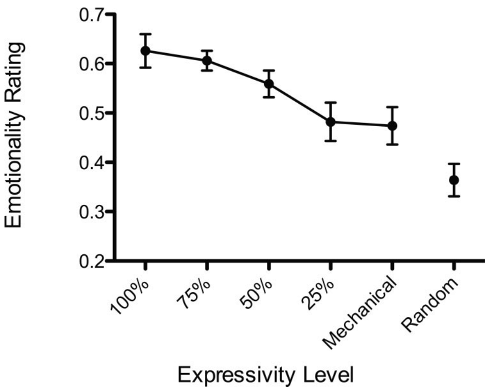

*Figure 1.* Experiment 1: Mean ratings of emotionality by tonality across expressivity levels. Error bars indicate ± *SEM*.

We performed a two-way repeated measures ANOVA on the participants' responses, with expressivity level and tonality as factors. Expressivity level, $F(5, 70) = 12.54$, $p < .001$, and tonality, $F(1, 14) = 16.89$, $p < .001$, were both significant. The main effect of tonality occurred because participants rated the minor nocturnes on average as being more emotionally expressive than the major ones. The effect sizes (as measured by $\eta^2$; see Levine & Hullett, 2002 for a discussion of $\eta^2$ vs. partial $\eta^2$) were 0.32 for expressivity level and 0.13 for tonality, demonstrating that expressivity level accounted for a much larger proportion of the variance than did tonality. The interaction between expressivity level and tonality was marginally significant, $F(5, 70) = 2.12$, $p = .07$.

We performed post-hoc comparisons (Tukey's HSD) to determine which mean ratings were significantly different from one another (at $p < .05$), and we found one adjacent pair that showed significance: 50% was rated significantly higher than 25%. In addition, the *random* condition was significantly different (rated lower) from all levels, and *mechanical* was rated significantly lower than *expressive*, 75, and 50%. In addition, the ratings for *expressive* were significantly greater than 50 and 25% and the ratings for 75% were significantly greater than 25%. These comparisons are summarized in Table 1. Examination of the raw values of these means confirms that the ratings were monotonically decreasing from *expressive* to *mechanical* and to *random*, indicating that the rank ordering of the acoustic versions was entirely recovered by the participants' ratings.

Table 1
*Experiment 1: Difference in Mean Ratings Between Expressivity Levels*

| | 100% | 75% | 50% | 25% | 0% | Random |
|---|---|---|---|---|---|---|
| 100% | X | .02 | .07$^*$ | .14$^*$ | .15$^*$ | .26$^*$ |
| 75% | | X | .05 | .12$^*$ | .13$^*$ | .24$^*$ |
| 50% | | | X | .08$^*$ | .09$^*$ | .20$^*$ |
| 25% | | | | X | .01 | .12$^*$ |
| 0% | | | | | X | .11$^*$ |

$^*$ $p < .05$.

To characterize the nature of the psychophysical curve we obtained, we fit the data with linear, polynomial, and the a priori hypothesized sigmoidal (ogive) models. The sigmoid made the best fit (Sum of Squared Residuals, SSR = .007) by an order of magnitude better than the next best fit, the line (SSR = .045), $F(2, 6) = 6.42$, $p < .05$. This indicates that judgments followed a psychophysical threshold function and is consistent with the Tukey's HSD test, finding that the only significant difference in adjacent values was between values in the middle of the curve (50 vs. 25%).

Finally, we tested whether a regression line fit to the observed values had a significant downward slope from 100% to 0%, that is, whether the slope was a significant departure from 0 (a 0 slope would have occurred if, even though the values were monotonically decreasing, and fit an ogive, they were decreasing in a trivial fashion). To test this, we used a planned orthogonal contrast, and confirmed that the slope of the line is significantly different from 0 and pointing downward, $F(1, 14) = 8.09$, $p < .02$.

**Ratings as a function of musical experience.** Before the following analysis, we excluded one additional participant whose high level of musical experience was greater than two standard deviations from the mean, to prevent any musical experience effects being driven by this one participant. We performed a linear regression for each of the remaining participants of their ratings by the five linearly scaled expressivity levels (excluding *random*). The β values resulting from these regressions can be thought of as "sensitivity to expressivity differences" because they indicate the slope of the line resulting from emotionality ratings spanning *expressive* to *mechanical*. We correlated these values with each participant's years of musical experience. A scatterplot of these values is shown in Figure 2, and the correlation was 0.59 ($p < .001$). Thus, the ratings by people with more musical experience had steeper slopes, indicating that these individuals were more sensitive to differing levels of expressivity. The excluded participant with a large amount of musical experience fit with this pattern, showing a β value of .91, while the other participants' values ranged from .02 to .54.

## Discussion

Experiment 1 demonstrated that people with varying levels of musical experience were sensitive to differences in expressivity as

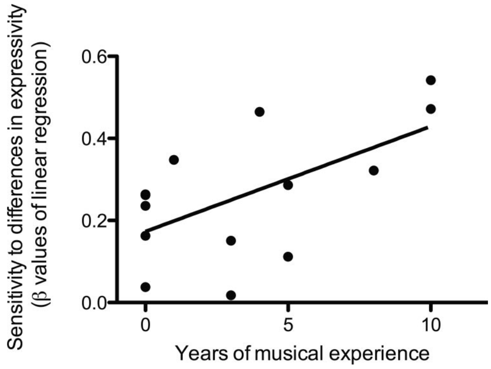

*Figure 2.* Experiment 1: Scatterplot of sensitivity to expressivity manipulations (individual β values from linear regressions of ratings on expressivity level) by years of musical experience.

conveyed by timing and amplitude variation in these performances of Chopin nocturnes. The lower emotionality ratings of the *random* condition indicates that they are not simply responding to the overall complexity or information content of the different versions, and that listeners are sensitive to the correct placement of timing and amplitude variations.

We found that greater musical experience correlated with greater sensitivity to the manipulations, as indexed by the steeper slope for the participants with greater musical experience. An additional factor contributing to the participants' ratings was the tonality of the pieces. Two were in a major key and two were in a minor key, and the minor nocturnes were rated as more emotional than the major ones. However, one should be cautious about overgeneralizing in attributing these differences only to tonality; a number of features differed among the nocturnes. One could artificially create minor versions of major nocturnes (or vice-versa) by manipulating the melody and harmony to better control and thus more rigorously examine these differences. However, doing so would also alter the pieces significantly from the composer's original intent and take them very far from ecological validity. Nevertheless, the observed differences could be due, at least in part, to the general association of major tonality with happiness or calm and of minor tonality with sadness; if participants are already in a generally calm, happy state, they may have interpreted the "sad" pieces as being more emotional than the "happy" ones because the sadness offers a more contrasting emotion to their state at the time. In addition, brain imaging evidence shows greater activation of the amygdala during passive listening to minor chords than to major chords in both musicians and nonmusicians (Pallesen et al., 2005). Alternatively, it may be because of qualities unique to these particular nocturnes or participants. To generalize, a much larger sample of excerpts would be needed, or the mode of the same excerpt would have to be manipulated.

