A microcosm of musical expression. III. Contributions of timing and dynamics to the aesthetic impression of pianists' performances of the initial measures of Chopin's Etude in E Major

Bruno H. Repp
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511-6695

(Received 11 January 1999; revised 8 April 1999; accepted 15 April 1999)

Four judges repeatedly assessed the overall aesthetic quality of more than 100 recorded performances of the opening of Chopin's Etude in E major on a 10-point scale. The judgments, which exhibited reasonable reliability and modest intercorrelations, were entered into regression analyses with 16 independent variables derived from earlier objective analyses of the expressive timing and dynamics of the performances [Repp, J. Acoust. Soc. Am. **104**, 1085–1100 (1998); **105**, 1972–1988 (1999)]. Only between 9% and 18% of the variance in the judges' ratings was accounted for. By contrast, timing variables accounted for 53% of the variance in one judge's ratings of synthesized performances that varied in timing only and mimicked the timing patterns of the original performances. These results indicate, first, that the aesthetic impression of the original recordings rested primarily on aspects other than those measured (such as texture, tone, or aspects of timing and dynamics that eluded the earlier analyses) and, second, that very different patterns of timing and dynamics are aesthetically acceptable for the same music, provided that other, aesthetically more crucial performance aspects are present. © *1999 Acoustical Society of America.* [S0001-4966(99)04607-X]

PACS numbers: 43.75.St, 43.75.Mn [WJS]

But all beautiful melodies share essential and multiple properties. Not least is a psychological property which perhaps can best be described as representing a kind of spiritual alienation. For all melodies tend to remove themselves from the common ground of mundane experience, revealing their special spice and character by the usual devices of touch and rubato . ... *If they are too passively integrated into the general mesh of sound, they sacrifice the prophetic, tender, and urgent qualities which mark them as independent of the collective reality.*

(Sherman, 1997, p. 45)

INTRODUCTION

The initial measures of Chopin's Etude in E major, op. 10, No. 3 (Fig. 1)[1] contain one of the most famous and beautiful melodies in the whole piano literature, and the epigraph from Russell Sherman's collection of insightful pianistic observations seems to apply perfectly to it. Many great pianists have struggled to impart to this melody the subtle and exalted expression it demands, but few have succeeded to every listener's satisfaction.

In two preceding articles, Repp (1998, 1999) has presented a detailed analysis of the expressive timing and dynamics of 115 commercially recorded performances of this famous melody. The *timing profile* of a performance was defined as its sequence of tone inter-onset interval (IOI) durations (without the initial upbeat), and its *dynamic profile* was defined as the sequence of peak sound levels immediately following tone onsets in the amplitude envelope. The grand average timing and dynamic profiles were considered typical and representative because many individual performances had timing and/or dynamic profiles that were similar to the grand average profiles. Principal components analyses were conducted to identify different ways in which individual performances deviated from these typical expressive patterns. These analyses yielded principal components (i.e., mutually uncorrelated profile shapes, four for timing and five for dynamics) that were interpreted as expressive strategies underlying the individual differences among the performances.

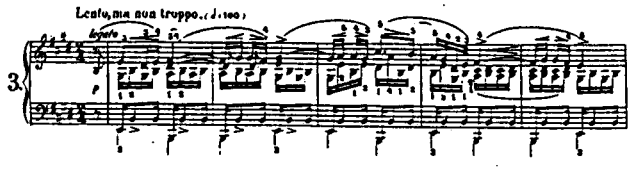

FIG. 1. The initial five measures of Chopin's Etude in E major, op. 10, No. 3 (Herrmann Scholtz edition, Leipzig: Peters, 1879). For the present purposes, the except ends on the second beat of bar 5.

While some individual performances instantiated these strategies in pure form, as it were, the majority had profiles that represented a mixture of several strategies. Indeed, the most typical individual profiles (those most similar to the grand average profile) by definition result from a weighted combination of all strategies. There was no evidence of distinct clusters or major gaps in the distributions of weights given to these strategies, which led to the conclusion that the

 0001-4966/99/106(1)/469/10/$15.00

individual differences in expression did not arise from categorically distinct structural interpretations. Rather, it was argued that they represent different expressive shapings or characterizations of a single musical structure (cf. Shaffer, 1995). Interestingly, there were few significant relationships between expressive timing and dynamics across the many performances. These two important parameters seem to be controlled independently at the phrase level, which opens up a wealth of expressive possibilities.

