1049

Music Perception
Summer 1997, Vol. 14, No. 4, 419–444

The Aesthetic Quality of a Quantitatively Average Music Performance: Two Preliminary Experiments

BRUNO H. REPP
Haskins Laboratories

A statistically average or prototypical member of a category (e.g., human faces) is often perceived as more attractive than less typical category members. Two experiments explored to what extent this may hold for music performance, an artistic domain in which individuality (i.e., deviation from prototypicality) is highly valued. In Experiment 1, graduate student pianists judged 11 student performances of Schumann's "Träumerei," one of which was created by forming the mathematical average of the other 10. The average performance was rated second highest in quality, even though it was judged second lowest in individuality. In Experiment 2, pianists judged 30 performances of the beginning of Chopin's Etude in E major, synthesized so as to vary only in expressive timing and tempo. The timing patterns were derived from expert pianists' recordings and from casual student performances, and they included separate and combined averages. All three averages received high quality ratings, and the expert average was rated highest of all 30 performances. There was a negative linear correlation between rated quality and individuality. Paradoxically, therefore, the students' expressive timing patterns were preferred over the experts'. Possible explanations of this finding are discussed, such as interactions between timing and other performance parameters and conditions under which conventionality tends to be favored over individuality.

Introduction

The present study was inspired by research on the perceived attractiveness of human faces. In a modern replication of work done more than a century ago by Francis Galton and others (Galton, 1883; Stoddard, 1887; Treu, 1914), Langlois and Roggman (1990) presented subjects with photographs of individual faces as well as with a composite face, constructed by aligning the digitized images of the individual faces and averaging their pixel gray values. Subjects judged the composite face to be more attractive

Address correspondence to Bruno H. Repp, Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695. (e-mail: repp@haskins.yale.edu)

than the majority of the individual faces, and the relative attractiveness of the composite face increased with the number of faces in the sample. Langlois, Roggman, and Musselman (1994) replicated these original findings and ruled out explanations in terms of image-processing artifacts or secondary attributes such as facial symmetry, familiarity, or youthfulness. These authors argued that it was only the mathematical averageness of the composite face that made it attractive. Additional support for this proposition has come from a recent study that used line drawings of faces (Rhodes & Tremewan, 1996).

The explanation of this interesting finding seems to be that artificially generated average faces resemble an ideal or prototype of the human face that individuals have abstracted from the many faces they encounter in their daily lives. Numerous psychological studies have demonstrated the capability of both human adults and infants to form such mental prototypes from exposure to a variety of exemplars of the same category (e.g., Grieser & Kuhl, 1989; Posner & Keele, 1968; Rosch & Mervis, 1975; Strauss, 1979). The prototype usually corresponds to the center of the multivariate category space, and it serves as a point of reference in classifying new category members (Rosch, 1975). The results of Langlois and Roggman (1990) and others (Martindale & Moore, 1988; Martindale, Moore, & West, 1988; Whitfield & Slatter, 1979) indicate that the prototype may also serve as an aesthetic standard. In the case of faces, a biological reason for this finding has been suggested, as even young infants show a preference for faces that are judged attractive by adults (Langlois et al., 1987). More generally, however, it could be a consequence of the fact that the prototype has the smallest average distance from other category members (Barsalou, 1985; Rosch & Mervis, 1975). It also has the smallest average distance from individuals' aesthetic ideals, which are formed from exposure to members of a category and thus are likewise distributed around the prototype. Although any given individual with an off-center aesthetic ideal will find some nonprototypical stimulus more attractive than the prototype, the prototype will be the most attractive stimulus, *on the average.* This may be called the minimal-distance hypothesis.

The present study investigates the applicability of this idea to the domain of classical music performance. As a cultural phenomenon, music performance obeys implicit norms that are handed down from teachers to pupils and are far from arbitrary. In Western art music, a particular composition calls for a performance that respects the notated score as well as the performance tradition(s) associated with its style (Meyer, 1989, 1996). Although musicians generally believe that there are many equally valid ways of playing the same music, there are nevertheless strong constraints on what is acceptable in music performance. Certain eccentric performers may choose to violate these constraints and get away with it, but the majority of performances may be thought of as variations around an implicit norm. Although much more needs to be learned about the distribution of actual performances in the hypothetical multidimensional space spanned by their distinctive characteristics, the performance at the center of that space may be the best instantiation of the hypothetical norm (if there is indeed a single norm). Such a prototypical performance could be approximated by averaging a number of individual performances, and the averaging process would also reduce unintended variation due to musicians' imperfect motor control. Therefore, an average performance should sound not only typical but also especially smooth and controlled, hence aesthetically appealing.

There are reasons, however, for believing that the situation is not quite so simple. Unlike the anatomic features of a face, the expressive characteristics of a music performance are under human control.[1] Therefore, it is possible to *strive toward* an ideal performance and to come ever closer to it as the artist's level of skill and musical insight increases. Most likely, this is what music students do. If all musicians had similar ideals, however, the best artists' performances would approach the prototype and hence become nearly indistinguishable from each other. Clearly, this is not what happens: The most distinguished musicians produce performances that are strikingly different from each other, not maximally similar (see, e.g., Repp, 1992, 1995). Thus there must be a force that counteracts the potential convergence upon a single ideal. This force is a striving towards diversity and individuality, which is characteristic of Western civilization and the romantic tradition in particular (Meyer, 1989).

In the domain of linguistic and literary expression, Bakhtin (1981) has postulated centripetal and centrifugal forces that, respectively, ensure communication and individual expression. Taking his lead from Bakhtin, Bowen (1993) has argued that a similar tension of opposing forces exists in music performance: "Each performance . . . is a unique moment during which the individual struggles to convey both a unique message and a specific musical work" (p. 144). Bowen had especially jazz performance in mind, where the work itself is fluid and undergoes change over time. In the case of a fixed score, "performance tradition" may be substituted for "musical work" in Bowen's statement. In other words, performers must adhere to a conventional norm of expression (to communicate with their audience) and at the same time try to make an individual statement (to sound interesting and different from other artists). Individual musicians will differ in the extent to which they yield to one or the other of these conflicting forces, and also in the ways in which they deviate from the norm and whether they do so

[1] It is worth noting that the studies of facial attractiveness kept expression neutral; they were concerned with composition rather than performance, as it were.

consciously or involuntarily. It is also likely that there are individual differences in musicians' concepts of the norm itself, which may depend on their training and the performances they have been exposed to, which prominently includes their own. Individual and collective performance norms may also change over time (Bowen, 1993).

The aesthetic judgment of performance quality is likely to be subject to similar conflicting forces. Even if the judges had similar notions of an ideal performance and would find such a performance appealing if they heard it, they may also be searching for evidence of an individual statement, perhaps in proportion to the strength of their own desire to deviate from the norm in musical or other activities. Because an average or prototypical performance, by definition, lacks individuality, it may therefore not be completely satisfactory, even if everything "seems right" in it. The relative importance of individuality may be situation-dependent, however: As Levinson (1990) has pointed out, different perspectives may be adopted in evaluating musical performances, depending on the performer's and listener's experience and goals.

The present experiments were a first attempt to investigate the aesthetic appeal of an average music performance relative to that of the individual performances that go into the average. The study was made feasible through MIDI technology, which yields parametric representations of music performances—specifically, of piano music—that can be manipulated and resynthesized. However, several methodological problems had to be faced right away.

First, some music performances differ so radically that it does not make sense to average them: Opposing tendencies in timing or dynamics will cancel, resulting in an average that is unrepresentative and probably unappealing as well. This problem was reduced (if not entirely eliminated) in the present study by making sure through appropriate statistical analyses that the performances averaged were reasonably similar to each other, so that they could be considered "members of the same category". Second, the averaging procedure was restricted to timing (including basic tempo) and—in Experiment 1—dynamics (including dynamic level), arguably ~~the two~~ most important dimensions of expressive performance. Other performance parameters (articulation, onset asynchronies, pedaling, dynamics in Experiment 2) were "regularized" and held constant, mainly for practical reasons. A third problem that can only be acknowledged is that, for a convincing test of the hypothesis under investigation, a larger number of MIDI-recorded performances of the same music may be needed than were available to the author at this time.[2] The temporal extent of music and the

[2] Langlois and Roggman (1990) found that at least 16 faces were needed in order for their composite to be significantly more attractive than the average individual face. However, Rhodes and Tremewan (1996) found no dependence on sample size.

boredom that sets in when the same piece is heard many times also impose practical constraints on sample size. Finally, the competence and consistency of the judges is a major concern. Ideally, they should be experienced concert artists or music critics, but these experts are very busy and difficult to recruit in sufficient numbers. Therefore, the present judges were mostly graduate student pianists, and this will have to be considered in the interpretation of the results.