In this experiment, our highest level of expressivity in Experiment 1 can be thought of as arbitrary. That is, these amplitude and timing values were obtained from a particular performer on a particular day, and presumably would vary across time both within and across performers. The MIDI and Disklavier technology permits us to extrapolate the expressive variation to extend past the particular 100% expressive performance we obtained. Thus, in Experiment 2 we sought to answer the question of whether or not 100% expressive constitutes an actual "top end" or upper boundary of musical expression for these particular pieces, given the overarching stylistic and interpretive approach taken by our pianist.

## Experiment 2

### Method

**Participants.** The participants were 11 adults (8 women, 3 men) between the ages of 18 and 29 (mean age 21.9, *SD* = 3.3), recruited from McGill University and the surrounding community. Participants received \$5 for ~15 min of their time. Participants had an average of 6.4 years of musical training (*SD* = 4.7).

**Stimuli.** To reduce participant fatigue, participants heard a subset of the nocturnes used in Experiment 1 (Op. 32, in a major key, and Op. 55, in a minor key). We extrapolated performance parameters from the expressive version to systematically add variability in timing and amplitude to the original performance, resulting in 2 new conditions: 125 and 150% expressivity. The original conditions from Experiment 1 were also presented, with the exception of *random*, resulting in 7 levels of expressivity for each of the two nocturnes, and a total of 14 stimuli. The independent variables for this experiment were, as in Experiment 1, expressivity level (150%, 125%, *expressive* [100%], 75%, 50%, 25%, and *mechanical* [0%]) and tonality (major or minor key).

**Procedure.** Participants heard each expressivity level of each nocturne once, in random order, and rated the emotional expressivity of each excerpt on the same slider scale employed in Experiment 1. The remainder of the procedure was identical to Experiment 1.

### Results

One participant had to be excluded from the analysis because he did not follow instructions. The mean of the participants' ratings by expressivity level can be seen in Figure 3. As in Experiment 1, the 100% condition was rated higher than 75, 50, and 25%, following a downward trend to *mechanical*, $F(1, 147) = 31.7, p < .001$.

We performed a two-way repeated measures ANOVA on the participants' responses, with expressivity level and tonality as factors. Expressivity level, $F(6, 54) = 13.16, p < .001$ and tonality, $F(1, 9) = 29.41, p < .001$ were both significant; participants rated major nocturnes on average as being more emotional than the minor ones. The effect sizes were $\eta^2 = 0.31$ for expressivity level and $\eta^2 = 0.19$ for tonality, demonstrating that expressivity level accounted for a larger proportion of the variance than did tonality. The interaction between expressivity level and tonality was also significant, $F(6, 54) = 3.97, p < .01$, with $\eta^2 = 0.07$, with the major nocturne rated as more emotionally expressive than the minor one at the higher levels of expressivity (150, 125, 100%) but not rated differently from the minor at the lower levels of expressivity (see Figure 4).

Tukey's HSD post hoc comparisons are summarized in Table 2. There were no pairs of adjacent expressivity levels for which mean ratings differ, but most other combinations of expressivity levels differed from each other. An exception to this, important for this

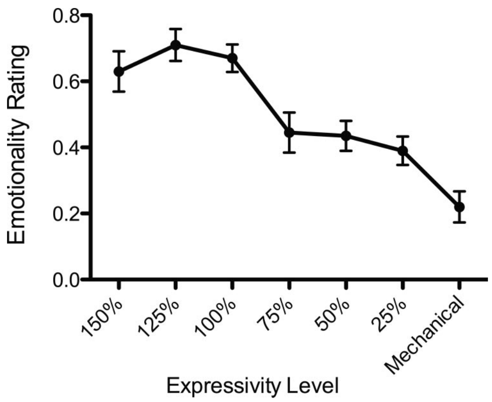

*Figure 3.* Experiment 2: Mean ratings of emotional expressiveness across expressivity levels. Error bars indicate ± *SEM*.

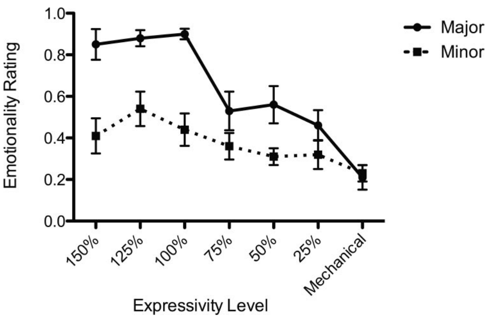

*Figure 4.* Experiment 2: Mean ratings of emotional expressiveness across expressivity levels by tonality. Error bars indicate ± *SEM*.

study, was that the 100, 125, and 150% versions did not differ significantly from each other.

To characterize the nature of the psychophysical curve we obtained, we fit the data as in Experiment 1 with linear, polynomial, and sigmoidal models. The sigmoid again made the best fit (SSR = .09) compared to the line (SSR = .12). Although the difference did not reach statistical significance $F(2, 6) = 1.33$, $p = .33$, it still represents an improvement in quality of fit.

## Discussion

The sigmoidal shape and lack of difference between ratings for the 100, 125, and 150% conditions suggests that there is a decreased ability to perceive changes in expressive nuances beyond 100%, "concert-level" expressivity. This implies that the amount of expressivity the performer chose to add to these pieces may well have been at the optimal level, beyond which extra expression (as implemented by increased variation in amplitude and timing) would not be perceived or preferred by the listener.

It is important to reiterate that the present results are from two performances by one performer, and may not be generalizable. Further work will be required to examine "beyond expressive" performances of other pieces.

In Experiments 1 and 2, timing and amplitude variation varied with each other, as they do naturally in real-life performance. In Experiment 3, we sought to examine the relative contributions of each to the listeners' judgments. Additional motivation comes from Todd's (1992) model of expressive dynamics, which emphasized the need for such data. Further, Juslin & Madison (1999) found that removing only timing variations impaired listeners' judgment of *which* emotion was being expressed in piano pieces, but to a lesser extent than did removing both timing and amplitude variation. However, these authors did not remove amplitude variation while leaving the timing variation intact, so it is unclear which of these was the more important dimension for listeners judging which emotion was being expressed.

