What does seeing the performer add? It depends on musical style, amount of stage behavior, and audience expertise

Jennifer Huang and Carol Lynne Krumhansl
Cornell University, USA

Abstract
The purpose of this study was to examine the effects of musical style, amount of stage behavior, audience expertise, and modality of presentation on structural, emotional, and summary ratings of piano performances. Twenty-four musically trained and 24 untrained participants rated two-minute excerpts of pieces by Bach, Chopin, and Copland, each performed by the same pianist, who was asked to vary his stage behavior from minimal to natural to exaggerated. Participants rated the performances under either audio-only or audiovisual conditions. The composer’s style had a consistently strong effect on the performance evaluations, highlighting the importance of careful repertoire selection. Moreover, the preferred degree of stage behavior depended on the style. The interaction between expertise, modality, and stage behavior revealed that non-musicians perceived differences across the three degrees of stage behavior only audiovisually and not in the audio-only condition. In contrast, musicians perceived these differences under both audiovisual and audio-only conditions, with the lowest ratings for minimal stage behavior. This suggests that varying the degree of stage behavior altered the quality of the performance. In addition, participants were asked to select two emotions that best characterized each performance. They preferentially chose more subtle emotions from Hevner’s (1936) Adjective Circle over the five general emotions of happiness, sadness, anger, fear, and tenderness traditionally used in music cognition studies, suggesting that these five emotions are less apt to describe the emotions conveyed through musical performance.

Keywords
audience expertise, emotion, modality, music performance, piano, stage behavior

Introduction
Performance expression in music refers to anything the score leaves wholly or partially unspecified, including such actions as slowing the tempo, adding durational accents, and varying the dynamics (sound level). It has been found to serve two main functions: to convey structure and to convey emotion. Principle structural aspects of Western classical music communicated through performance expression include phrasing, rhythm, meter, melody, and modulations of
key (Clarke, 1982, 1985; Gabrielsson, 1999; Palmer, 1989). Studies examining these aspects of musical structure tend to focus on variations in timing and dynamics, but different instruments offer the possibility of varying other qualities such as attack, decay, and timbre.

Regarding emotion, an influential paradigm of Gabrielsson, Juslin and colleagues focused on a few general emotions such as happiness and sadness (e.g., Gabrielsson & Juslin, 1996; Juslin, 1997). In one study, Juslin (1997) instructed professional guitarists to play a melody and to communicate the general emotions of anger, happiness, sadness, or fear, then to play the same melody without expression. Listeners successfully distinguished all of the intended emotions from each other, illustrating how strongly the performer’s emotional intent influenced the perceived emotion. This study found the expressive cues of tempo, dynamics, timing, timbre, and articulation to be important in decoding the emotional expression of music. Effects of this kind have been replicated in many experiments using this paradigm, summarized in Juslin (2001), which also provides a theoretical framework.

This previous research has considered music primarily as an aural experience; consequently, it has focused on acoustic cues. However, visual information can provide powerful cues for a wide range of perceptual, cognitive, and emotional responses. Quite a few studies have considered the effect of seeing the performance on how it is evaluated. This has obvious practical value for performing musicians. Lang Lang, whose performances are sometimes described as “flamboyant”, is a current example of how controversial the issue of performance expression can be (Elias, 2010; Pomfret, 2009; Smith, 2010; Sterns, 2009). Moreover, from a psychological point of view, music provides a complex and interesting means to explore how the modalities of vision and audition combine.

The literature has taken various approaches to understanding the effects of seeing the performance. In some cases, comparisons are made between both seeing and hearing the performance (audiovisual), and only hearing it (audio-only). In other cases, a condition in which the participant only sees the performance (visual-only) is also included. In conjunction with varying the modality of presentation, other variables have been manipulated. One is whether the performers are given explicit instructions as to the degree of expression, from minimal to exaggerated. The studies also vary in terms of the expertise of the performer, from young music students to professional musicians. Some studies have investigated possible differences between expert musicians’ and non-musicians’ judgments. The studies differ quite widely in terms of the specific music that is performed. And, finally, the kinds of evaluations that are made cover a wide range, from very specific technical aspects to overall quality of the performance.

**Performance and evaluation**

A seminal paper by Davidson (1993) investigated the effect seeing the performance had on how expressive it was judged to be. Participants were assigned to one of three conditions: audio-only, visual-only, or audiovisual. Performers were asked to play the same excerpt in three ways: “projected” as they would normally perform in public, “deadpan” with minimal expressive interpretation, and “exaggerated” with overstated expressive interpretation. For the visual-only and audiovisual presentation modes, Davidson (1993) adapted Johansson’s (1973) reflective ribbons technique to create a stimulus based solely on the musicians’ movements, so that the participants saw the point-light images from their joint movements. Only in the audiovisual and visual-only performances were the participants able to distinguish between the three performance manners. This was true for performances of both a violin quartet and a solo pianist, demonstrating the pertinent role of visual cues in judging the perceived expressivity of a performance.
Gillespie (1997) conducted a study comparing experts’ ratings of audio-only and audiovisual presentations of violin and viola students’ vibratos on five factors: width, speed, evenness, pitch stability, and overall sound. As expected, in both conditions, experienced players rated higher than inexperienced players for all five factors. However, inexperienced players were rated significantly higher for pitch stability, evenness, and overall sound under the audiovisual condition than they were under the audio-only condition. Experienced players were rated higher only for pitch stability under the audiovisual condition than they were under the audio-only condition. These findings suggest that, when given visual information, adjudicators based their ratings more on seeing the correct motion of producing a vibrato than on hearing the proper sound from the vibrato. Inexperienced players in particular seemed to benefit more from audiovisual presentation than from audio-only presentation.