Experiment 1

METHODS

Materials

Ten individual performances of Robert Schumann's "Träumerei", op. 15, no. 7, complete with the initial repeat, were obtained from a small MIDI data base recorded 12 to 18 months previously. The score is shown in Figure 1. The pianists (P1, P2, . . . , P10) were all piano graduate students at the Yale School of Music and were recorded on an upright Yamaha MX100A Disklavier after a brief rehearsal. Each pianist played "Träumerei" three times (twice in one case), in alternation with three other pieces (see Repp, 1995, for details). The performances were all fluent and expressive but contained a few inaccuracies, owing to the limited preparation (see Repp, 1996c). Their expressive timing and dynamics, although reliably different from individual to individual, followed similar patterns. This was demonstrated in principal components analyses on the respective measurements, each of which yielded a single significant component that accounted for more than 80% of the variance (see Repp, 1995, 1996a).

The MIDI files contained the pitches of the notes played, the times of note onsets (key depressions) and offsets (key releases), their relative intensities (MIDI velocities), and pedal onsets and offsets. To facilitate line-up for averaging and to avoid possible artifacts, the performances were edited and "regularized" as follows: (1) After importing the MIDI data as text into a spreadsheet program, all note offsets and pedal events were filtered out, leaving only the note onsets. (2) All pitch errors were identified and corrected: Wrong notes (very rare) were assigned the correct pitch, missing notes were inserted, and extra notes were deleted. (3) The onsets of nominally simultaneous notes were synchronized, using the onset time of the highest note as the reference. The only exception to this was the fermata chord in bar 22, which was played partially or wholly arpeggio by most pianists and whose notes were allowed to retain their idiosyncratic timing pattern. (4) The note onsets and MIDI velocities of each pianist's three (or two) performances were lined up and averaged, resulting in a single average performance for each pianist. (5) Note offsets were created that coincided with following note onsets, according to the nominal values of the notes in the score. In other words, the articulation was legato throughout. (6) A uniform pedaling pattern was imposed on all 10 performances. This pattern was derived from a performance that was not part of the set (pianist LPH, as analyzed by Repp, 1996b) and was scaled to the expressive timing of each performance so that pedal offsets and onsets occupied constant relative positions within the note interonset intervals (IOIs) in which they occurred.[3] (7) A grand average performance was created by averaging the MIDI data of the 10 individual performances. (8) The MIDI data were converted back into sound, and it was verified through careful listening that all 11 performances were free of artifacts.

[3] Steps 5 and 6 were motivated by the likely interdependence of note offsets and pedal timing and by the difficulty of averaging pedaling data, which often differ qualitatively across performances (see Repp, 1996b).

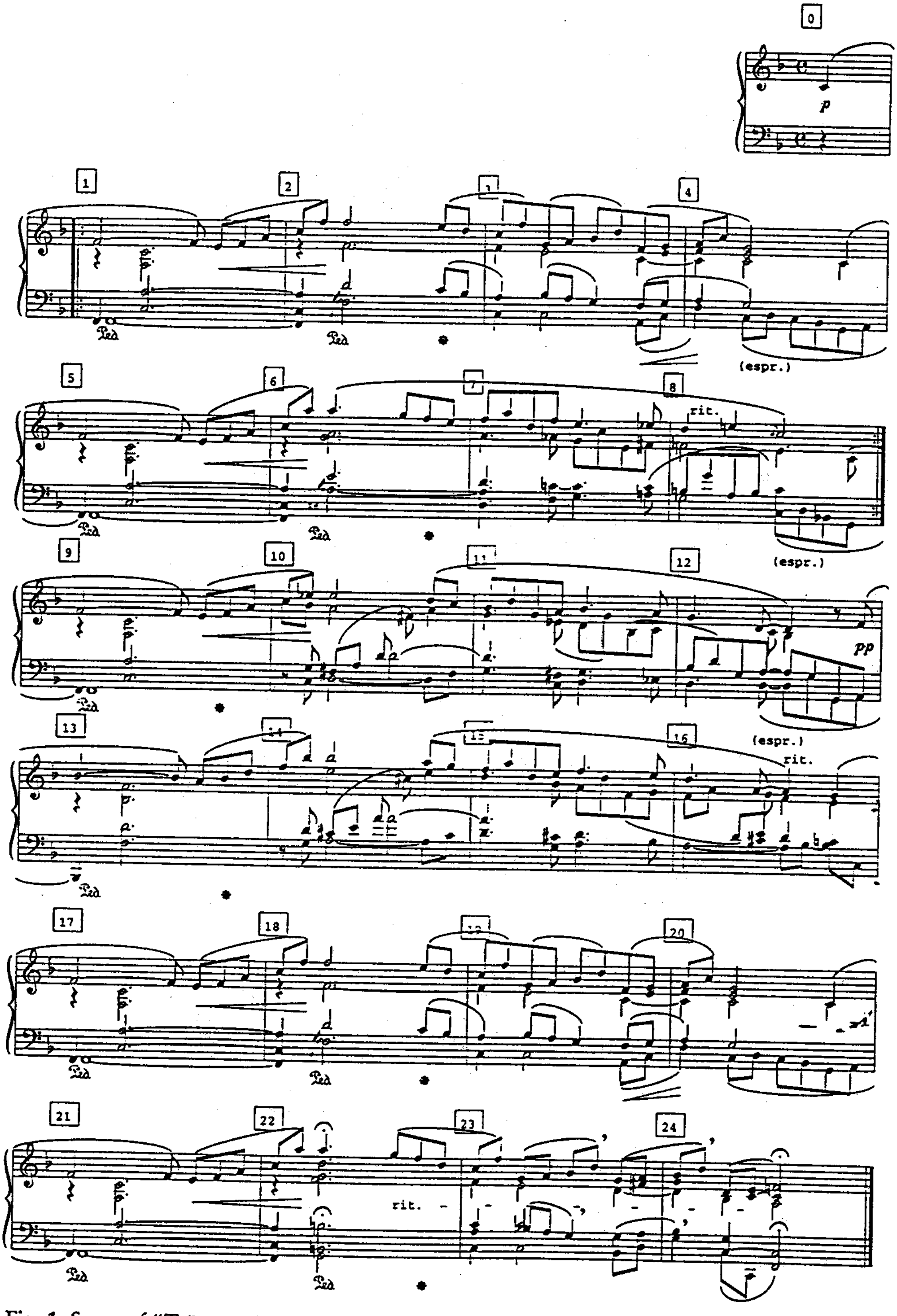

Fig. 1. Score of "Träumerei," op. 15, no. 7, by Robert Schumann.

The performances thus differed only in terms of expressive timing (which subsumes differences in basic tempo) and dynamics (which subsumes differences in overall dynamic level). Chord synchronization, legato articulation, and pedaling were held constant. Moreover, by first averaging across the three (or two) individual performances of each pianist (which were highly similar; see Repp, 1995), uncontrolled variability was reduced in the individual performances. Although no formal comparison was conducted between the original performances and their regularized versions, it is believed that the latter retained the specific individual and expressive characteristics of the former, with little if any loss in quality. On the contrary, whatever significant expressive details may have been lost through regularization were probably more than compensated for by the elimination of unintended irregularities. Graphs of the expressive timing and dynamics of the average performance may be found in Repp (1995) and Repp (1996a), respectively.

The 11 performances were reproduced on a Roland RD-250s digital piano with "Piano 1" sound.[4] This realization was preferred over acoustic recordings from the original Yamaha Disklavier, which would have been subject to ambient noise, room acoustics, and changes in the condition of the instrument since the original recordings were made. It was assumed that the slightly synthetic sound would not seriously impair the expressive qualities of the performances and, in any case, was just another constant factor. The analog output of the digital piano was recorded onto digital audio tape (DAT). The DAT recorder could be programmed to deliver different predetermined sequences to the judges. Twelve random sequences were constructed such that each performance appeared in each serial position, and each performance followed each other performance once, to the extent possible.[5]

Judges

Twelve pianists were enlisted as judges and were paid for their participation. Eight of them were graduate students in piano performance at the Yale School of Music, including five from the original group who had played "Träumerei" (P2, P3, P4, P6, P7). The remaining four judges included one gifted undergraduate (performance certificate student) pianist, one music theorist and piano teacher, and two Ph.D. candidates in musicology, one of whom also reviewed CDs for a classical music magazine.