In Experiment 3, we thus sought to compare the relative effects of timing variation and amplitude variation, again as measured by participants' ratings of the performances' emotionality. We also introduced an additional modification to the design. Based on the findings of Experiments 1 and 2, there appeared to be nonlinearities in the mapping between acoustic changes and their psychological representations. We were especially interested to learn more about the psychometric function at the upper expressivity levels (between 75 and 100% expressive) to test an initial hypothesis that musicians would be more sensitive in this region because it represents the most likely portion of the curve in which they themselves attempt to learn to convey expressivity. To explore this, without making the experiment longer (which could tax our participants' attention), we omitted the 25% condition and replaced it with a condition halfway between 75 and 100%, the 87.5% condition.

## Experiment 3

### Method

**Participants.** The participants were 20 adults (16 women, 4 men) between the ages of 18 and 33 (mean age 22, *SD* = 3.8), recruited from McGill University and the surrounding community. Participants received $5 for a half hour of their time. Ten were experienced musicians with 8 or more years of training on a musical instrument, and 10 were nonmusicians with less than 1 year of musical training. None had participated in the previous experiments.

**Stimuli.** The stimuli were two of the Chopin's nocturnes used in Experiment 1: Op. 15, No. 1 (major) and KK IVa, No. 16 (minor). As in Experiment 2, we used only two of the nocturnes, one of each tonality, to keep the experimental session to a reasonable length and prevent participant fatigue. Because each participant heard both conditions in this within-subjects design (as described below), the experiment length would have been double that of Experiment 1.

The stimuli were prepared in the same manner as in Experiment 1, except that the timing and amplitude manipulations were kept distinct, one value was varied while the other was held constant. For the timing-varied stimuli, six versions were created for each of the two nocturnes (100, 87.5, 75, 50, 0% or *mechanical*, and *random*) with timing information interpolated between 0 and 100% as before, while the amplitude was held constant at its mean value for the piece. Thus, the amplitude profile for all six timing-varied stimuli was identical to the amplitude profile for the *mechanical* version, and the timing information was allowed to vary. The amplitude-varied versions were created in an analogous fashion, with the timing held constant across versions. Note that the *mechanical* versions of the timing-varied and the amplitude-varied stimuli were necessarily identical to one another as neither had any note variability, apart from what was written in the score. Conse-

Table 2
*Experiment 2: Difference in Mean Ratings Between Expressivity Levels*

| | 150% | 125% | 100% | 75% | 50% | 25% | 0% |
|---|---|---|---|---|---|---|---|
| 150% | X | −.08 | −.04 | .19* | .20* | .24* | .41* |
| 125% | | X | .04 | .27* | .28* | .32* | .49* |
| 100% | | | X | .23* | .24* | .28* | .45* |
| 75% | | | | X | .01 | .06 | .23* |
| 50% | | | | | X | .05 | .22 |
| 25% | | | | | | X | .17 |

* $p < .05$.

quently, these conditions were combined for the statistical analyses reported below.

As an alternative to this, we had considered varying one parameter while keeping the other unaltered from the original performance. In this case, the amount of timing information would have varied while the amplitude information was fully expressive, and vice versa. However, we wished to examine the separate contributions of each of these two parameters, with as little influence as possible from the other. Therefore, we decided to instead minimize the variability in the parameter not of interest and keep the amplitude at its mean (for the timing-varied condition) or the timing at its written values (for the amplitude-varied condition).

**Pedaling.** For both the timing- and amplitude-varied stimuli, the pedaling was altered according to the same procedure as Experiments 1 and 2. Thus, there was no difference in pedaling variability between the two conditions, though there were differences among expressivity levels within each condition.

**Procedure.** Each participant heard all of the stimuli. Two blocks (Varied Timing and Varied Amplitude) were presented in counterbalanced order, and the order of the stimuli was randomized within each block. The *mechanical* version was the same in both blocks (see *Stimuli* above). Thus, the participants heard a total of 12 versions of each nocturne (*mechanical*, 50%-T, 75%-T, 87.5%-T, *expressive*-T, and *random*-T for the Varied Timing block and *mechanical,* 50%-A, 75%-A, 87.5%-A, *expressive*-A and *random*-A for Varied Amplitude block), with 11 of the 12 being distinct from the others. Each of the 12 versions of each of the 2 nocturnes was presented 2 times for a total of 48 trials.

The testing procedure was the same as in Experiment 1 with two exceptions. First, the participants heard the stimuli over AKG 240 headphones at 73 dB ± 2 dB and not over loudspeakers because of construction noise near the laboratory. We confirmed that the headphones eliminated this distraction without interfering with the quality of judgments: pilot testing confirmed that judgments made with headphones and speakers were not statistically different. Second, to help participants remember more easily where on the scale their previous ratings fell, we included numbers above the tick marks on the slider scale; 1 meant "not at all emotional" and 7 meant "very emotional." This was added in the hopes of increasing consistency of participants' ratings. If, during the experiment, they were able to remember where they had previously rated an excerpt similar to the currently presented one, perhaps they would find it easier to rate that current one. As before, participants were informed that they could place the slider anywhere, between or on numbers, and they should try to use the entire range of the slider. The slider settings were again converted to ratings ranging from 0 to 1.

## Results

**Timing variations.** The grand mean of ratings was 0.54 (*SE* = 0.06), again around the center of the scale (which ranged from 0 to 1). Individual participants' means ranged from 0.41 to 0.76 (*SD* = 0.1). The mean ratings are shown as a function of stimulus conditions in Figure 5. The independent variables for this analysis were expressivity level, tonality, and musical experience.

We performed a 3-way repeated measures ANOVA on the participants' responses in the Varied Timing condition, with expressivity level and tonality as within-subject factors and musical experience as a between-subject factor. Expressivity level, $F(5, 90) = 18.88$, $p < .001$, tonality, $F(1, 18) = 8.31$, $p < .01$, and musical experience, $F(1, 18) = 13.99$, $p < .001$, were all significant factors. The effect sizes as measured by $\eta^2$ were 0.32, 0.03, and 0.01, respectively, indicating as in Experiment 1 that, though statistically significant, musical experience and tonality together accounted for a small proportion of the variance in the results, while expressivity level accounted for a large proportion of the variance. The interaction between tonality and expressivity level was also significant, $F(5, 90) = 6.26$, $p < .001$ ($\eta^2 = .06$), as was the interaction between tonality and musical experience $F(1, 18) = 7.92$, $p < .01$, ($\eta^2 = .04$).