The results of Gillespie’s (1997) study suggest that visual information is capable of superseding auditory information, but not all previous research has found an effect of modality of presentation. In a cross-cultural study exploring Argentine and American musicians’ evaluations of a conductor and choir, an experienced conductor was asked to vary between good and poor conducting (Madsen, 2009). The results showed consistently high agreement between Argentine and American participants’ evaluations, but no significant differences in overall ratings across the three presentation modes (audio-only, visual-only, and audiovisual). The researcher cautioned against generalizing too much from these results, due to the small number of participants and the fixed presentation order of the conditions.

More recent studies suggest that mere arousal effects do not adequately account for the influence of visual information on performance evaluation. Broughton and Stevens (2009) conducted a study asking marimba players to play in either a projected performance condition or a deadpan performance condition. Participants were randomly assigned to either the audio-only or the audiovisual condition and asked to provide ratings of their level of interest in the performance. There was a significant difference between ratings given under audio-only and audiovisual conditions for the projected performances, but not for the deadpan performances. Specifically, additional visual information only elevated participants’ interest when it contained expressive movements.

Similar results were found in a study by Wapnick, Mazza, and Darrow (1998). Graduate students and faculty members in music were asked to evaluate the performances of undergraduate violinists on intonation, dynamic range, phrasing, sound quality, and overall performance. The researchers found that violinists who rated high on stage behavior and on dress received higher performance ratings under the audiovisual condition than they did under the audio-only condition. In contrast, the performance evaluations of violinists who rated low on these two attributes did not differ significantly between the audiovisual and audio-only conditions.

Wapnick et al. (2000) replicated these results in another study using children’s musical performances. Musically trained participants were asked to rate children’s solo piano performances on rhythmic accuracy, dynamic range, phrasing, and overall performance, as well as how talented the child seemed to be. Children who rated high on the visual attributes were given higher performance ratings than children who rated low on these attributes, under both the audiovisual and audio-only conditions. The difference between the high-attribute and low-attribute groups was significantly greater in the audiovisual condition than in the audio-only condition, largely carried by the scores for high-attribute pianists being higher in the audiovisual condition than in the audio-only condition. Again, the scores for low-attribute pianists did not differ significantly between the audiovisual and audio-only conditions.
Whereas the series of studies by Wapnick et al. (2000) used performances by novice pianists and non-professional violinists, a study by Ryan, Wapnick, Lacaille, and Darrow (2006) utilized selected excerpts of pianists’ performances from the prestigious International Van Cliburn Competition, controlling for potential effects of tempo and musical style. The participants, who were all musically trained, were asked to evaluate the performances on tone quality, note accuracy, rhythmic accuracy, expressiveness, adherence to style, and overall impression. A general increase in scores from audio to audiovisual conditions was found, consistent with most previous findings (Davidson, 1993; Gillespie, 1997; Wapnick et al., 1998). However, pianists exhibiting low stage behavior actually received higher performance ratings than pianists exhibiting high stage behavior, contradicting the results of previous studies by Wapnick et al. (1998, 2000, 2004), perhaps owing to differences in the degrees of stage behavior exhibited by each pianist, differences in the performers’ level of expertise, or other differences between the experiments.

In a separate study using excerpts from the 2001 Van Cliburn Competition, researchers found a significant effect of participant expertise on performance evaluation (Wapnick, Ryan, Lacaille, & Darrow, 2004). Non-pianists were favorably influenced by the added visual component, and this effect was so strong that it led to a significant main effect of modality, with performances in the audiovisual conditions rated higher than those in the audio-only conditions. In contrast, pianists’ performance ratings did not differ significantly between the two conditions. These results suggest that the expertise of the listener affects his or her evaluation of a musical performance. However, other researchers have found that musically trained participants generally gave higher ratings than non-musicians did (Broughton & Stevens, 2009).

**Performance and emotion**

As noted above, the research applying Gabrielsson and Juslin’s paradigm (Gabrielsson & Juslin, 1996; Juslin, 1997) has focused on general emotions, such as anger, happiness, sadness, or fear. In a study of how these emotion judgments are affected by seeing the performance, Dahl & Friberg (2007) had participants rate the emotional content of musical excerpts on a seven-point scale for the emotions of fear, anger, happiness, and sadness. In visual-only presentations of marimba performances and wind performances, the emotions of happiness, sadness, and anger were all communicated, whereas that of fear was not. This suggests that movement alone can convey specific information about the performer’s emotional intent, albeit not equally well for all emotions.

Thompson, Russo, and Quinto (2008) examined the role of singers’ facial expressions in the emotional connotation of sung intervals. Participants, presented audiovisually with the stimulus materials, were asked to rate intervals of either major or minor thirds on a five-point scale from “very sad” to “very happy”. Major thirds presented with the singer’s matching facial expression were rated “very happy”, and minor thirds presented with the congruent facial expression were rated “very sad”, whereas incongruent audiovisual information resulted in intermediate emotion ratings, between “very sad” and “very happy”. Thus, not only the melodic interval but also the facial expression influenced the judgment of emotion. This effect was not diminished by asking participants to base their judgments solely on the auditory information. Nor was it affected by asking participants to perform a secondary task during the experiment, which was done in order to manipulate attentional load. These results suggest that the visual components of the performance were automatically registered and integrated with the auditory components.

Vines, Krumhansl, Wanderley, and Levitin (2006) had participants judge tension continuously in real time during performances of a solo clarinet piece. Tension was used in this study
as a measure of emotional response. Audiovisual and audio-only judgments differed considerably. In performances where the tension conveyed visually contrasted with the tension conveyed aurally, participants combined the two sources of information rather than ignoring one or the other, which supports Thompson et al.’s (2008) conclusions regarding integrating audio and visual information. In contrast, the phrasing judgments, which were also made continuously during the performances, were nearly identical in the audio-only and visual-only conditions, showing that performance gestures convey much of the same structural information to an audience as audio does. The gestures were particularly important at the beginning and ending of phrases, where they outlined the musical structure.