Procedure

The judges listened to the performances individually in a quiet room over Sennheiser HD 540 reference II earphones at a comfortable intensity. They were given printed instructions, the musical score, and a rating sheet. The instructions described the experiment as a mock piano competition with 11 competitors, from whom 6 finalists should be selected. On the rating sheet, space was provided for five judgments and additional comments about each performance. Judges were asked to provide rough evaluations of each performance in four categories, using the symbols --, -, √, +, and ++, and then to give an overall numerical rating on an 11-point scale ranging from 0 (mediocre) to 10 (outstanding), using decimals or extending the scale if necessary. The four categories were: tempo (ranging from "much too slow" to "much too fast"), dynamics (ranging from "much too weak" to "much too strong"), expression (ranging from "very inexpressive" to "very exaggerated"), and individuality (ranging from "very conventional" to "very unusual"). The midpoint of each scale was labeled "just right."

After a judge had listened to and judged the 11 performances (the "semifinalists"), there was a break of 10–15 minutes during which he or she was recorded playing music on the

[4] An audio file of the average performance (which lasts about 2.5 minutes) as well as the MIDI instructions for all 11 performances may be found on the internet at http://www.haskins.yale.edu/Haskins/MISC/REPP/AP.html.

[5] It would have required 12 performances to do this exactly.

digital piano for a different study. Then the six performances with the highest overall ratings (the "finalists") were arranged in a new quasi-random order and presented for a second evaluation, again taking care to vary the serial position of each performance across different judges. The purpose of this second round was to check on the consistency of the ratings and to give judges an opportunity to revise their first evaluations, if they were so inclined. Of course, different judges generally selected different sets of finalists. It would have been methodologically preferable to present all 11 performances for a second time, but this was considered too taxing for the participants; instead, each judges' ratings for his or her five nonfinalists were carried over from the first round.

RESULTS AND DISCUSSION

Although most judges used the 11-point rating scale efficiently, it was decided to first normalize the data by converting the overall ratings to inverse rankings ranging from 10 to 0, with ties. Figure 2 shows the average semifinal and final rankings, with standard errors for the latter. (Semifinal ratings for nonfinalists were carried over to the finals before computing averages and standard errors.) It can be seen that the highest average score was achieved by P10, with the average performance (AP) in second place, tied with P3.[6] The average scores of the next six performances were lower,

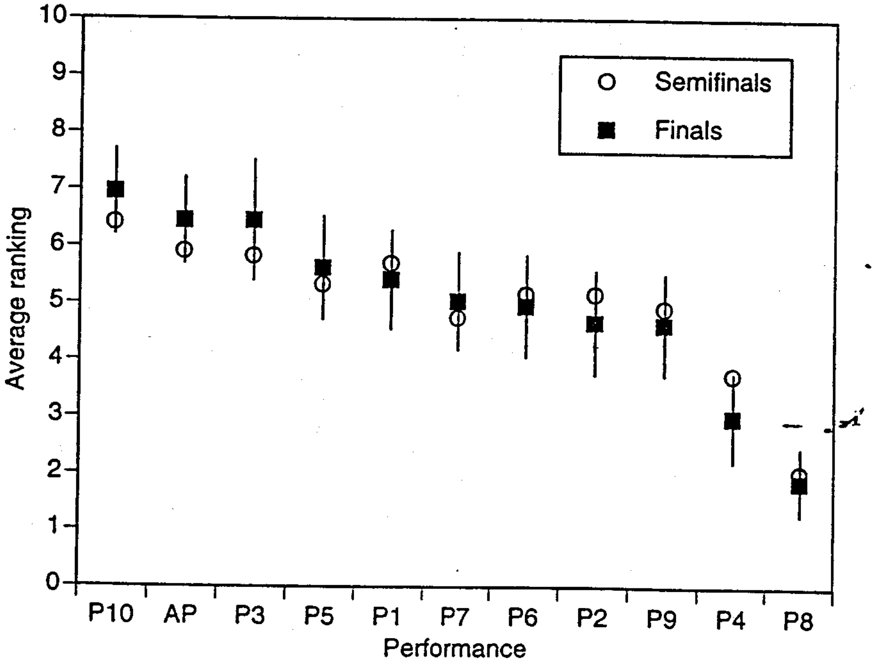

Fig. 2. Average semifinal and final rankings of the 10 individual performances and the average performance (AP), with standard error bars for the final rankings.

[6] The author also served as a pilot subject; he ranked AP second after P10 in the semifinals, but fourth after P10, P7, and P5 in the finals.

but there was considerable statistical overlap. Only P4 and P8 received distinctly lower ratings, evidently because of their tempos. (P8 was the slowest performance by far, whereas P4 was the fastest.) The three top-ranked performances extended their lead in the finals, relative to the semifinals.

The relatively narrow range of the average rankings for 9 of the 11 performances and the overlap of their uncertainty regions indicate considerable individual differences in the judges' preferences and possibly low reliability of judgments. However, the correlation between the average semifinal and final rankings (not including any carried-over semifinal judgments) of the 11 performances was 0.81 ($p < .01$). Eight of the 12 judges showed moderate to high positive correlations between their semifinal and final rankings (although, with only four degrees of freedom, most fell short of significance), three showed little relationship, and one a negative correlation. However, it probably would be wrong to conclude that the latter judges did not know what they were doing. Given the length and considerable similarity of the performances, a change in preferences among the six top-rated performances could occur very easily on repeated hearing, and there may also have been effects of order of presentation, which differed in the semifinal and final rounds.

Of the five pianists whose own performances of 12–18 months ago were in the set (unbeknownst to them), two (P2, P7) preferred their own performance over all others, both in the semifinals and in the finals. One (P4) ranked her own performance second in the semifinals and third in the finals, which is also noteworthy in view of the low overall ranking of this very fast rendition. One pianist (P6) ranked her own performance second in the semifinals and fourth in the finals. Only P3, whose performance was generally well liked, paradoxically gave it the lowest rating.[7] With this exception, the results indicate that, despite the regularization, the performances preserved individual characteristics that the pianists tacitly recognized as being close to their own aesthetic ideal.

The qualitative judgments in the four specific categories (tempo, dynamics, expression, and individuality) were converted into numerical ratings ranging from 0 (--) to 4 (++), with 2 corresponding to "just right" (√). The average semifinal and final ratings are shown in Figure 3. The tempo ratings (Figure 3a) faithfully reflect the actual tempos of the individual performances, as reported in Repp (1995); the linear correlation between final ratings and average tempo (beats per minute) is 0.98. The dynamics ratings (Figure 3b) likewise reflect the objectively measurable dynamic levels of the individual performances, as reported in Repp (1996a); the correlation between final ratings and average MIDI velocity is 0.91. It is clear from

[7] Because pianists were expected to like their own performance, inclusion of their self-ratings was not expected to bias the data in favor of AP. Because of P3's unexpected dislike of her own performance, however, AP slipped back into third place after P10 and P3 when the self-ratings were excluded from the data.