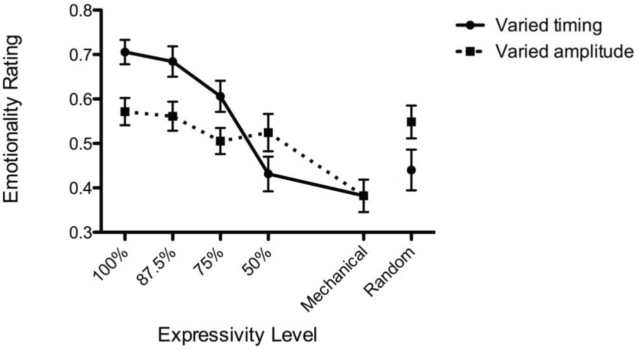

*Figure 5.* Experiment 3: Mean ratings of emotionality for Varied Amplitude and Varied Timing conditions across expressivity levels. Error bars indicate ± *SEM*.

Tukey's HSD post hoc comparison among expressivity levels revealed that *random-T, mechanical* and 50%-T were not significantly different, and that 75%-T, 87.5%-T, and *expressive-T* were not significantly different from each other, though the ratings for these two groups of levels differed from each other (see Table 3). The ratings were again monotonically decreasing, and a planned linear contrast confirmed that the slope of the line connecting the points is significantly different from zero, $F(1, 19) = 55.31$, $p < .001$. Additionally, as in Experiment 1, the random version was rated less emotional than the expressive version.

To characterize the nature of the psychophysical curve we obtained, we fit the data as in Experiments 1 and 2 with linear, polynomial, and sigmoidal models. The sigmoid again made the best fit (SSR = .018) compared to the line (SSR = .021), although the difference did not reach statistical significance $F(2, 6) = 1.17$, $p = .37$, it still represents an improvement in quality of fit.

The significant main effect of musical experience showed that overall, musicians rated all of the stimuli as more emotional than did the nonmusicians, but there was no interaction between musical experience and expressivity level, indicating that they followed a similar pattern of ratings for each expressivity level. The significant main effect of tonality arose because the nocturne in a minor key (KK IVa) was generally rated as more emotional than the nocturne in a major key (Op. 15). The interaction between tonality and expressivity level is illustrated in Figure 6a.

Tukey's HSD post hoc test showed that minor was rated as more emotional than major for 100%-T and 87.5%-T, but there were no significant differences between the two for the remaining expressivity levels. To investigate the interaction between tonality and

Table 3
*Experiment 3: Varied Timing, Difference in Mean Ratings Between Expressivity Levels*

| | 100% T | 87.5% T | 75% T | 50% T | 0% | Random T |
|---|---|---|---|---|---|---|
| 100% T | X | .02 | .10 | .28* | .35* | .27* |
| 87.5% T | | X | .08 | .25* | .32* | .24* |
| 75% T | | | X | .18* | .25* | .17* |
| 50% T | | | | X | .07 | −.01 |
| 0% | | | | | X | −.08 |

* $p < .05$.

musical experience we again performed Tukey's HSD post hoc test. We found that nonmusicians rated the minor nocturne as more emotional than the major one, but the musicians did not rate them as emotionally different. This is shown in Figure 6b.

**Amplitude variations.** The grand mean of ratings was 0.52 ($SE = 0.04$). Individual participants' means ranged from 0.22 to 0.65 ($SD = 0.1$). The independent variables for this analysis were expressivity level, tonality, and musical experience. We performed a three-way repeated measures ANOVA on the participants' responses from the Varied Amplitude condition, with expressivity level and tonality as within-subject factors and musical experience as a between-subject factor. Only expressivity level, $F(5, 90) = 4.73$, $p < .001$, had a significant main effect with an effect size $\eta^2 = 0.11$. No main effects were found for tonality, $F(1, 18) = .50$, $p = .49$, or musical experience $F(1, 18) = .67$, $p = .44$, but there was a significant interaction between tonality and musical experience, $F(1, 18) = 4.80$, $p < .05$, $\eta^2 = 0.05$. This arose because nonmusicians generally rated the minor nocturne as more emotional than the major one, but musicians tended to rate the major one as more emotional than the minor (see Figure 7). However, Tukey's HSD post hoc test did not reveal a significant difference between mean ratings for either musicians or nonmusicians.

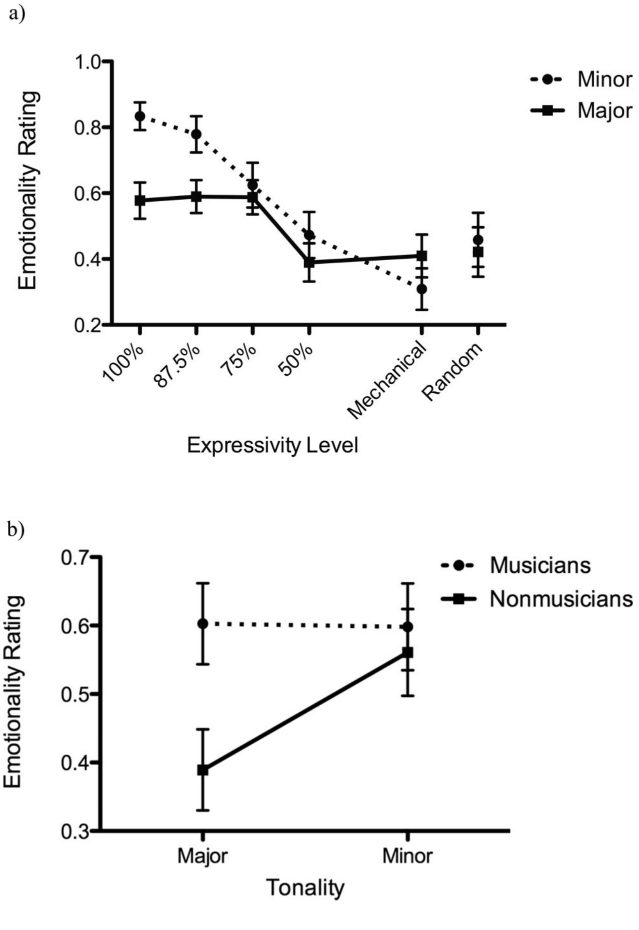

*Figure 6.* Experiment 3: (a) Mean ratings of emotionality for Varied Timing condition across expressivity levels by musical experience and (b) Mean ratings of emotionality for Varied Timing condition across musical experience by tonality. Error bars indicate ± *SEM*.