Vines et al. (2010) presented emotion judgments that were made on the same performances in audiovisual, audio-only, and visual-only presentation modes. The participants judged the performances on nineteen emotion adjectives that have been used in previous research. The evaluative scales received the highest ratings: expressivity, movement, familiarity, intensity, quality, and interest. In contrast, the general emotion terms received lower ratings; these included happiness, anxiety, anger, sadness, fear, and disgust. In the visual-only condition, the ratings dropped sharply for the adjectives that were generally rated the highest. In other words, just seeing the performance resulted in decreases in the evaluative emotion adjective. The audiovisual condition was intermediate between the visual-only and the audio-only conditions.

The present study

To summarize, the current literature finds a wide variety of effects involving the modality of presentation, the type or degree of stage behavior, and the expertise of the judges. The findings are somewhat mixed, probably owing to variations in the particulars of the experimental designs, such as stimulus materials, the expertise of the performances, and the qualifications of the judges. Moreover, it remains unclear whether and how these variables might interact with one another. For example, expertise might have a different effect depending on the modality of presentation. Also, few previous studies have explicitly varied the composer or style of music in the same experiment. Different pieces of music may have different inherent stylistic qualities. This would be expected to have a variety of effects because, for example, it may be more difficult to convey certain emotions through the performance (Juslin, 1997).

The studies also vary considerably in how the performances are evaluated. Some ask participants to judge overall quality, degree of expressivity, level of interest, dynamics, phrasing, sound quality, and even specific technical aspects such as vibrato. Other studies focus on the emotions that are communicated through the performances. These include the general emotions such as happiness, sadness, fear, and anger. However, other adjectives have been used – conveying for example movement, intensity, and pleasantness – which relate to the idea that emotions can be represented in a two-dimensional space with axes corresponding to valence (positive to negative), and arousal (from active to passive). It might be noted, however, that an analysis of the data in Vines et al. (2010) found four independent categories: Active/Positive, Passive/Positive, Active/Negative, and Passive/Negative, such that positive is not the opposite of negative, and active is not the opposite of passive.

With this literature in mind, we designed an experiment that included all combinations of the modality of presentation (audio-only and audiovisual), degree of expressivity (minimal, natural, and exaggerated stage behavior), the expertise of the evaluators (non-musicians and musicians), and the style of classical music (baroque, romantic, and modern). We also sought to consider a wide range of characteristics that might be affected by modality, degree of stage
behavior, expertise, and musical style. We obtained ratings on three scales concerning structural aspects of the music (form, melody/harmony, and phrasing), three scales associated with emotions (dynamics, unexpected events, and amount of emotion expressed), and three summary measures (interest, appropriateness of interpretation, and overall rating).

In creating the stimulus materials for the present study, we asked a professional pianist to play three different pieces from the following periods of classical music: baroque, romantic, and modern. Baroque music is typified by composers such as Johann Sebastian Bach, who wrote music with complex tonal counterpoint and basso continuo, or a continuous bass line. From the baroque era, the pianist chose Bach’s Prelude No. 8 in E-flat Minor (BMV 853), because his personal ideal for the piece was not to emote. Expressive melodies and harmonies were the focus of the romantic era, exemplified by the lush and dramatic music of Frédéric Chopin. The pianist chose Chopin’s Ballade No. 1 in G Minor (Op. 23) because he felt that the intensity of the music necessitated a great deal of flexibility. Finally, from the modern period of classical music, the pianist chose to play Variations 3–8 of Aaron Copland’s Piano Variations. Twentieth-century composers writing in the modern style, such as Copland, abandoned tonality, and the pianist described the Variations as very dissonant, with coiled energy that was repressed at some moments and released at others. The piano was chosen as the focal instrument, both for practical reasons and because pianists face the additional communicative hurdle of not being able to face the audience during performance.

In addition to the nine evaluative scales, the present study sought to determine whether the five emotions of happiness, sadness, anger, fear, and tenderness, often used in studies involving musical performance and emotion, were apt to describe the emotions conveyed through the music. In addition to these five emotions, we offered the choices of more subtle emotions from Hevner’s Adjective Circle (1936), shown in Figure 1. Hevner (1936) found that these clusters of emotions expressed the variety of emotional responses to music.

The performer selected two emotions from the circle that best conveyed the emotional content of each piece: “serene” and “plaintive” for the Bach piece, “yearning” and “melancholy” for the Chopin piece, and “emphatic” and “agitated” for the Copland piece. On the one hand, because these emotions are selected specifically for the excerpted pieces from adjectives that apply to music in particular, it might be expected that they would be chosen more often. On the other hand, the emotions may appear too specific, causing the more general emotions to be chosen more often. We also examined whether the choice of emotions was affected by the modality of presentation, the degree of stage behavior, or the participant’s level of musical expertise.

**Methods**

**Preparation of materials**

Stimulus tapes were created with the help of pianist Xak Bjerken, professor of piano performance at Cornell University. We controlled for potential effects of performer attractiveness, ability, and dress by using one expert pianist. The excerpts were recorded on the same day so that Professor Bjerken’s appearance differed as little as possible across all conditions. For each of the three pieces, Professor Bjerken was videotaped from the same camera angle showing his full profile from the right side, as an audience member sitting in the center of a concert hall would typically view the pianist. He was instructed to play the same two-minute excerpt of each of the three chosen pieces, varying his stage behavior in three ways: minimal stage
behavior, in which he was asked to move as little as possible; natural stage behavior, as he would normally perform the piece in concert; and exaggerated stage behavior, in which he was asked to exaggerate his movements as much as possible.