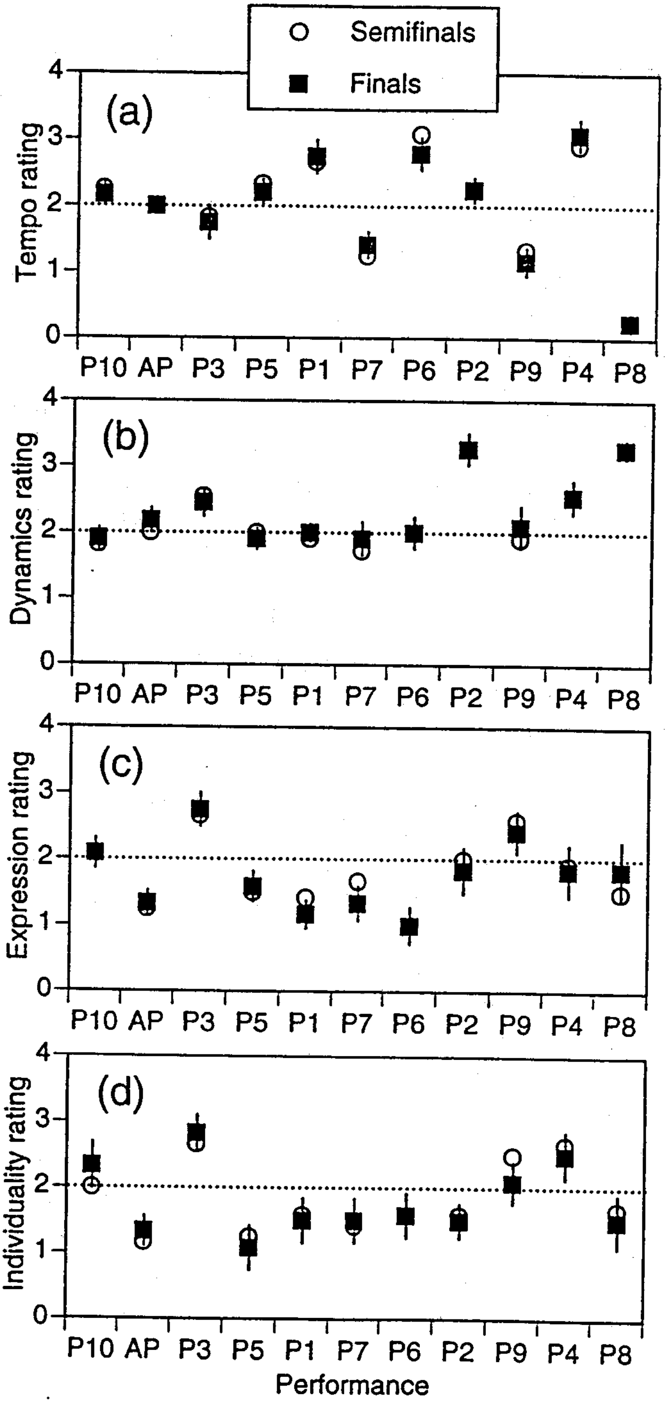

Fig. 3. Average semifinal and final ratings on four specific categories, with standard error bars for the final ratings. The dotted horizontal line indicates the center of each scale ("just right"). The dimensions are (a) tempo (slow–fast), (b) dynamics (soft–loud), (c) expression (inexpressive–exaggerated), and (d) individuality (conventional–unusual).

Figures 3a and 3b that the less preferred performances were deviant in tempo, dynamics, or both.

The final expression ratings (Figure 3c) and individuality ratings (Figure 3d) were correlated (r = 0.74, p < .01), and several judges mentioned that these dimensions were difficult to keep separate. What is noteworthy here is the low rating of AP on both dimensions. Evidently AP was perceived as relatively inexpressive and conventional, yet its quality was ranked as highly as that of P3, which received the highest expression and individuality ratings. P10, the most preferred performance overall, was rated only slightly above average in expression and individuality. Its special quality may have been one of dynamic differentiation and subtlety, which was not well captured by the rating categories used here.[8]

The judges' additional written comments were also revealing. They were able to point out specific strengths and weaknesses of P3 and P10, the most highly rated individual performances, such as: "a little exaggerated at times in terms of timing," "too stoppy before downbeats, but very expressive," "beautiful leading into B-flat cadence," "doesn't hold dotted quarter notes long enough," and so on. By contrast, AP elicited fewer and more general comments: "too 'straight' for me," "good flow," "boring—sounds rather predictable," "very sensitive—mood is right, too" (by the only judge who gave AP the highest final rating), "dullness not offensive because degree of variation reduced," "very beautiful, nice and plain but not at all boring," and "very simple but good."

In summary, the results of this first experiment demonstrate that the procedure of linearly averaging the note IOIs and MIDI velocities of a number of reasonably similar individual performances generates a performance that is musically meaningful and free of distortions. Musically trained listeners evidently perceive it as quite pleasing, even though it lacks individuality. Numerically, the average performance ranked higher than 8 of the 10 individual performances, which is consistent with the minimal-distance hypothesis outlined in the Introduction. On the other hand, the advantage of the average performance over most individual performances was not statistically reliable, owing to large variability in judgments, some of which presumably reflects individual differences in aesthetic criteria.[9] Although greater statistical reliability of results could have been achieved by testing additional judges, pianists are difficult to lure into the laboratory in sufficient numbers. Therefore, it was considered more efficient and informative

[8] Interestingly, P3 and P10 were the least accurate performances originally, which benefited most from editing and regularization. It seems that these pianists cared more about expression than about accuracy. The opposite type is represented by P4, P6, and P8, who were very accurate but somewhat inexpressive.

[9] In order to determine how many different preference patterns underlie the 12 pianists' judgments, the semifinal rankings were subjected to a principal components analysis. Five components emerged as significant (i.e., had eigenvalues greater than 1). It is likely that different judges gave different weights to different aspects of performance.

to conduct a new experiment focusing on a single expressive dimension in a shorter music excerpt, in the hope that this would reduce the variability of aesthetic judgments.

Experiment 2

The task in Experiment 1 was quite difficult, owing to the length and relative similarity of the student performances. However, it is known from earlier studies (Repp, 1992, 1995, in press) and from informal observations that famous pianists' performances are much more diverse than student performances, especially in their expressive timing. Experiment 2 was a first attempt to incorporate expert performances in a test of the minimal-distance hypothesis. However, because these performances were available only in the form of acoustic recordings, Experiment 2 had to be restricted to variations in timing, with dynamics held constant across performances, as there is currently no way of getting accurate estimates of expressive dynamics in several voices from an acoustic waveform. Because a larger number of performances than in Experiment 1 was to be used, the music was restricted to a short excerpt. Thus, the listeners in Experiment 2 judged the quality of renditions of a passage, varying in timing and tempo only.

The experiment included the timing patterns of both student and expert performances and thus allowed a comparison between the two, as well as between their averages. Perrett, May, and Yoshikawa (1994) have shown that the average of faces judged to be highly attractive is perceived as even more attractive than a grand average face. The present comparison between student, expert, and grand averages had a similar goal. It was expected that the expert average would be preferred over the student average, with the grand average in between. This prediction was based on the implicit assumption that the expressive timing of commercially recorded expert performances is generally superior to that of minimally rehearsed student performances. (It will be seen that this was a simplistic assumption.)

In order to learn more about the criteria that pianists use in making aesthetic judgments, the judges in Experiment 2 were also asked to play the test excerpt, to illustrate their own preferred tempo and timing. It was hypothesized that their evaluations would reflect, to a considerable extent, the relative similarity of the test stimuli to their own performance. A question of secondary interest was whether the pianists' performances would be influenced in any way by their exposure to the test stimuli; therefore, samples of their playing were obtained at three different points during the experiment.

METHODS

Materials

The excerpt chosen was the beginning (measures 1–5) of Chopin's Etude in E major, op. 10, no. 3, a piece as popular as Schumann's "Träumerei" and familiar to virtually all pianists. The score is shown in Figure 4a. The final chord was extended through the second beat of bar 5 to provide a proper conclusion to the excerpt. Fifteen expert performances