A Tukey's HSD post hoc comparison investigating the main effect of expressivity level revealed a significant difference between *mechanical* and 50%-A. No other pairs of adjacent expressivity levels were significantly different. The *random-A* version was significantly different from (higher than) *mechanical*, but not significantly different from any others. In addition, *mechanical* was significantly different from both 87.5%-A and *expressive-A* (see Table 4). The ratings for Varied Amplitude were not monotonic, as shown by the slight rise in the graph for the 50%, but this was not statistically significant. Even with this bump, the slope of the regression line is downward and significantly different from 0 by planned linear contrast, $F(1, 19) = 21.23$, $p < .001$.

**Comparing timing and amplitude variations.** Given that one of the original purposes of this study was to determine whether timing or amplitude would have a greater effect on perceived emotionality, the effect sizes (as indicated by $\eta^2$ values) for expressivity level in the two blocks were compared. In Varied Timing, expressivity level had an $\eta^2$ of .32 (.43 for musicians and .25 for nonmusicians), while in Varied Amplitude the $\eta^2$ was .11 (.11 for musicians and .13 for nonmusicians). Musicians tended to give higher ratings than nonmusicians when timing was varied, but there was no significant difference between groups when amplitude was varied.

The means for Varied Timing ratings ranged from .38 (on a scale of 0 to 1) for *mechanical* to .71 for *expressive-T*. For Varied Amplitude, the mean ratings ranged from .38 for *mechanical* to .57 for *expressive-A*.

For both conditions, there is a decreasing linear trend from *expressive* to *mechanical*, but the slope of a line fitted to the data

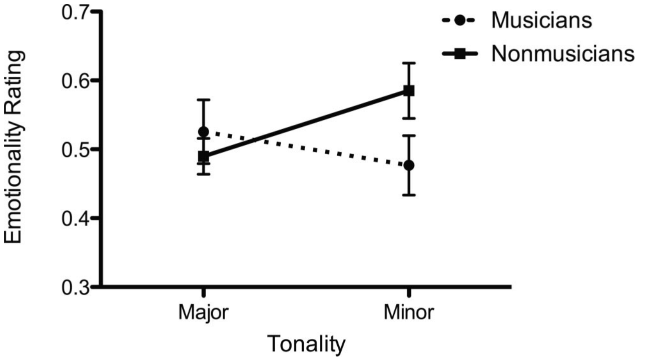

*Figure 7.* Mean ratings of emotionality for Varied Amplitude condition across musical experience by tonality. Error bars indicate ± *SEM*.

Table 4
*Experiment 3: Varied Amplitude, Difference in Mean Ratings Between Expressivity Levels*

| | 100% A | 87.5% A | 75% A | 50% A | 0% | Random A |
|---|---|---|---|---|---|---|
| 100% A | X | .01 | .07 | .05 | $.17^{*}$ | .02 |
| 87.5% A | | X | .06 | .04 | $.16^{*}$ | .01 |
| 75% A | | | X | .18 | .25 | .17 |
| 50% A | | | | X | $.07^{*}$ | −.01 |
| 0% | | | | | X | $-.08^{*}$ |

$^{*}$ $p < .05$.

is steeper for the Varied Timing condition (y = −.05) than the Varied Amplitude (y = −.02). *Random-T* is rated much less emotional than *expressive-T*, but there is no significant difference between *random-A* and *expressive-A*.

## Discussion

Experiment 3 showed that timing and amplitude variations both affect emotionality judgments. Generally, more systematic (but not random) variation in timing or amplitude translates to greater subjective ratings of emotionality. Though amplitude variation in music has not been previously studied in isolation, its importance as an expressive factor on its own is not surprising given that performers normally change amplitude (typically together with timing), playing more loudly as they speed up and more softly as they slow down (Clarke, 1999; Repp, 1996). Based on implicit learning over many years of exposure to music, listeners would come to associate both timing and amplitude variation with increased emotionality. In addition, Todd (1992) predicted that amplitude variation would be perceived similarly to timing variation. Thus, experienced listeners should be sensitive to the amplitude variations in a similar manner as they are to the timing variations.

This does not mean that timing and amplitude variation are equally important to the perception of emotion in a performance. To compare the two manipulations, we looked at the effect size (as indexed by $\eta^2$) for expressivity level and found that it was larger in the Varied Timing condition than in the Varied Amplitude condition, suggesting that timing variations alone are more effective in communicating emotion than amplitude variations alone. Indeed, the ratings were higher overall for the Varied Timing condition, and there were bigger differences between each successive expressivity level in the Varied Timing condition than in the Varied Amplitude condition. This suggests that, in the absence of timing variations, expressively motivated variations in amplitude are limited in their ability to convey emotion.

One possible limitation of the present experiment is that performances we obtained might have had less amplitude variation than is typically present in piano performances, causing the influence of amplitude variation on ratings of emotionality to be smaller. We attempted to minimize this possibility by using a professional pianist, and our own subjective evaluation was that the performance was fully expressive in both dimensions of amplitude and timing. Indeed, the range of amplitude values captured in the MIDI file was near the maximum allowable. Further research utilizing other performances will be necessary to generalize our findings.

As in Experiment 1, listeners were sensitive to the different quantities of variations in amplitude and timing that were used, but there was not a one-to-one correspondence between the physical amounts of variation and the psychological translation into emotionality. There was a linear decrease in variation in the physical parameters we manipulated from the *expressive* down through the *mechanical* levels, but the participants' ratings of the stimuli did not follow this linear pattern, showing possible threshold effects at the top of the range in Experiments 2 and 3, and at the bottom of the range in Experiments 1 and 3 (and the suggestion of a bottom-of-range effect in Experiment 2 as well, yielding a sigmoid-shaped curve).

A difference between this experiment and Experiments 1 and 2 is that participants heard the stimuli through headphones instead of from a free-field speaker. It is unlikely that this significantly affected our data. Across a large number of unrelated, independent experiments, involving a range of listeners and material, we have found no difference in listening judgments obtained from using headphones versus loudspeakers, provided that both are of sufficiently high quality, such as those used in the present experiments (Guastavino, 2007; Pras, Zimmerman, Levitin, & Guastavino, 2009; Salimpoor, Guastavino, & Levitin, 2007).

An interesting difference between the Varied Timing and Varied Amplitude conditions is that, in the Varied Amplitude condition, *random-A* was rated as more emotional than *mechanical* and no different from *expressive-A*. In the Varied Timing condition, this was not the case; *random-T* was rated as no different from *mechanical* but was rated as less emotional than both *expressive-T* and 75%-T. A possible explanation for this is that the reduced sensitivity participants show when only amplitude is varied (as evidenced by the differences in effect size and the reduced slope for the Varied Amplitude condition) also causes them to be less sensitive to the "correctness" of the placement of the amplitude variations. Alternatively, if there is less variation overall in amplitude, perhaps these relatively small variations do not provide enough information for participants to judge where they "correctly" belong, that is, the stimuli may have been at or near the perceptual threshold for amplitude variation in this real-world musical context. Thus, participants may be able to perceive that the timing variation is incorrectly placed in the *random* performances, but be unable to perceive that the amplitude is not varying in a conventional way.