In addition to its specific pattern over time, each expressive parameter has two more global properties (mean and standard deviation) that are also continuously variable. For timing, these are the basic tempo (inversely related to the average IOI duration) and the within-profile variation in IOI duration; for dynamics, they are the basic dynamic level and the within-profile dynamic variation. Although basic dynamic level could not be assessed meaningfully in the recorded performances, individual differences in the other three aspects were also discussed in the preceding articles (Repp, 1998, 1999).

These earlier analyses were objective and descriptive; nothing at all was said about the aesthetic impressions made by the various patterns of timing and dynamics. Nevertheless, these patterns clearly serve an aesthetic goal: Each pianist is trying to realize a particular conception of the music, and listeners are invited to share this conception and derive pleasure from it. It is often believed, especially among music theorists and music psychologists, that the most important function of expressive variation is to clarify and disambiguate musical structure (see, e.g., Berry, 1989; Clarke, 1985; Palmer, 1989). There may be specific contexts in which this is true, but it is hardly the case in the present musical excerpt, whose structure seems quite unambiguous. Moreover, it is not obvious what aesthetic benefits can be derived from the perception of structure as such. Music theorists and analysts may find intellectual pleasure in being able to hear structural detail or resolved ambiguities, but this is very different from the intuitive, almost visceral response of the ordinary music lover to musical expression (see Cook, 1990; Levinson, 1997). Scruton (1997) has argued convincingly that the perception of expression involves relating musical processes to real-life activities via imaginative metaphors that often remain implicit and ineffable (see also Sloboda, 1998). Musical motion, resulting from the temporal unfolding of its structure, is perceived in terms of life experience, just as a performer's life experience may provide metaphors that help give his or her performance an individual expressive shape. The listener's internal resonance to the musical motion is a kind of "sympathetic dance" (Scruton, 1997) that need not be, and often cannot be, described in words.

Clearly, aesthetic perception is a highly subjective and incompletely understood process, as is the creative process underlying the individual expression produced by a performer. However, aesthetic evaluation is somewhat more straightforward and more amenable to scientific study. It is the process of expressing one's relative liking for or preference among different art objects, here different performances of the same musical composition. Musical performance evaluation is usually engaged in by persons who are familiar with the work and its expressive possibilities; often they are musicians who have played the music themselves. Just as artists have a conception of the music that they are trying to realize in performance, so judges may have a conception that they expect to hear realized. In the simplest case, therefore, a judge's liking of a performance will reflect the extent to which the artist's conception (as conveyed by the performance) matches his or her own conception—that is, the extent to which he or she "resonates" to the performance.

However, this account is oversimplified in at least two ways. First, while the resonance metaphor may apply to any single expressive dimension, performances have multiple dimensions which may meet the judge's expectations in different degrees and which may be assigned different weights (see Levinson, 1990; Thompson *et al.*, 1998). The overall aesthetic impression, then, is a complex function of the listener's resonances to each of these dimensions. Consequently, two very different performances may be liked equally. Second, one must take into account that individuality (originality) is valued in artistic performance, at least in the standard Western classical repertoire. On purely statistical grounds, the best match to most listeners' expectations will be provided by performances that are close to the norm, because expectations are abstracted from experience with many performances and thus are close to the norm, too (Repp, 1997). Nevertheless, and paradoxically, a performance that merely delivers what is expected will often not be fully satisfactory, perhaps because this is too common an occurrence. True, one occasionally hears a performance that realizes the norm so perfectly that it may be considered definitive (by some listeners, at least). However, listeners usually expect to hear some evidence of originality, something unexpected and surprising, for which there are many possibilities, more or less convincing. What exactly makes an unexpectedly original performance convincing to one particular listener, but perhaps not to another, is a very complex question. In some cases, it could perhaps be traced back to the listener's own conception being biased in the performance's direction, but often a listener is simply "won over" by a great artist's very personal interpretation because it reveals as yet unimagined expressive possibilities.