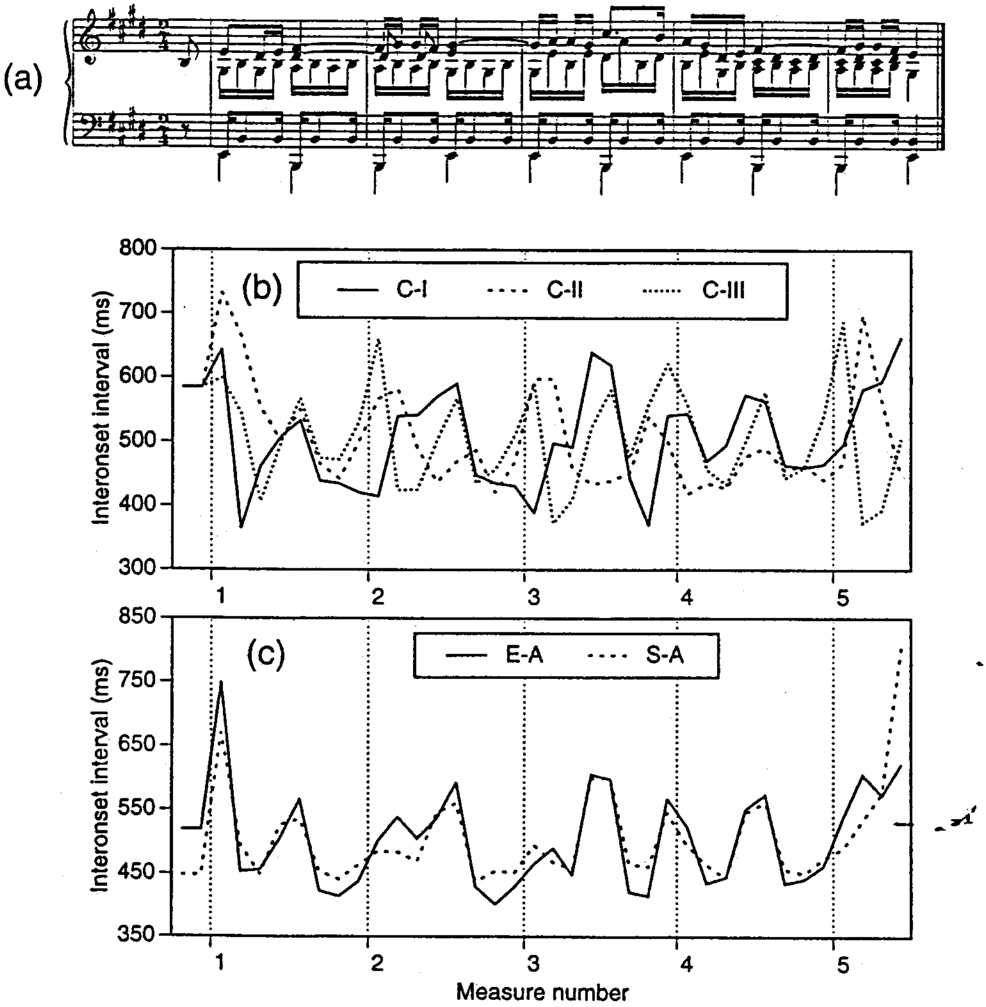

Fig. 4. (a) The beginning of Chopin's Etude in E major, op. 10, no. 3, with a terminal chord. (b) The timing profiles representing the three rotated principal components of the expert performances (component scores rescaled into milliseconds). (c) The timing profiles of the expert average (E-A) and student average (S-A) performances. All initial eighth-note upbeats are plotted at half their total durations (i.e., as two sixteenth notes).

TABLE 1
Varimax-Rotated Principal Component Loadings of 15 Expert Performances

Code	Pianist (record label)	C-I	C-II	C-III
E13	Maurizio Pollini (DGG 413 794-2)	0.911	0.293	0.028
E14	Sviatoslav Richter (Philips 420 774-2)	0.895	0.132	0.235
E5	Van Cliburn (RCA Victor 60358-2-RG)	0.810	0.151	0.408
E1	Vladimir Ashkenazy (London 414 127-2)	0.795	0.296	0.168
E4	John Browning (RCA Victrola 60131-2-RV)	0.782	0.228	0.325
E9	Adam Harasiewicz (Philips Classics 422-282-2)	0.726	0.483	0.218
E3	Idil Biret (Naxos DDD 8.550364)	0.699	0.414	0.171
E10	Vladimir Horowitz (RCA Victor 60376-2-RG)	0.687	0.035	0.541
E7	Yuri Egorov (Peters International PLE 121)	0.632	0.201	0.629
E8	Philippe Entremont (Sony Classical MLK 64057)	0.615	0.575	0.268
E6	Alfred Cortot (EMI Classics 2905401)	0.601	0.658	–0.172
E12	Vlado Perlemuter (Nimbus NI 5095)	0.179	0.867	0.184
E2	Wilhelm Backhaus (Pearl GEMM CD 9902)	0.175	0.744	0.276
E11	Louis Lortie (Chandos CHAN 8482)	0.214	0.656	0.596
E15	Tamás Vásáry (DGG 2535 266)	0.193	0.309	0.865

Note.—Highest loadings are in boldface.

(E1, E2, ..., E15) of the Etude were obtained from various LPs and CDs (see Table 1). Their beginnings were sampled into a Macintosh Quadra 660AV computer, and their successive note onset times were hand-measured in a waveform display, using the highest note in each chord as the reference. With the exception of the initial eighth-note upbeat, the music moves in sixteenth notes, so all IOIs are nominally equal. Their actual durations, however, varied a great deal, as was to be expected in performances of such a slow, highly expressive piece from the romantic period.

To determine whether there was more than one pattern of timing variation, a principal components analysis was done on the timing profiles (series of IOIs) of the expert performances.[10] The analysis yielded three principal components with eigenvalues greater than 1 (the conventional criterion of significance), which together accounted for 80% of the variance. Varimax rotation was applied to these components, leading to three rotated components that accounted for 42%, 22%, and 16% of the variance, respectively. These orthogonal components represent different timing patterns underlying the performances. Component timing profiles (Figure 4b) were obtained by multiplying the standardized component scores by the average within-performance standard deviation of the original IOIs and adding these products to the grand mean IOI duration. Component I (C-I) is characterized by deceleration during each of the five melodic gestures and acceleration during the long melody notes that terminate these gestures. Component II (C-II) instead shows deceleration during the long melody notes and acceleration during melodic gestures. Component III (C-III) combines these two strategies and thus shows acceleration-deceleration patterns within each beat. The three components are mutually uncorrelated.

[10] Because this analysis was based on correlations among timing profiles, differences in basic tempo did not play a role. The initial eighth-note upbeat was excluded from the analysis. There was remarkable variation in initial upbeat duration; several pianists played it effectively as a sixteenth note.

The individual pianists' component loadings (i.e., the correlations of their timing profiles with the three statistically extracted component profiles) are shown in Table 1. It is evident that 10 pianists had their highest loading on the first component, four on the second, and one on the third component. Several pianists had nearly equally high loadings on two components. However, there was a clear break between the first 11 and the last 4 pianists with regard to their loadings on the first component, which is the most common underlying timing pattern. Therefore, to avoid averaging artifacts as much as possible, it was decided to form an expert average (E-A) timing pattern by averaging the note IOIs of the first 11 performances only. The last four performances were included in the experiment but not in the average.

Naturally, the E-A timing profile (Figure 4c) was similar to the C-I profile (Figure 4b). Although both profiles represent estimates of typical expert timing, it was suspected that C-I might contain statistical artifacts due to the rotation of the principal components (note the very short IOIs in bars 1 and 3) and therefore be less appealing aesthetically than E-A. In order to gain information about the viability of the component profiles as performances, all three (C-I, C-II, C-III) were included as stimuli in the experiment. They were given initial upbeats of the same duration, representing the average duration of the upbeats of all expert performances.

In addition to these expert-derived timing patterns, the experiment included the timings of nine student performances (S1, S2, . . . , S9), their average (S-A), and a grand average (G-A) formed by averaging the expert and student averages. The student performances had been recorded in the intermission of Experiment 1. The pianists played the beginning of the Chopin Etude from the computer-generated score (Figure 4a) on the Roland RD-250s digital piano after a brief warm-up. Each student played the excerpt three times, and the performances were recorded in MIDI format. Subsequently, the performances were edited and regularized as in Experiment 1, and the three performances of each student were averaged to reduce random variation.

The nine student timing profiles were subjected to a principal components analysis, which yielded a single significant component, accounting for 85% of the variance. Thus the student performances were much more homogeneous than the expert performances (see also Repp, 1995, in press), and their average could be formed without problems. The S-A timing profile (Figure 4c) differed from the E-A profile in three ways: its initial upbeat and downbeat were shorter, the tempo modulations during the passage were somewhat less pronounced, and the final sixteenth note was much longer. This last difference was due to the fact that the students had played the excerpt with a final chord whereas the experts had played the passage in the context of the complete Etude, where there is no final chord but continuing sixteenth-note motion. However, the ending of the E-A performance (which did end with a chord) did not sound unnaturally abrupt.

Because information about expressive dynamics was not available for the experts, it was decided to hold dynamics constant for all performances. The constant dynamic pattern was obtained by averaging the MIDI velocities of the nine student performances (each of which was the average of three original performances). The MIDI velocity values were aligned with the MIDI pitches and the IOIs in a spreadsheet program, and note onset times for resynthesis were calculated by cumulating the IOIs. The MIDI files were synthesized on the Roland RD-250s digital piano with "Piano 1" sound. Careful listening confirmed that all stimuli were free of audible artifacts.

In summary, this experiment included 24 individual performances (E1, . . . , E15; S1, . . . , S9), three averages (E-A, S-A, and G-A), and three expert component performances (C-I, C-II, and C-III), a total of 30 stimuli that differed only in tempo and timing.[11] They were

[11] Readers are invited to listen to the stimuli on the internet by accessing http://www.haskins.yale.edu/Haskins/MISC/REPP/AP.html. The performances are provided both as audio files and as MIDI files.

recorded onto DAT twice in the same random order, with several seconds of silence between stimuli.

Judges

Twelve pianists served as judges. Eight of them were graduate students and two were performance certificate students at the Yale School of Music. Three of them had been judges in Experiment 1 and thus had contributed a performance of the Chopin excerpt. The two remaining judges were the author and his research assistant, a postdoctoral musicologist.