Three pieces played in three different ways resulted in a total of nine excerpts. The chosen excerpts from each take were converted via HandBrake to QuickTime video format, and the stimulus materials for the experiment were created using Microsoft PowerPoint. For the practice session at the beginning of the experiment, a Youtube video of a different pianist playing the second movement of Beethoven’s *Piano Sonata in E-flat Major* (Op. 31, No. 3) was downloaded and trimmed to 45 seconds using MPEG Streamclip.

**Participants**

A total of 61 participants completed the study. They consisted of two groups, distinguished on the basis of their formal music training: non-musicians from Cornell University and musicians from Eastman School of Music. Thirteen subjects were not included in the data analysis because they did not meet the criteria for their participant group. Specifically, they volunteered as non-musicians but in fact had more music training than the 4-year maximum.
Non-musicians

Twenty-four undergraduate students from Cornell University qualified as non-musicians for the purposes of this study. None of these participants was currently playing an instrument. The mean number of years of formal music training was 1.02 years, with a range of 0–4 years. The mean number of concerts attended per month was 0.04 concerts.

Musicians

Twenty-four undergraduate students from Eastman School of Music qualified as musicians for the purposes of this study. All of these participants were currently playing one or more instruments. The mean number of years of formal music training on the primary instrument was 9.17 years, with a range of 3–13 years. The mean number of concerts attended per month was 6.59 concerts.

Procedure

All participants were randomly assigned to either the audio-only or the audiovisual condition. The first excerpt presented was the 45-second extract of the practice session video clip. Participants were instructed to wait until the end of each excerpt to begin filling out the response form. Following the practice clip, participants were instructed to fill out the first page of the response form, and to ask if they had any questions.

For each excerpt, participants were asked to respond to nine questions on a Likert scale of 1–7, with 1 being “Not at all” and 7 being “Very much”. The following nine questions were asked of the participants:

1. How well does the performance bring out significant melodic/harmonic events?
2. How well does the performance bring out the dynamics (softness/loudness) of the piece?
3. How well does the performance bring out the phrases in the piece?
4. How well does the performance bring out unexpected events?
5. How well does the performance bring out the organic units of the piece?
6. How appropriate is the performer’s interpretation of this piece?
7. How appropriate is the performer’s expressed emotion?
8. How well does the performance maintain your interest?
9. What is your overall performance rating for this piece?

Questions 2, 4, and 7 pertained to the emotional aspects of the performance, and questions 1, 3, and 5 (intended to convey musical “form” to non-musicians) pertained to the structural aspects of the performance. Questions 6, 8, and 9 were summary measures of the performance.

In addition, participants were asked to indicate the two emotions that applied to the performance overall. They could choose from the five conventional emotions, happy, sad, anger, fear, and tenderness, as well as the six emotions chosen by the pianist from the Hevner Adjective Circle (1936) to describe each piece: serene and plaintive for Bach, yearning and melancholy for Chopin, and emphatic and agitated for Copland. The eleven emotions were intermixed and presented concurrently. At the end of the response form, participants were given space to write additional comments.
**Non-musicians**

Both the audio-only and the audiovisual conditions of the experiment were presented via PowerPoint on an iMac desktop computer. Participants completed the experiment either individually or in groups of 2 or 3, depending on how many participants signed up for a given timeslot. Each participant heard (audio-only) or both viewed and heard (audiovisual) all nine excerpts. To create the audio-only version of the experiment, a black screen was placed over the video clips in PowerPoint, so that the audio heard in the audio-only condition did not differ at all from that of the audiovisual condition. The excerpts were ordered so that no two composers or degrees of stage behavior would be presented consecutively.

**Musicians**

These participants completed the experiment in groups of between four and nine individuals. The video clips used were the same as those shown to the non-musicians. Each clip was projected onto a screen. The participants in the audio-only condition had their backs facing the screen for the duration of the experiment, so that the audio heard by these participants was identical to the audio heard by the participants in the audiovisual condition. These participants completed the same practice session and response form as the non-musicians.

**Results**

**Scale type**

Participants rated each performance on nine scales covering a variety of qualities. In order to gain some insight into the relationship between the scales, and to reduce the complexity of the analysis, a principal components analysis was conducted on the ratings for each combination of expertise, composer, stage behavior, and modality. The solution is shown in Figure 2.

Intuitively, phrases, melody/harmony, and form refer to structural aspects of music; interest, interpretation, and overall judgments are summary measures; and dynamics, unexpected events, and expressed emotion are associated with emotional responses. Thus, the nine scales can be clustered into three scale types, which we will call structural, summary, and emotional. In subsequent analyses, we will denote these as scale types.

Using the collapsed data, a 2 (expertise) x 2 (modality) x 3 (composer) x 3 (stage behavior) x 3 (scale type) MANOVA was performed, with the first two factors between-subjects and the last three within-subjects. Post-hoc Students’ t-tests were performed on the significant main effects and interactions. The main effect of scale type ($F(2, 3406) = 26.14, p < .0001$) was significant, with emotional scales receiving higher ratings than structural and summary scales. Scale type interacted with expertise ($F(2, 3406) = 5.31$) as shown in Figure 3a. Musicians gave approximately equal ratings on the three scale types, whereas non-musicians used emotional scales like musicians, but gave lower ratings on summary and structural scales.