Both a performer's and a listener's conception of a musical work may change over time, sometimes from one performance or hearing to the next. Some experienced artists may have multiple conceptions available at the same time, being able to choose one or the other at will, and the same may be true for some listeners, especially if they are performing artists as well. This possibility of multiple standards further complicates aesthetic judgment. Nevertheless, such judgment is rarely arbitrary. While it may be expected to be less reliable and to show larger individual differences than many other kinds of perceptual judgment, there are real differences in aesthetic quality among music performances, even at the highest level. Although it is difficult to prove through objective analysis alone that some performances are better than others, there will often be some degree of consensus among competent judges (cf. Thompson *et al.*, 1998). Given such a consensus—but even in its absence—one may then ask whether there are any objectively measurable per-

formance aspects that correlate with the judges' evaluations. Different correlations for different judges will reflect their different criteria and biases.

The aim of the present study was to determine the extent to which objectively measured aspects of expressive timing and dynamics contribute to the overall aesthetic impression of the initial phrase of Chopin's Etude in E major.[2] Repeated evaluations of most or all previously analyzed performances were obtained from four judges, on a 10-point rating scale. First, as a prerequisite to the subsequent statistical analyses, the reliability of each judge's ratings was determined across two or more sessions, and the intercorrelations among the judges' ratings were calculated. Musical aesthetic judgment, perhaps because of the aura of subjectivity that envelops it, has rarely been subjected to even this kind of basic analysis (though see Wapnick *et al.*, 1993; Thompson *et al.*, 1998), certainly not for such a large sample of expert performances. Subsequently, the judges' ratings were entered into a regression analysis, with the objective performance measurements as the independent variables, to determine how much of the systematic variance in the ratings could be accounted for. Finally, the relation of the judges' evaluations to several sociocultural characteristics of the artists was examined briefly. Although these characteristics had been found to be only weakly related to some aspects of expressive timing and dynamics, it was considered possible that they were related to other aesthetically relevant aspects of the performances and hence to the judges' ratings.

One of the judges also evaluated a set of synthesized performances that differed in timing only and mimicked the timing patterns of the original performances. The purpose of this exercise was to determine to what extent the timing variables account for aesthetic ratings when all other performance aspects are held constant. In a much smaller set of synthesized performances of the same Chopin excerpt, Repp (1997) had found that listeners preferred typical over atypical timing profiles. However, he had not obtained any ratings of the original performances from which the timing profiles had been derived. Now a direct comparison could be made, albeit for only a single listener, to examine whether there is indeed a preference shift in favor of typical timing in synthesized versus original performances.

I. METHOD

A. Judges

In view of the difficult and time-consuming nature of the task, only four judges participated. Initially, the author (BR) thought of undertaking the evaluation all by himself and therefore proceeded to collect a large amount of data for himself. To increase the generality of the findings, however, he then enlisted three additional judges (MC, NN, OS) who volunteered to participate but could not be expected to invest quite as much time and effort.

All judges were well qualified for the task and knew the Chopin Etude well. BR is an amateur pianist and music lover, 53 years old at the time, with some 35 years of experience in the informal judgment of piano performances in concert and on recordings. MC is a well-known scientist and musician, 71 years old. He has a master's degree from the Juilliard School and was a successful concert pianist in his youth. He possesses great insight into composers' individual rhythmic characteristics and into performance expression generally. NN is a pianist and musicologist, 57 years old, with special expertise on historical piano performance. OS is a 30-year-old music critic and amateur pianist who reviews live and recorded performances in several periodicals. His special interest is early music, but he is certainly familiar with Chopin's idiom. He provides a useful contrast to the three older judges, in that he is more representative of contemporary aesthetic tastes. The judges also provide a broad international spectrum: BR grew up in Austria and came to the U.S. at age 25; MC was also born in Austria but left at an early age and lived mainly in Australia and the U.S.; NN is Australian; OS is Israeli.