## General Discussion

The results of these three experiments provide converging evidence that average listeners are able to detect subtle variations in the expressive performance of piano pieces. Musicians demonstrate a greater sensitivity to these performance variations than nonmusicians (Experiments 1 and 3).

Experiments 1 and 2 showed that listeners are attuned to such subtle cues as changes in timing and amplitude. The results of Experiment 3 further expand upon these findings and confirm that both musicians and nonmusicians can detect the difference between levels of expressivity when the two dimensions of timing and amplitude are decoupled and manipulated separately.

The "random" condition in Experiments 1 and 3 served as an important control, to probe whether listeners were basing their emotionality ratings merely on the amount of information content or variability contained in the signal. Our findings confirmed that they were not: both musicians and nonmusicians discerned that the randomly permuted timing sounded both different from and less expressive than the other versions of the piece. This suggests that the general listener has internalized knowledge of expressive conventions in Western music, and recognizes when these conventions are not being observed. This extends the findings by De Poli (2003) that listeners can reliably interpret a performer's expressive intentions (c.f. Vines et al., 2005).

We also found that listeners can detect differences between some but not all of the adjacent levels of expressivity. In Experiments 1 and 3, the data suggest that listeners are more sensitive to differences in the middle of the curve than at the ends; Experiment 1 listeners heard more of a difference between 50 and 25% expressive than they did between other adjacent pairs. In Experiment 3, in the Varied Timing condition, participants differentiated between the top half (100, 87.5, and 75%) and the bottom half (50%, mechanical, and random) of the expressivity levels but did not differentiate within these halves. In the Varied Amplitude condition participants showed reduced sensitivity, only differentiating mechanical from the other expressivity levels. The listeners in Experiment 2 did not differentiate among excerpts above 100% expressive, showing that this top end of our experimental stimuli was not arbitrary. The ratings overall do show a general linear trend for expressivity level, and this is consistent with the linear interpolation we used in our manipulation of the expressive factors in the piano performance.

## Effects of Musical Training

In both Experiments 1 and 3, we found that musical training affected the way participants rated the emotionality of the stimuli. In Experiment 1, the direction of the ratings was not affected by musical experience, but the degree of difference between levels of expressivity was greater for people with more musical experience, demonstrating that musical experience was associated with an increase in listeners' sensitivity to these manipulations.

In Experiment 3, there were significant interactions between musical experience and tonality for both the Varied Timing and Varied Amplitude conditions. This is because the nonmusicians tended to rate KK IVa (the minor nocturne) as being more emotional than Op. 15 (the major nocturne) while musicians' ratings were more consistent between the two. This indicates that the nonmusicians, relative to musicians, were basing a greater proportion of their judgments on the tonality of the piece rather than on the expressivity levels, while musicians were basing their judgments primarily on the expressivity levels, the performances' variations in timing and amplitude.

## Parallels With Previous Research

One set of findings that motivated our experimental design concerns the systematic way in which timing and amplitude are varied in actual performance. Although no two performances are exactly alike, instrumentalists tend to stay within certain boundaries or constraints that define an acceptable musical performance within a particular idiom or genre (Repp, 1998; Shaffer & Todd, 1994). For example, it is common for performers to play more quietly as their playing slows (Repp, 1996), and this coupling occurs most often at phrase boundaries (Palmer, 1996; Todd, 1985), which is also where major increases in emotionality tend to occur (Sloboda & Lehmann, 2001; Vines et al., 2005). Melodic contour and amplitude are also correlated: higher pitches within a melody tend to be played louder than lower pitches, and this correlation remains small but consistent across the entire piece (Palmer, 1996). The KTH rule system models many of these structure- and intended emotion-based performance strategies, including these phrase-dependent and pitch-dependent tempo and amplitude changes (Friberg, Bresin, & Sundberg, 2006).

The ways in which different pianists employ expressive variation can be captured within a relatively small number of general types or schemas. In a study by Repp (1992), the variability of professional pianists (performing Schumann's Träumerei) could be accounted for by four factors (as extracted through Principal Components Analysis; two of these factors were largely because of artists' personal style and were named the "Cortot" and "Horowitz" factors). The performances of students, on the other hand, could be accounted for by only one factor (Repp, 1992, 1995b; for further examples, see Juslin & Laukka, 2003, and Palmer, 1996). Repp (1995a) proposes that this difference stems from the students' not yet having developed an individual style, with their performances being more mainstream and conservative than those of the experts.

## Potential Confounds and Future Directions

This experiment used piano performances of compositions from the standard practice period of classical music. Further research on other instruments (e.g., those in which timbre and intonation can be varied) and other musical genres will be required before one can make any general claims. It would also be interesting to study pieces of music that are not as reliant on pedaling for performance expression. In our studies, we varied the pedal along with the timing and amplitude variation to prevent it from overshadowing the other performance cues. Thus, it was always covaried with timing and amplitude.

In addition to examining other types of music, it could be fruitful to study additional performers or additional performances from a single performer. Though performers follow many of the same conventions in their performances (Juslin, 2000, 2001; Repp, 1992), professional musicians do demonstrate considerable individuality of style (Repp, 1995b) and performances may be affected by a performer's current emotional state (De Poli, 2003), so these parametric manipulations may affect their performances in different ways. Another future experiment would be to combine all different levels of the amplitude and timing variation. For example, how would a piece with 25% of timing variation and 75% of amplitude variation compare with a piece with 50% variation of each? An experiment leading in another direction would be to create a longer performance that has changes in expressivity within a single piece, similar to a previous study of moment-to-moment perceptions of expressivity within a piece (Sloboda & Lehmann, 2001), but while incorporating explicit variation in expressivity. Participants could then rate the emotion in the piece continuously, perhaps using the continuous tracking technique (Krumhansl, 1987; Schubert, 2004; Vines, Krumhansl, Wanderley, & Levitin, 2006), helping to control for the inherent differences in expressivity across different portions of a piece.