Scale type also interacted with composer, ($F(4, 3406) = 27.99, p < .0001$) as shown in Figure 3b. As can be seen, the ratings for Chopin were uniformly high. The ratings on the emotional scales for Copland were higher than for Bach; conversely, the ratings on the structural scales were higher for Bach than for Copland. That these scales show this pattern as a function of composer is consistent with the inherently different musical qualities of the three pieces, and further supports the idea that the nine scales can be collapsed into the three scale types, emotional, summary, and structural.
Scale type also interacted with modality (expertise ($F(2, 3406) = 6.77, p = .0012$), as shown in Figure 3c. The structural scales received equally high ratings in the audio-only and audiovisual conditions, suggesting that structural information is carried equally across modality. However, for the other two scale types, emotional and summary, the ratings for the audiovisual condition were lower than the audio-only condition. This summarizes all significant effects involving scale type; no other interactions with scale type were significant.

**Composer effects**

We then performed a 2 (expertise) × 3 (composer) × 3 (stage behavior) × 2 (modality) MANOVA without scale type, because all its effects have already been presented. There was a very large main effect of composer ($F(2, 338) = 13.52, p < .0001$). Chopin was rated significantly higher than both Bach and Copland. However, composer interacted significantly with stage behavior ($F(2, 3725) = 9.98, p < .0001$), shown in Figure 4. As can be seen, the ratings for Chopin and Bach increased with increasing degrees of stage behavior, but the values for Copland were highest for the natural performances.
Figure 3a. **Expertise × scale type interaction.** Non-musicians’ ratings of summary and structural measures were significantly lower than all other ratings.

Figure 3b. **Composer × scale type interaction.** Chopin rated significantly higher than Bach and Copland for all three scale types. Copland rated significantly higher than Bach for emotional measures, and Bach rated significantly higher than Copland for structural measures.

Figure 3c. **Modality × scale type interaction.** Ratings for emotional and summary scales were significantly lower in the audiovisual condition than in the audio-only condition.
There was also a three-way interaction between composer, stage behavior, and modality ($F(4, 3725) = 3.98, p = .0032$). The lowest ratings were given to the minimal performances in the audiovisual condition for all three composers. In other words, seeing the performance with minimal Stage Behavior had a detrimental effect independent of composer. However, there were some differences. The Bach ratings increased with stage behavior, independently of modality. The Chopin ratings were higher in the audio-only condition than in the audiovisual condition for natural and exaggerated performances. In contrast, Copland was rated higher in the audiovisual condition. This summarizes all significant effects involving composer; no other interactions with composer were significant except its interaction with scale type discussed above.

**Stage behavior, modality, expertise**

We then performed a 2 (expertise) × 3 (stage behavior) × 2 (modality) MANOVA, because all effects of scale type and composer have been discussed. There was a significant main effect of stage behavior ($F(2, 338) = 8.91, p = .0002$), with no significant difference between the natural and exaggerated performances, but both of these performances rated significantly higher than the minimal performances. There were no main effects of modality or expertise.

There were significant stage behavior × expertise ($F(2, 3732) = 16.32, p < .0001$) and stage behavior × modality ($F(2, 3732) = 15.87, p < .0001$) interactions. The best way to visualize these interactions is by plotting the audio-only and audiovisual conditions separately. Figure 5a shows the results for the audio-only condition. Without seeing the performances, non-musicians were unable to distinguish between the minimal, natural, and exaggerated performances. In contrast, musicians rated both the natural and exaggerated performances higher than the minimal performances, even without seeing the performer.

Figure 5b shows the results for the audiovisual condition. Musicians again rated both the natural and exaggerated performances higher than the minimal performances. This difference is larger than it was in the audio-only condition, especially in the low ratings for the minimal performances. Now, non-musicians’ ratings increased from minimal to natural and exaggerated, although not as much as musicians; only the difference between minimal and exaggerated was significant for non-musicians. Thus, it appears that the non-musicians needed to see the performances in order to distinguish the three different degrees of stage behavior.

**Figure 4. Composer × stage behavior interaction.** Ratings for Chopin and Bach increased with increasing stage behavior. Ratings for Copland were highest for natural stage behavior.
There was no main effect of modality or interactions with the other variables, except with scale type, as discussed in the first section of the results. There was no main effect of expertise or interactions with other variables, except for its interaction with scale type, discussed in the first section of the results, and its interaction with stage behavior, just discussed. Thus, all significant results on the rating scales have been described.

**Emotions**

The percentage of responses for the eleven emotion choices was computed for all combinations of the independent variables. The result of a principal components analysis is shown in Figure 6.

The general emotions of anger and fear were closest to the Hevner emotions of agitated and emphatic, and the conventional emotions of sadness and tenderness were closest to the Hevner emotion of yearning (see Figure 1). Happiness is included in Hevner Group 6, but none of the chosen pieces conveyed happiness to the pianist. Nonetheless, happiness in the principal

---

**Figure 5a. Audio-only.** Non-musicians’ ratings did not differ significantly across the three degrees of stage behavior. Musicians’ ratings for natural and exaggerated performances were higher than for minimal performances.

**Figure 5b. Audiovisual.** Non-musicians’ ratings for exaggerated performances were higher than for minimal performances. Musicians’ ratings for natural and exaggerated performances were higher than for minimal performances.
A variable was constructed for emotion type: general (happy, sad, fear, anger, tenderness), Hevner adjectives specific to Bach (serene, plaintive) denoted Hevner (Bach) in Figure 7, Hevner adjectives specific to Chopin (yearning, melancholy) denoted Hevner (Chopin), and Hevner adjectives specific to Copland (emphatic, agitated) denoted Hevner (Copland). From there, a 2 (expertise) × 3 (composer) × 3 (stage behavior) × 2 (modality) × 4 (emotion type) multivariate analysis of variance was performed on the tally of times each emotion was selected.