Procedure

The experiment was again set up like a miniature competition, this time in three rounds. In the first round, each judge listened to all 30 performances, starting at some randomly determined point on the DAT, and rated each performance in terms of overall quality and individuality. The quality judgment was made by circling a number on a 10-point scale whose endpoints were labeled "mediocre" and "outstanding"; the individuality rating was made by placing a check mark below one of five attributes ("highly conventional," "conventional," "somewhat individual," "individual," or "highly individual"). In the "semifinals," the 12 performances with the highest ratings were presented again in a different random order, constructed by programming the DAT recorder. (If there were ties for 12th place, candidate performances were taken in serial order.) In the "finals," the six performances rated most highly in the semifinals were presented a third time, again in a different random order. (To break ties for 6th place in the semifinals, the first-round ratings were consulted.) The judges were told that only tempo and timing varied in the stimuli.

Before each of the three rounds, each of the 10 student judges played the Chopin excerpt (Figure 4a) three times on the digital piano. A brief warm-up period was provided at the beginning. The students were told that they were expected to illustrate their preferred tempo and expressive timing of the music. The performances were recorded in MIDI format. Later, the note onsets were filtered out, successive notes were labeled (in each chord, the note with the highest pitch was chosen), and IOIs were calculated.

RESULTS AND DISCUSSION

The judges' ratings were again translated into inverse rankings (from 30 to 1, with ties), and the lower rankings were carried over from earlier to later rounds. Figure 5 shows the average rankings. The spread was much wider here than in Experiment 1, indicating better agreement among judges. The top-ranked performance was the expert average, although it was virtually tied with three other performances, one by an expert and two by students. Figure 5 shows that the latter three stimuli gained substantially in their ratings as the experiment progressed, whereas E-A did not. Lower (although not significantly lower) ratings were given to a group of six performances containing the student and grand averages, three students, and only one expert. The next group of four, ranked significantly lower than the four leaders by the criterion of nonoverlapping standard-error bars, contained two students and two experts. Thus, surprisingly, most of the students were ranked higher than the experts. In particular, only two students (S5, S8) received lower-than-average rankings; all other performances

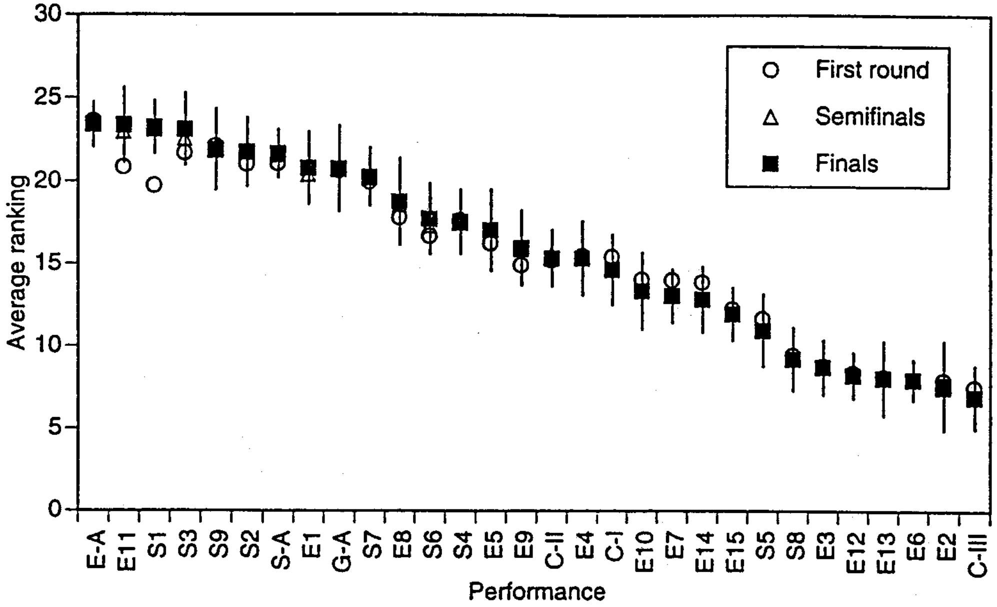

Fig. 5. Average rankings of 30 test performances in three successive rounds, with standard error bars for the final rankings.

with low rankings were expert-derived. Among the least appreciated timing patterns were those of experts resembling C-II (E2, E6, E12). Paradoxically, however, E11, which also correlated with C-II (see Table I), was the most highly rated expert profile. The C-II profile itself received an average ranking. C-I was clearly rated lower than E-A, which confirms that E-A is a better estimate of typical timing than the statistically extracted and rotated component pattern. C-III was the lowest-ranked performance, whereas the single expert profile resembling it (E15) was judged somewhat more positively.

The average first-round and semifinal rankings (here without carried-over lower rankings) showed a significant correlation (r = .49, p < .01), especially after two "outliers" (stimuli that made it into the semifinals only once and twice, respectively, but received unusually high ratings there) were removed (r = .82, p < .001). There was no significant correlation, however, between either first-round or semifinal average rankings and average final rankings. This may have been because many performances entered the finals just once or twice or because of the judges' difficulty in distinguishing among the most highly rated performances. At the individual level, a somewhat different picture emerged. Only two judges showed a significant correlation between their first-round and semifinal rankings, whereas three others showed moderate positive correlations. No judge showed a significant correlation between semifinal and final rankings, but 11 of the 12 correlations were positive, which suggests some weak relationship. On the whole, these comparisons suggest considerable instability, although not necessarily randomness, of individual judgments. Nevertheless, as Figure 5 suggests, not much systematic change in average rankings occurred across the three rounds, S1 and E11 being the main exceptions.

Of the three judges who (unknowingly) evaluated their own timing patterns, S2 gave it the final rank 27 (30 in the semifinals), S9 gave it 26, and S8 gave it 21.5. Each of these rankings is higher than the average ranking of the same stimuli (cf. Figure 5) and thus provides evidence that the pianists resonated to their own expressive timing. Clearly, however, these three judgments contributed little to the overall preference for student over expert performances.

The average individuality ratings are shown in Figure 6. It is evident that the judges considered the expert timings more individual than the student timings; there is hardly any overlap. They also recognized that performances resembling C-II (E2, E12) and C-III (E15) as well as C-III itself and, to some extent, C-II were atypical. (The most highly rated performance, E2, was also extremely slow.) A clear exception was again E11, whose timing profile resembled C-II but nevertheless received a relatively low individual-

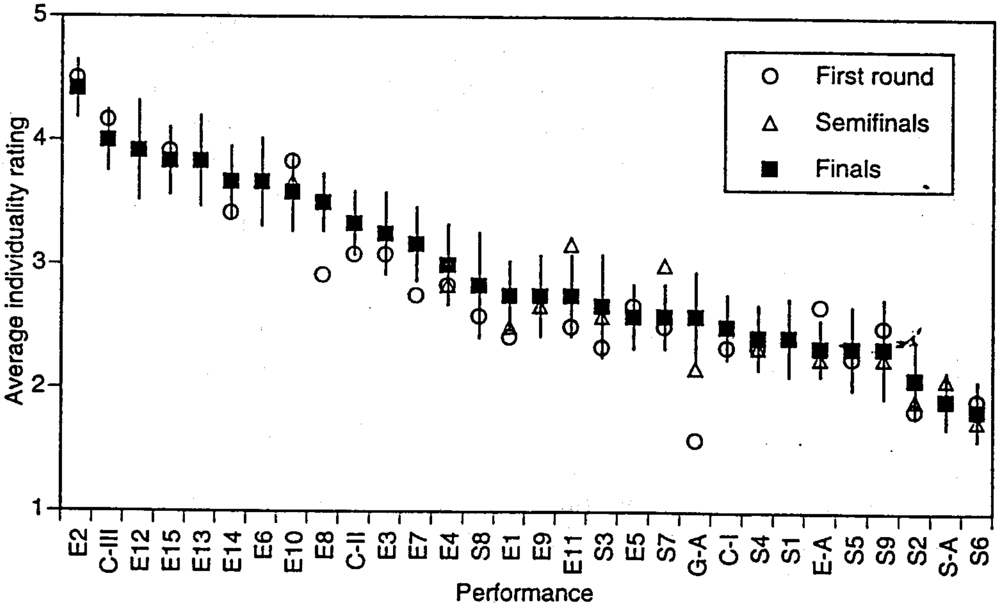

Fig. 6. Average individuality ratings in the three rounds, with standard error bars for the final ratings.

ity rating. The average performances received even lower individuality ratings. The grand average was rated lowest in the first round but gained remarkably in the next two rounds. The student average remained very low on the scale, however, with the expert average in between. These results show that the judges were well aware of the greater individuality of the expert performances but nevertheless gave them relatively low quality ratings. The correlation between the final quality rankings and individuality ratings was –.73 ($p < .001$).