Researchers have suggested that some of the timing variation inherent in a human performance of music is due not to expressiveness but to perceptual coding of low-level features of the music—if notes are grouped together based on Gestalt (Wertheimer, 1923/1938), auditory scene analysis (Bregman, 1990) or other grouping principles, performers may hear some time intervals within and around these groups as being shorter than others and play them longer to compensate (Penel & Drake, 1998, 2004). Listeners may then have trouble hearing these intervals as being longer than the others because it sounds "right" to them. Our results could be interpreted as supporting this hypothesis; the mechanical version sounded very unemotional to listeners, and the random version, in which the time intervals were lengthened and shortened but in the incorrect places, sounded even less emotional. Even if these are perceptual and not musically expressive timing modulations, the lack of them makes a performance sound less human and less familiar, perhaps leading it to be less emotionally communicative.

Another possible objection is that our question of "how emotional" the music was is vague, and we cannot be sure of how people are interpreting it and on what they are basing their answers. However, this is the case with any research based on self-report, participants' reports of themselves may be biased in unpredictable ways. We did not use physiological measures such as galvanic skin response because we were not concerned with how the participant felt; we were concerned with what they perceived in the music, and these are two different phenomena (Gabrielsson, 2002). We know that the participants are not simply answering the question "which has *more*?" (whether it refers to variability, interest, or anything other than emotion) because they typically rate the random version as the lowest, and it has as much variability, novelty and information content as the expressive version, which was rated highest (excepting the Varied Amplitude condition, which may have been a statistical anomaly).

The imprecision inherent in any mechanical device was also a consideration in these experiments. Goebl & Bresin (2003) have shown that timing of the Yamaha Disklavier Pro Grand Piano can vary by as much as 30 ms in reproduction, yet this is still considered to be highly accurate, so much so that it is used for judging piano performance in international competitions (Tommasini, 2002). This is one reason we used recordings of performances, to minimize the effects of mechanical variability during the playback of successive trials.

## Conclusion

In this series of experiments, we attempted to quantify the relation between physical (acoustic) parameters of musical performance and the psychological representation of musical emotion. To our knowledge, this is the first study to present a continuum of carefully controlled variations of music in the physical domain from which were obtained the corresponding responses in the psychological domain. We found that objective changes in timing and amplitude variation do indeed affect subjective judgments of expressivity, and the psychophysical function relating these judgments to the varied parameters is generally monotonic, but contains some nonlinearities. In particular, threshold effects were observed at the ends of the rating curve, and greater discriminability occurred in the middle of the range. Musicians were found to be more sensitive than nonmusicians to covaried timing and amplitude variation. For all listeners, timing variation alone carries more expressive information than amplitude variation alone, although amplitude variation does significantly contribute to the emotional expressiveness of a performance. Finally, we believe that the technique we have introduced here, involving the parametric manipulation of real performances by real musicians, has shown itself to be useful for the study of music and emotion perception.

## References

Balkwill, L. L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception, 17,* 43–64.

Bernstein, L. (1976/1981). *The unanswered questions: Six talks at Harvard (The Charles Eliot Norton Lectures).* Cambridge, MA: Harvard University Press.

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge, MA: M. I. T. Press.

Clarke, E. F. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 473–500). San Diego: Academic Press.

Clynes, M. (1983). Expressive microstructure in music, linked to living qualities. In J. Sundberg (Ed.), *Studies of music performance* (pp. 76–181). Stockholm: Royal Swedish Academy of Music.

De Poli, G. (2003). Analysis and modeling of expressive intentions in music performance. *Annals of the New York Academy of Sciences, 999,* 118–123. doi:10.1196/annals.1284.012

Drake, C., Penel, A., & Bigand, E. (2000). Tapping in time with mechanically and expressively performed music. *Music Perception, 18,* 1–23.

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology, 2*(2–3), 145–161. doi:10.2478/v10053-008-0052-x

Gabrielsson, A. (1999). The performance of music. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 501–602). San Diego: Academic Press.

Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae, Special Issue,* 123–147.

Gaser, C., & Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *The Journal of Neuroscience, 23*(27), 9240–9245 doi:0270-6474/03/239240-06

Goebl, W., & Bresin, R. (2003). Measurement and reproduction accuracy of computer-controlled grand pianos. *Journal of the Acoustical Society of America, 114*(4, Pt. 1), 2273–2283. doi:10.1121/1.1605387

Gomez, P., & Danuser, B. (2007). Relationships between musical structure and psychophysiological measures of emotion. *Emotion, 7*(2), 377–387. doi:10.1037/1528-3542.7.2.377

Guastavino, C. (2007). Categorization of environmental sounds. *Canadian Journal of Experimental Psychology, 60*(1), 54–63. doi:10.1037/cjep2007006

Halpern, A., Martin, J. S., & Reed, T. D. (2008). An ERP study of major-minor classification in melodies. *Music Perception, 25,* 181–191. doi:10.1525/mp.2008.25.3.181

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice, 21,* 531–540. doi:10.1016/j.jvoice.2006.03.002

Handel, S. (1993). *Listening.* Cambridge, MA: MIT Press.

Juslin, P. N. (1997). Emotional communication in music performance: A functionalist perspective and some data. *Music Perception, 14,* 383–418.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*(6), 1797–1812. doi:10.1037/0096-1523.26.6.1797

Juslin, P. N. (2001). Communicating emotion in music performance: A review and theoretical framework. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 309–337). New York: Oxford University Press.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*(5), 770–814. doi:10.1037/0033-2909.129.5.770

Juslin, P. N., & Madison, G. (1999). The role of timing patterns in recognition of emotional expression from musical performance. *Music Perception, 17*(2), 197–221.

Kendall, R. A., & Carterette, E. C. (1990). The communication of musical expression. *Music Perception, 8*(2), 129–164.

Koelsch, S., Schröger, E., & Tervaniemi, M. (1999). Superior pre-attentive auditory processing in musicians. *NeuroReport, 10,* 1309–1313. doi:10.1097/00001756-199904260-00029

Krumhansl, C. L. (1987). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology, 51*(4), 336–352.

Krumhansl, C. L., & Kessler, E. J. (1992). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review, 89*(4), 334–368. doi:10.1037/0033-295X.89.4.334

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*(4), 612–625. doi:10.1111/j.1468-2958.2002.tb00828.x

Levitin, D. J., & Menon, V. (2003). Musical structure is processed in "language" areas of the brain: A possible role for Brodmann Area 47 in temporal coherence. *NeuroImage, 20*(4), 2142–2152. doi:10.1016/j.neuroimage.2003.08.016

Levitin, D. J., & Tirovolas, A. K. (2009). Current advances in the cognitive neuroscience of music. *The Year in Cognitive Neuroscience 2009: Annals of the New York Academy of Sciences, 1156,* 211–231. doi:10.1111/j.1749-6632.2009.04417.x

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

Moore, B. C. J. (1997). *An introduction to the psychology of hearing* (4th ed.) San Diego: Academic Press.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America, 93,* 1097–1108. doi:0001-4966/93/021097-12

Musacchia, G., Strait, D., & Kraus, N. (2008). Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. *Hearing Research, 1–2,* 34–42. doi:10.1016/j.heares.2008.04.013

Pallesen, K. J., Brattico, E., Bailey, C., Korvenoja, A., Koivisto, J., Gjedde, A., & Carlson, S. (2005). Emotion processing of major, minor, and dissonant chords: A functional magnetic resonance imaging study. *Annals of the New York Academy of Sciences, 1060,* 450–453. doi:10.1196/annals.1360.047

Palmer, C. (1996). Anatomy of a performance: Sources of musical expression. *Music Perception, 13*(3), 433–453.