All of the significant effects involved emotion type. There was a main effect of emotion type \((F(3, 252) = 29.78, p < .0001)\), shown in Figure 7a, with the Hevner emotions chosen by the pianist to describe Chopin and Copland chosen more frequently by the participants than those the pianist chose for Bach. Notably, the general adjectives were chosen the least often. There was also a significant interaction between emotion type and composer \((F(6, 252) = 44.70, p < .0001)\), shown in Figure 7b.

As expected, the adjectives specific to the Chopin and Copland pieces were also chosen most often for those pieces. The adjectives specific to the Bach piece were also chosen quite often for the Bach piece, yet slightly (but not significantly) less often than the adjectives specific to the Chopin piece. None of the other variables or their interactions was significant.
This experiment investigated influences on how a musical performance is evaluated. We designed the experiment to include three factors that have been used quite frequently in previous studies: the degree of stage behavior (minimal, natural, or exaggerated), the modality of presentation (audio-only or audiovisual), and the expertise of the participants judging the performances (non-musicians or musicians). All possible combinations of these three variables were used so that not only their main effects but also their interactions could be examined.

**Figure 7a. Main effect of emotion type.** Hevner emotions chosen by the pianist to describe the pieces by Chopin – Hevner (Chopin) – and Copland – Hevner (Copland) – were chosen significantly more often by the participants than the pianist’s Hevner emotions for the Bach piece – Hevner (Bach). General emotions were chosen significantly less frequently than the Hevner emotions.

**Figure 7b. Composer × emotion type interaction.** Hevner adjectives specific to Chopin and Copland were chosen most often for those pieces. Hevner adjectives specific to both Bach and Chopin were chosen most often for the Bach piece.

**Discussion**

This experiment investigated influences on how a musical performance is evaluated. We designed the experiment to include three factors that have been used quite frequently in previous studies: the degree of stage behavior (minimal, natural, or exaggerated), the modality of presentation (audio-only or audiovisual), and the expertise of the participants judging the performances (non-musicians or musicians). All possible combinations of these three variables were used so that not only their main effects but also their interactions could be examined.
These factors and their interactions are shown in the shaded area in the top left-hand corner of Table 1. We discuss these later and compare them with previous findings after we highlight the novel elements of the study.

A factor that has not received systematic treatment in most previous studies is the composer’s musical style. The pianist played different pieces, one each from three different periods of classical music: baroque (Bach), romantic (Chopin), and modern (Copland). This introduces many new questions concerning how musical performances are evaluated. One is whether there are simply overall differences depending on the style of the piece (or the piece itself). What is the optimal degree of stage behavior for the different composers? Does musical style change the influence of seeing and not just hearing the piece? How does musical style interact with the expertise of the observers, who may vary considerably in their exposure to these styles and to concert performances?

In addition, we introduced a wide range of evaluative measures. These concerned the quality of the performance in terms of structural or formal musical characteristics (phrasing, harmony/melody, form), emotional musical characteristics (unexpected events, dynamics, overall emotional expression), and summary measures (interest, appropriateness of interpretation, overall performance). Would ratings on these different kinds of scales depend on the degree of stage behavior, modality of presentation, expertise of the participant, and/or the particular piece being played?

Finally, we investigated the possibility that these factors might influence not only the overall amount of emotion conveyed, but also the particular emotion participants felt expressed in the music. To this end, we invited the performer to identify two emotions from the Hevner Adjective Circle (1936) that best fitted each of the three pieces. These more subtle musical emotions were intermixed with general emotions of the type that are often used in music cognition studies—in this case, happy, sad, fear, anger, and tenderness. Are the more subtle adjectives chosen over the more general ones? How do the different experimental conditions affect these choices?

As can be seen outside the shaded area of Table 1, there were strong effects of composer, scale type, and emotion type. These are effects that have not generally been considered or have not been systematically studied in previous research. Moreover, these factors interact with stage behavior, modality, and expertise, factors that have been more systematically studied in

Table 1. Summary of significant results. The table summarizes the statistical effects of stage behavior (minimal, natural, exaggerated), modality of presentation (audio-only, audiovisual), expertise (musicians, non-musicians), composer (Bach, Chopin, Copland), type of evaluative scale (emotional, summary, structural), and emotion type (general, designated for Bach, designated for Chopin, designated for Copland). The values show the level of significance (values less than .01 are considered significant). Values in bold on the diagonal show the main effects of each variable.

<table>
<thead>
<tr>
<th></th>
<th>Stage behavior</th>
<th>Modality</th>
<th>Expertise</th>
<th>Composer</th>
<th>Scale type</th>
<th>Emotion type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage behavior</td>
<td>.0002</td>
<td>&lt; .0001</td>
<td>&lt; .0001</td>
<td>&lt; .0001</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Modality</td>
<td>&lt; .0001</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.0012</td>
<td>—</td>
</tr>
<tr>
<td>Expertise</td>
<td>&lt; .0001</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.005</td>
<td>—</td>
</tr>
<tr>
<td>Composer</td>
<td>&lt; .0001</td>
<td>—</td>
<td>—</td>
<td>&lt; .0001</td>
<td>&lt; .0001</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Scale type</td>
<td>—</td>
<td>.0012</td>
<td>.005</td>
<td>&lt; .0001</td>
<td>&lt; .0001</td>
<td></td>
</tr>
<tr>
<td>Emotion type</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>&lt; .0001</td>
<td>—</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>
the performance literature. These interactions suggest that the effects of these factors needs to be qualified by the musical style of the composition, what aspect of the performance is being evaluated, and the particular emotional quality of the piece of music.