The timing profiles of the nine performances provided by each of the 10 student judges during the experiment were subjected, without the upbeat, to analyses of variance with the factors Position (36 IOIs), Take (3 levels), and Repetition (random, 3 levels). Four pianists showed a significant Position by Take interaction, indicating a change in the preferred timing profile in the course of the experiment. However, inspection of the profiles revealed only minor differences in profile shape which had reached significance because of the high number of degrees of freedom (70,140). Therefore, a single average timing profile was calculated from each pianist's nine performances. A principal components analysis on these 10 individual profiles showed that a single component accounted for 87% of the variance, just as in the analysis on the student performances employed in the experiment.

A multiple regression analysis was subsequently conducted, to determine to what extent the judges' rankings of the 30 test performances could be predicted from the relative similarity of the test performance to their own performance. The first-round rankings were taken as the dependent variable here. The two predictor variables were the absolute difference in basic tempo and the profile similarity (product-moment correlation) of the judge's performance and the performance judged. The results of this analysis are shown in Table 2. The *average* values of the two predictor variables accounted for 67% of the variance in *average* first-round rankings (multiple $R = .82, p < .001$), with profile similarity making a slightly higher contribution than tempo difference. At the individual level, the regression analysis yielded significant multiple correlations for eight of the 10 student judges. For four of them, both predictor variables made a significant contribution, for two tempo difference only, and for two profile similarity only. Figure 7 compares the average first-round rankings of the 10 student judges and the average individual multiple-regression predictions. It is evident that three test performances (E-A, E11, C-II) were liked much better than predicted, whereas several others (all expert performances) received significantly lower rankings than predicted. Although the multiple regression analysis did predict higher-than-average rankings of the average performances (which supports the minimal-distance hypothesis), it is interesting to note that E-A in

TABLE 2
Standardized Regression Coefficients (Correlations) and Multiple *R* for Multiple-Regression Predictions of First-Round Rankings from Profile Similarity and Absolute Tempo Difference

Judge	Profile	Tempo	Multiple *R*
MYL	0.214	−0.403*	0.450*
MM	0.506**	−0.432**	0.666***
PYW	0.549**	−0.351*	0.628**
PYL	−0.011	−0.544**	0.545**
RK	0.429**	−0.533**	0.680***
HS	0.367*	−0.263	0.475*
AT	0.359	−0.110	0.372
LB	0.399**	−0.651***	0.734***
TC	0.392*	0.035	0.394
AS	0.493**	−0.122	0.511*
Average rankings	0.643***	−0.528***	0.816***

*$p < .05$; **$p < .01$; ***$p < .001$

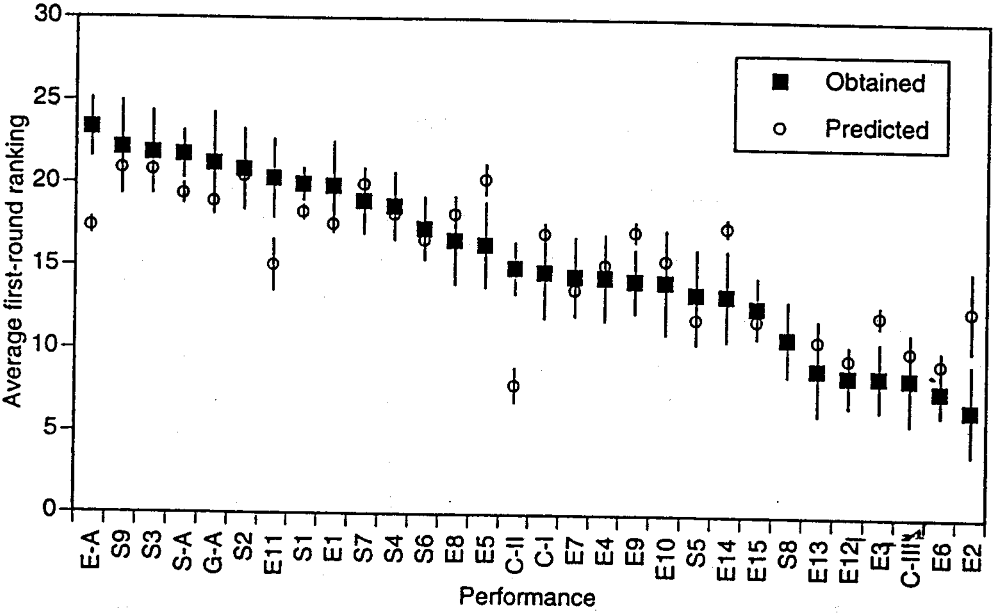

Fig. 7. Average first-round rankings by the 10 student judges, and average individual regression predictions based on tempo difference and profile similarity, with standard error bars.

particular fared much better than predicted. Evidently, some additional factors play a role; they may include the degree of profile modulation and the relative smoothness of these modulations.

General Discussion

The present experiments represent a first attempt to assess the aesthetic quality of a quantitatively average music performance, with special attention given to tempo and timing. The results provide tentative support for the hypothesis that an average performance can sound more appealing than the majority of the performances that go into the average. The effect was more striking for the set of expert performances in Experiment 2 than for the student performances in both experiments. It seems possible that the latter effect, like that obtained for faces (Langlois & Roggman, 1990), will increase in magnitude when a larger number of performances is used. The findings lend support to the idea that an average music performance represents a prototype and an aesthetic ideal, at least in the sense that it is most similar, on the average, to individual aesthetic ideals.

However, the present results may be specific to the materials and the methodology that has been used, and this will have to be investigated in further research. In particular, the results of Experiment 2 present a paradox to which the remaining discussion will be devoted: Although the expert average was the highest-ranked performance (which seems to parallel the finding of Perrett et al., 1994, that the average of attractive faces is more attractive than a grand average face), the individual expert timing profiles were ranked lower than those of most student performances. This was quite unexpected, and a number of possible explanations come to mind.

To some extent, the result must have been due to the fact that the judges were student or amateur pianists whose own performances were more similar to the student than to the expert performances. This is the minimal-distance hypothesis again. If the judges in Experiment 2 had been famous pianists, would they have preferred the expert timings over the student timings? There may never be an opportunity to find out, but the following armchair experiment is instructive: Let us assume that the 15 famous pianists whose performances formed the basis of the expert stimuli served as judges, that their individual aesthetic ideals for the Chopin excerpt are instantiated in their performances, and that they give the same relative weights to tempo and expressive timing as the average student judge. Their rankings can then be predicted by computing the absolute tempo differences and profile similarities between their performances and the test performances, and by inserting the averages of these numbers into the multiple regression equation for the student judges' average rankings. The results of this exercise showed, first of all, a very narrow range of average rankings, implying that the results of an experiment conducted with expert judges would probably yield unclear results, owing to large individual differences in aesthetic criteria corresponding to large individual differences in expressive timing and tempo. Second, the predictions did show a slight average advantage of expert over student performances and of the expert average over the student average. Thus, the expert performances indeed fared somewhat better when the judges were hypothetical experts than when they were real students. Third, both expert and student averages were predicted to receive higher rankings than any of their respective constituent performances. Thus, the hypothetical data would support the minimal-distance hypothesis. This exercise shows then that the low rankings of most expert performances may be partially due to the fact that the judges were not expert pianists themselves, but this can hardly be the whole story. Surely, student and amateur pianists are able to appreciate great pianists' performances, and well-prepared famous pianists also generally play better than minimally rehearsed student pianists. Most of the original expert recordings certainly sounded quite beautiful, even though they had recognizably the same timing as the experimental performances derived from them. The fact of the matter is that most expert timings did not sound as appealing once they were removed from their original context and imposed on a synthetic "carrier." The question is why. The following explanations are discussed in order of *increasing* plausibility.

One factor that comes to mind immediately is that the expert stimuli were derived from recordings of the complete Chopin Etude and thus did not have the large final ritard that the student performances exhibited (see Figure 4c). Such a large ritard may have been perceived as more appropriate in the experimental stimuli, which ended with a chord. This would be a rather trivial explanation. However, it seems highly unlikely that the extent of the final ritard had much of an influence on the judges' evaluations. The ritard occupied only the final 5% or so of a performance, and the aesthetic impression was largely based on the timing during the body of the passage. The high ranking of E-A also contradicts this explanation. Readers not convinced by this argument will probably agree after listening to the stimuli (see footnote 11).