Palmer, C. (1997). Music performance. *Annual Review of Psychology, 48,* 115–138. doi:10.1146/annurev.psych.48.1.115

Palmer, C., & Hutchins, S. (2006). What is musical prosody? In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 46). Boston: Academic Press.

Palmer, C., Jungers, M. K., & Jusczyk, P. W. (2001). Episodic memory for musical prosody. *Journal of Memory and Language, 45*(4), 526–545. doi:10.1006/jmla.2000.2780

Pantev, C., Ooostenveld, R., Engelien, A., Ross, B., Roberts, L. E., & Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature, 392,* 811–814. doi:10.1038/33918

Parncutt, R., & Troup, M. (2002). Piano. In R. Parncutt & G. E. McPherson (Eds.), *The science and psychology of music performance: Creative strategies for teaching and learning* (pp. 285–302). New York: Oxford University Press.

Penel, A., & Drake, C. (1998). Sources of timing variations in music performance: A psychological segmentation model. *Psychological Research, 61*(1), 12–32. doi:10.1007/PL00008161

Penel, A., & Drake, C. (2004). Timing variations in music performance: Musical communication, perceptual compensation, and/or motor control? *Perception & Psychophysics, 66*(4), 545–562.

Pierce, J. R. (1961/1980). *An introduction to information theory: Symbols, signals and noise.* New York: Dover.

Pitt, M. A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of Experimental Psychology: Human Perception and Performance, 20*(5), 976–986. doi:0096-1523/94

Pras, A., Zimmerman, R., Levitin, D. J., & Guastavino, C. (2009). Subjective evaluation of mp3 compression for different musical genres. *Proceedings of the 127th Convention of the Audio Engineering Society (AES).* New York, October 9-12, 2009.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America, 88*(2), 622–641. doi:10.1121/1.399766

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei." *Journal of the Acoustical Society of America, 92*(5), 2546–2568. doi:10.1121/1.404425

Repp, B. H. (1995a). Expressive timing in Schumann's "Träumerei": An analysis of performances by graduate student pianists. *Journal of the Acoustical Society of America, 98*(5), 2413–2427. doi:10.1121/1.413276

Repp, B. H. (1995b). Quantitative effects of global tempo on expressive timing in music performance: Some perceptual evidence. *Music Perception, 13*(1), 39–57.

Repp, B. H. (1996). The dynamics of expressive piano performance: Schumann's "Träumerei" revisited. *Journal of the Acoustical Society of America, 100*(1), 641–650. doi:10.1121/1.415889

Repp, B. H. (1998). A microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America, 104*(2, Pt. 1), 1085–1100. doi:10.1121/1.423325

Salimpoor, V., Guastavino, C., & Levitin, D. J. (2007, October). *Subjective evaluation of popular audio compression formats.* 123rd Meeting of the Audio Engineering Society, New York.

Schellenberg, E. G., Krysciak, A. M., & Campbell, R. J. (2000). Perceiving emotion in melody: Interactive effects of pitch and rhythm. *Music Perception, 18,* 155–171.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception, 21*(4), 561–585. doi:10.1525/mp.2004.21.4.561

Shaffer, L. H., & Todd, N. P. M. (1994). The interpretive component of musical performance. In R. Aiello (Ed.), *Musical perceptions* (pp. 258–270). New York: Oxford University Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana, IL: University of Illinois Press.

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*(1), 87–120. doi:10.1525/mp.2001.19.1.87

Smith, B. (1995). *Psiexp: An environment for psychoacoustic experimentation using the IRCAM musical workstation.* Paper presented at the Society for Music Perception and Cognition Conference, University of California, Berkeley.

Sridharan, D., Levitin, D. J., Chafe, C. H., Berger, J., & Menon, V. (2007). Neural dynamics of event segmentation in music: Converging evidence for dissociable ventral and dorsal networks. *Neuron, 55,* 1–12. doi:10.1016/j.neuron.2007.07.003

Strait, D. L., Kraus, N., Skoe, E., & Ashley, R. (2009). Musical experience and neural efficiency – effects of training on subcortical processing of vocal expressions of emotion. *European Journal of Neuroscience, 29*(3), 661–668. doi:10.1111/j.1460-9568.2009.06617.x

Sundberg, J., Friberg, A., & Fryden, L. (1988). Musicians' and non-musicians' sensitivity to differences in music performance. *Kungliga Tekniska högskolan Quarterly Progress and Status Report, 29*(4), 77–81.

Sundberg, J., Friberg, A., & Fryden, L. (1991). Threshold and preference quantities of rules for music performance. *Music Perception, 9,* 71–92.

Taylor, C. A. (1965). *The physics of musical sounds.* Aylesbury, England: The English Universities Press Ltd.

Taylor, C. A. (1992). *Exploring music: The science and technology of tones and tunes.* Philadelphia, PA: Institute of Physics Publishing.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion, 4*(1), 46–64. doi: 10.1037/1528-3542.4.1.46

Timmers, R. (2002). *Freedom and constraints in timing and ornamentation.* Maastricht, The Netherlands: Shaker Publishing.

Todd, N. P. M. (1985). A model of expressive timing in tonal music. *Music Perception, 3,* 33–58.

Todd, N. P. M. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America, 91*(6), 3540–3550. doi:10.1121/1.402843

Tommasini, A. (2002). Critics notebook: An international e-competition relies on the high-tech e-piano. *New York Times,* June 13, 2002, p. E5.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Dalca, I. M., & Levitin, D. J. (2005). Dimensions of emotion in expressive musical performance. *Annals of the New York Academy of Sciences, 1060,* 462–466. doi:10.1196/annals.1360.052

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition, 101,* 80–113. doi:10.1016/j.cognition.2005.09.003

Wertheimer, M. (1923/1938). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of Gestalt psychology.* London: Routledge & Kegan Paul.

Received October 22, 2009
Revision received August 2, 2010
Accepted August 10, 2010 ■