**Composer effects**

Let us first consider the effects involving composer. The main effect of composer is shown in bold on the diagonal of Table 1. Overall, the Chopin piece was preferred to the Bach and the Copland pieces, perhaps reflected in the prevalence of Chopin pieces typically chosen for piano competitions and audition tapes. Moreover, the preferred degree of stage behavior depended on the composer. That is to say, there is an interaction between stage behavior and composer. This is seen in Table 1, at the top of the column labeled Composer. The ratings for both Chopin and Bach increased as a function of stage behavior. Apparently, increasing the amount of stage behavior had a generally positive effect on the ratings. However, for the Copland piece, ratings increased from the minimal to the natural performances, then decreased from the natural to the exaggerated performances. Thus, the ratings for the Copland piece exhibited the kind of inverted U-shaped pattern that would be found if exaggerating one’s stage behavior detracted from the evaluations (one possible interpretation of findings from Ryan et al., 2006).

As the penultimate entry in the Composer column shows, how the evaluative scales were used depended strongly on the composer. All three scale types were rated highly for Chopin. The structural scales were rated highly for Bach but not for Copland, reflecting the Copland piece’s inherent lack of musical structure relative to the other two pieces. In a complementary fashion, the emotional scales were rated highly for Copland but not for Bach. Thus, the pieces were evaluated differently depending on the dimensions being evaluated. The ratings on the summary measures were lower for both Bach and Copland than for Chopin, which also supports our interpretation of the main effect of composer. An implication of this is that general results across scales cannot be predicted, and there will be strong effects dependent upon the particular composer’s musical style.

The last composer effect was with the type of emotion adjective chosen for the different pieces, shown as the last entry in the Composer column. We designed the experiment so that the emotion choices offered to participants could naturally be classified into four groups: general emotions (happy, sad, fear, anger, tenderness), emotions designated for the Bach piece (serene, plaintive), emotions designated for the Chopin piece (yearning, melancholy), and emotions designated for the Copland piece (emphatic, agitated). For all three pieces, the general emotions (happy, sad, fear, anger, tenderness) were chosen less often than the more subtle emotions designated by the performer for each of the three pieces. (This is seen as the main effect of emotion type at the bottom right-hand corner of Table 1.) However, the specific emotions chosen by the participants depended on the piece. The emotions designated by the performer for Bach (serene, plaintive) were chosen often for the Bach performance (although slightly less often than the emotions designated for the Chopin piece). The emotions designated for Chopin (yearning, melancholy) were most often chosen for the Chopin piece. Finally, the emotions designated for Copland (emphatic, agitated) were most often chosen for the Copland piece. What this means is that the participants perceived the intended emotion, and again, they perceived these subtler emotions more often than the general emotions.

It may seem obvious that the emotions designated specifically for these pieces would be chosen most often. Viewed in another light, however, it means that listeners agreed on the finer-grained emotions, and this was independent of their level of expertise. That is, the non-musicians
did not, as might have been expected, choose the more general adjectives more often than the musicians did. Thus, there was remarkable agreement between musicians and non-musicians in the choice of the more subtle musical emotions specific to each of the three pieces, and this agreement is consistent with previous work involving emotions conducted on groups with different levels of musical expertise (Juslin, 1997).

To summarize briefly, many strong effects depended on the composer’s musical style. The musical style affected the preferred degree of stage behavior, the scale types applied to the different pieces, and the type of emotion perceived. The particular effects found here are undoubtedly quite specific to the particular pieces of music chosen, but they suggest that understanding the effects of different composers and pieces as they interact with other factors would lead to a fuller account of how performances are evaluated. Moreover, these results highlight the importance of careful repertoire selection, given how much the performance ratings varied from piece to piece. The particular effects found here are also undoubtedly quite specific to the particular performer. Further studies with different performers would probably show additional interacting factors but would lack the control of using a single performer as was done in this study.

**Scale type effects**

Next, we turn to the effects of scale type, shown in the column labeled Scale type. Given the relationships discovered between the nine response scales as a function of stage behavior, modality, expertise, and composer, it was possible to group the scales into three scale types, as previously indicated: structural or formal musical characteristics (phrasing, harmony/melody, form), emotional musical characteristics (unexpected events, dynamics, overall emotion expression), and summary measures (interest, appropriateness of interpretation, overall performance). The three summary measures of interest level, appropriateness of interpretation, and overall rating were influenced by both structural and emotional aspects of the performance, consistent with previous studies defining these two elements as the primary functions of performance expression (Clarke, 1982, 1985; Gabrielsson & Juslin, 1996; Juslin, 1997, 2007; Palmer, 1989).

There were overall differences in the ratings given on the different types of evaluative scales. Ratings were higher on the emotional scales than on the structural or summary scales. Notably, the scale type interacted with the modality of presentation. Ratings on structural scales did not differ between the audio-only and audiovisual conditions, indicating that the structural aspects of music can be appreciated independently of whether or not the performance is seen. In contrast, lower ratings were given on the emotional and summary measures when the performance was seen. This is consistent with previous work finding that structure, but not emotion, was conveyed equally well across modality (Vines et al., 2006). Whereas previous studies have found no significant differences between ratings of audiovisual and audio-only performances with minimal stage behavior (Broughton & Stevens, 2009; Wapnick et al., 1998), these results suggest that the visual components of minimal performances do contribute to performance ratings – they contribute negatively.

At first, the finding that audiovisual presentations detract from evaluations seems counterintuitive, because it would be expected that seeing the performance would contribute positively. However, previous research has shown that increased arousal alone does not contribute to higher ratings in music performance evaluations (Broughton & Stevens, 2009; Wapnick et al., 1998). In this case, non-musicians in particular may have given lower ratings when seeing the performer, perhaps owing to their unfamiliarity with pianists’ stage behaviors and
concert performances in general. Additionally, this finding could be carried by the results of the minimal stage behavior condition, in which the lack of expression negatively affected the evaluations – all participants gave low ratings when they saw the minimal performances. These considerations might have contributed to the detrimental effect of seeing the performance, although other factors may also have come into play.