A second possibility is that the expert timing profiles received low rankings because they contained more random variation than did the student performances. Each expert provided only a single performance, whereas the student profiles represented the averages of three individual performances. Expert timing profiles, moreover, were subject to human error in waveform measurement, whereas the student profiles were measured by computer. This explanation would be consistent with the high ranking of E-A, which could be attributed to the averaging-out of random variability. However, as pointed out earlier, the timings of the experts' original recordings sound very similar to those of the experimental stimuli, yet they seem more appealing in the original context. Also, considering the high degree of control that outstanding pianists possess, a single performance is surely representative of their intentions. Therefore, this explanation does not seem wholly convincing.

A third, more interesting possibility is that listening to a number of different renditions of the same music somehow induces a preference for conventional over less typical performances. In other words, the judges' aesthetic ideals may have drifted towards the prototype during the experimental session. Although this may be indeed so, it must be considered that the students' ideals (judging from their own performances) were in the vicinity of the prototype to begin with, so that increased conservatism could not have had much of an effect.

A fourth explanation that may be much closer to the truth and that needs to be investigated in future research is that there are interactions between timing and dynamics, both in performance and in aesthetic judgment. The stimuli of Experiment 2 had a constant dynamic pattern which represented the average MIDI velocities of the student performances (Repp, 1996a). This constant pattern, being itself highly conventional, may have provided a better match for the conventional timing of the students and the average timing profiles than for the less conventional timing of the experts. The methodology adopted in Experiment 2 was based on the implicit assumption that timing can be divorced from dynamics. There is evidence for a correlation between timing and dynamics within performances (Todd, 1992): Performers typically play louder when they speed up (in the middle of a phrase) and softer when they slow down (at phrase boundaries). However, the present study is concerned with individual differences: If a pianist exhibits an unusual timing profile, does this mean his or her dynamics were unusual, too? A relevant analysis of the "Träumerei" performances used in Experiment 1 (Repp, 1996a) revealed only a very weak connection between individual differences in timing and dynamics. Listening to the original expert recordings does not reveal striking individual differences in dynamics between performances exhibiting very different timing profiles. Nevertheless, research is planned to investigate possible interactions between timing and dynamics, which is an important issue in its own right.

A fifth possible explanation is that the synthetic sound of the digital piano, together with the "regularization" of the performances (elimination of onset asynchronies and of variations in articulation and pedaling), created a restricted and artificial sound environment within which only relatively conventional timing patterns sounded acceptable to listeners. Perhaps the appreciation of unconventional timing is possible only in the context of the sound of a fine concert grand piano, and conversely only a fine instrument may offer the incentive to produce original timing patterns. After all, it is the sound that is timed! This is a very intriguing possibility that deserves further study.

All of the explanations discussed so far regard the judges' preferences for the student over the expert performances as some kind of methodological artifact. One final possibility to consider, however, is that the results of Experiment 2 do tell the truth: Perhaps most experts' timing in the Chopin phrase (and only their timing) was indeed not as "good" as the students' timing. After all, the Chopin piece is extremely well known, almost hackneyed, and experienced pianists perhaps cannot stand any more hearing it played in a conventional way. Thus they deliberately distort its timing to give it a "new" shape that helps remove the staleness from the music and stimulates jaded listeners, *even though this new shape is less beautiful by conventional standards* (and the artists know it). Alternatively, experienced artists may develop ingrained habits of expressive timing that deviate more or less from the norm, either as a consequence of a deliberate effort to establish themselves as an individual voice within the artistic community or as the result of a kind of a natural selection process favoring diversity among concert artists. The consequences of such habits for the timing of a phrase would be largely unconscious, and the artist may in fact believe that his or her unconventional timing represents the norm.

The hypothesis that the aesthetic ideal coincides with the mean of some statistical distribution seems at variance with the common notion that an ideal music performance is almost impossible to achieve. How can these ideas be reconciled? One factor may be complexity and/or difficulty. A single dimension of a short piece of music is likely to have a central prototype. For example, the ideal tempo for a piece of music usually lies in the center of the distribution of actual tempos encountered, some of which are perceived as too fast by most listeners, others as too slow. It is relatively easy to get close to the prototypical tempo, keeping in mind that individual performers have different notions of the ideal tempo that are themselves distributed around "the" ideal tempo. However, in a complex multimovement work such as Bach's Goldberg Variations, it is much more difficult to approximate the ideal tempo for every single variation in the work while at the same time trying to establish proper tempo relationships and contrast among the variations. Some of these goals may be incompatible, so that a compromise may have to be reached. Because of the many possible solutions to such a complex problem, the distribution of actual performances in the multidimensional space of possibilities becomes extremely thin and therefore no longer implies a central prototype. There are then only individual ideals, which may be more or less clearly formulated and may no longer be points but areas in the hyperspace, permitting alternative solutions. Also, any aspect of performance that poses exceptional difficulties for musicians is unlikely to have a central prototype, simply because most performances fall short of the ideal, rather than being distributed around it.

A second possible reason for the elusiveness of performance ideals is the psychological phenomenon of adaptation. The same considerations or forces that lead experienced artists to deviate from a general norm in the first place may also lead them to deviate from their personal aesthetic ideals, so that their performances and ideals keep changing over time. Individual artists may differ greatly in the relative flexibility of their aesthetic goals. As the central tendency of a dynamic system, the norm or prototype itself may change over time (Bowen, 1993).

In conclusion, although the results of the present study must be considered very preliminary, they raise interesting questions about music performance in the Western classical-romantic tradition that warrant further study from both psychological and musicological perspectives.[12]

References

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (M. Holquist & C. Emerson, Trans.; M. Holquist, Ed.). Austin, TX: University of Texas Press.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. Journ*al of Experimental Psychology: Learning, Memory, and Cognition, 11,* 629–649.

Bowen, J. A. (1993). The history of remembered innovation: Tradition and its role in the relationship between musical works and their performances. *Journal of Musicology, 11,* 139-173.

Galton, F. (1883). *Inquiries into human faculty and its development.* New York: Macmillan.

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology, 25,* 577-588.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1,* 115-121.

Langlois, J. H., Roggman, L. A., Casey, R. J., Ritter, J. M., Rieser-Danner, L. A., & Jenkins, V. Y. (1987). Infant preferences for attractive faces: Rudiments of a stereotype? *Developmental Psychology, 23,* 363-369.

Langlois, J. H., Roggman, L. A., & Musselman, L. (1994). What is average and what is not average about attractive faces? *Psychological Science, 5,* 214-220.

Levinson, J. (1990). Evaluating musical performance. Chapter 15 in *Music, art, and metaphysics* (pp. 376–392). Ithaca, NY: Cornell University Press.

Martindale, C., & Moore, K. (1988). Priming, prototypicality, and preference. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 661–670.

Martindale, C., Moore, K., & West, A. (1988). Relationship of preference judgments to typicality, novelty, and mere exposure. *Empirical Studies of the Arts, 6,* 79–96.

Meyer, L. B. (1989). *Style and music.* Philadelphia: University of Pennsylvania Press.

Meyer, L. B. (1996). Commentary. *Music Perception,* 13, 455–483.

[12] This research was supported by NIH grant MH-51230. I am grateful to Linda Popovic and Lisa Robinson for assistance, to the participating pianists for their time and patience, and to José Bowen, Janet Hander-Powers, Nigel Nettheim, and Linda Popovic for helpful comments on an earlier draft of the manuscript.

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature, 368,* 239–242.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

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,* 2546–2566.

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

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

Repp, B. H. (1996b). Pedal timing and tempo in expressive piano performance: A preliminary study. *Psychology of Music, 24,* 199–221.

Repp, B. H. (1996c). The art of inaccuracy: Why pianists' errors are difficult to hear. *Music Perception,* 14, 161–184.

Repp, B. H. (in press). Expressive timing in a Debussy Prelude: A comparison of student and expert pianists. *Musicae Scientiae.*

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological Science, 7,* 105–110.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7,* 532–547.

Rosch, E., & Mervis, C. D. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Stoddard, J. T. (1887). Composite photography. *The Century, 33,* 750–757.

Strauss, M. S. (1979). Abstraction of prototypical information by adults and 10-month-old infants. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 618–632.

Todd, N. P. McA. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America, 91,* 3540–3550.

Treu, G. (1914). Durchschnittsbild und Schönheit. *Zeitschrift für Ästhetik und allgemeine Kunstwissenschaft, 9,* 433–448.

Whitfield, T. W. A., & Slatter, P. E. (1979). The effects of categorization and prototypicality on aesthetic choice in a furniture selection task. *British Journal of Psychology, 70,* 65–75.