In one of only two effects of expertise (the other being the interaction between stage behavior and expertise described below), the ratings on the different scale types depended on expertise. Musicians and non-musicians used the emotion scales similarly, but musicians gave higher ratings than non-musicians did on the structural and overall evaluation scales. This fits with the idea that formal training in music increases one’s attention to the structural aspects of music. Alternatively, it may be that non-musicians simply had difficulty applying these more technical terms to the performances. However, the non-musicians also did not apply the non-technical evaluative scales that would be accessible to them. Thus, it would seem that musical training also produces greater concern with quality of the performances.

**Stage behavior, modality, and expertise**

Lastly, we turn to the previously studied factors shown in the top left-hand corner of Table 1. The first effect to note, at the top of the first column labeled Stage behavior, is the strong main effect. Ratings on the evaluative scales were lowest for the minimal level of stage behavior, and highest for both natural and exaggerated levels of stage behavior, consistent with previous studies (Broughton & Stevens, 2009; Wapnick et al., 1998, 2000). The exaggerated level of stage behavior was not significantly higher than that of the natural level of stage behavior. This lack of difference invites two possible, non-exclusive interpretations. First, this particular pianist’s exaggerated behavior may contain only slightly more gestures than his natural performances. In other words, his exaggerated stage behavior is close to his personal natural stage behavior. Second, it might be that the exaggerated level of stage behavior is approaching a threshold point, after which higher levels of stage behavior may have negative effects on the evaluations. This is supported by the finding, discussed earlier, that exaggerated performances were rated lower specifically for the Copland piece.

There were no main effects of modality (audio-only and audiovisual were not rated differently overall) or expertise (non-musicians and musicians gave similar ratings overall). However, both of these variables interacted strongly with degree of stage behavior. These interactions are best understood by examining the effects for audio-only and audiovisual conditions separately. In the audio-only condition, non-musicians could not discriminate the different degrees of stage behavior; for them, the stage behavior made no impact when they were only hearing the music. The musicians, in contrast, could distinguish the minimal from the natural and exaggerated performances by just hearing the music. From this, it can be concluded that the pianist’s interpretation of the piece was affected by his intended degree of stage behavior, but these differences were detectable only by trained musicians.

In the audiovisual condition, the musicians showed an even greater effect of stage behavior than in the audio-only condition. The differences musicians noted upon hearing the piece were enhanced by also seeing the performance. When both seeing and hearing the performances, non-musicians now gave higher ratings as a function of the degree of stage behavior, whereas they had not when only hearing the pieces. For non-musicians, then, only the visual component carried information about the degree of stage behavior, confirming the impact of visual information found in previous studies (Dahl & Friberg, 2007; Davidson, 1993; Gillespie, 1997;
Thompson et al., 2008; Vines et al., 2003). Even with visual information, the non-musicians’ discrimination was still not as sharp as that of musicians.

Although differences across the degrees of stage behavior are obvious to everyone audiovisually, only discerning listeners (i.e., musicians) are able to detect these effects without any visual cues. First of all, this suggests that non-musicians attending a concert, and presumably both hearing and seeing the performance, may base their judgment more on what they see than on what they hear. Since the non-musicians did not rate the audio-only excerpts differently, the fact that they rated the excerpts differently in the audiovisual condition suggests that the impact of the added visual cues was so great as to change their evaluation of the entire performance. On the other hand, musicians who only heard the performances could already distinguish the different degrees of stage behavior. Visual information enhanced the magnitude of the differences across the three degrees of stage behavior but did not lead to a different pattern of results, as it did for the non-musicians.

A second implication of these results is that asking the pianist to change his stage behavior changed the quality of the music, in a manner perceptible to musicians. In other words, secondary performance gestures, referring to gestures not directly related to the production of sound, do influence the quality of playing. Although the pianist was asked to vary his playing as little as possible across the three stage behavior conditions, it appears that he was not able to consciously control for the effects of secondary gestures on his playing. This could be analogous to studies in social psychology showing that activating the facial muscles most associated with smiling produced higher ratings of amusement, even without participants’ conscious awareness of smiling (Strack et al., 1988). Restricting or exaggerating movements could similarly have caused the pianist to emote differently, without his conscious awareness of doing so. This information could be of didactic use to musicians in their preparation of repertoire. It may not be practical superficially to exaggerate their stage behavior in performance, but rather, to practice with exaggerated stage behavior and listen to the effect of this modification on their interpretation of the music.

To conclude, we presented here an experiment that investigated multiple factors influencing performance evaluation. This study uncovered overall effects of the individual factors, as well as a complex pattern of interrelationships. These findings raise new questions about how musical performances might be studied empirically, and attest to the vast complexity and subtlety of highly expert musical performances. Future research could focus in particular on the question of how best to describe the emotions conveyed through musical performance and how this depends on the style of the music or the particular piece. The agreement between musicians and non-musicians in choosing subtler emotions over the more general emotions of happiness, sadness, anger, fear, and tenderness, which have traditionally been used in music cognition studies, suggests that music expresses nuanced emotions perhaps more than other forms of communication and social interaction. This may be due to the complexity of music, or to other aspects of performance that are unique to music, such as the fact that a performance can convey not only the emotional intent of the composer but also that of the performer interpreting the piece.

Acknowledgements

This research would not have been possible without the generous help and support of Professors Xak Bjerken, David Dunning, and Kevin Ernste of Cornell University, as well as Professors David Temperley and Elizabeth Marvin, Lecturer William Marvin, and lab technician Trevor deClercq of Eastman School of Music. We thank the two anonymous reviewers for helpful comments on an earlier draft of this paper.
References