B. Procedure

The procedure followed by MC, NN, and OS shall be described first. These judges were mailed a tape containing 102 performances of the Chopin excerpt.[3] MC was sent a digital tape, NN and OS cassette tapes, together with instructions. The judges were asked to familiarize themselves with the range of performances by listening to at least part of the tape before starting the evaluation. The evaluative judgments were to be made on a 10-point scale, where "1" represented the lowest rating and "10" the highest, with decimal ratings a possibility. The judges were encouraged to use the whole range of the scale and to evaluate all 102 performances in a single session. The sound quality of the recordings was to be ignored as much as possible. In addition to the overall ratings, written comments on each performance were invited. An improvised answer sheet was used to record the ratings and comments. On the tape, each recorded excerpt faded out during bar 6 or 7. However, it was emphasized that only the first 5 bars were to be evaluated and that the tape should be stopped at that point for each excerpt. (Many pianists accelerate the tempo in the following bars, and the way in which this is done may have a considerable influence on the overall aesthetic impression.) A second evaluation session was conducted on another day (MC: 4 days later; NN: one week later; OS: one month later). The procedure was the same, but whereas the first session started with performance No. 1, the second session started with performance No. 52. The judges were urged not to look at their previous ratings. All three judges were completely blind to the pianists' identities, which were revealed to them only afterward.

BR had heard the performances previously when analyzing their timing and therefore could not avoid recognizing some of them. He also included 15 additional performances that he had used in an earlier study (Repp, 1997) and had not heard recently.[4] To make his evaluations as unbiased as possible, BR deleted the pianists' names from the computer files and retained only their serial numbers. He subsequently recognized only about 10% of the performances, which is not believed to have had an influence on his judgments. He evaluated the 117 performances 10 times, each time in a different random order. The first three sessions were on successive days, the fourth session three days later, the fifth session one week later, and sessions 6–10 on successive days

one week later.[5] He listened to each excerpt twice before giving a rating and added written comments during the last six sessions. Each rating session took about 2.5–3 h.

C. Synthesized performances

Following his evaluation of the original performances, BR evaluated an additional set of computer-generated performances that varied in expressive timing only. These performances were created by imposing the measured timing profiles onto a constant "carrier." In the course of earlier experiments, a number of graduate student pianists had played the Chopin Etude excerpt (ending with a chord on the second beat of bar 5) on a Roland RD-250s digital piano and had been recorded via a Musical Instrument Digital Interface (MIDI). The MIDI instructions of one of the best performances were imported as text into a spreadsheet program, and the onset of the highest note in each metrical (sixteenth-note) position was located. From the onset times of these "primary" notes, inter-onset intervals (IOIs) were calculated. All other MIDI events (other note onsets, note offsets, pedal depressions, and pedal releases) were then linked to the nearest primary note onset by expressing their times of occurrence relative to it. Next, the measured IOIs of one of the expert performances were substituted for the IOIs of the student performance. The primary note onset times were then re-calculated by cumulatively adding the new IOIs, and the times of occurrence of all other MIDI events were obtained by adding their (positive or negative) relative times of occurrence to the new onset times of the primary notes. Thus all secondary note onsets, note offsets, and pedal events remained in their original (absolute) temporal relationships to the nearest primary note onset. The MIDI velocities (relative intensities) of all notes remained unchanged. The result was a synthesized performance with a timing pattern closely resembling that of an original performance.[6] In that way, 115 synthesized performances were created and played back under computer control on the Roland RD-250s digital piano (with "Piano 1" sound). Their aesthetic quality, although inferior to that of most original excerpts, was satisfactory and, moreover, their sound was free of noise or distortion. Their range of aesthetic quality seemed somewhat reduced relative to that of the original performances, since only a single expressive dimension varied.

BR evaluated the synthesized versions in five additional sessions spread over eight days, two months after his evaluation of the original excerpts. He first created a set of randomly renumbered copies of the MIDI files, so that he was totally blind to their identity. He recognized the timing of only a few highly individual performances. In each session, he listened to the excerpts in a different random order, gave each excerpt a rating on the 10-point scale (usually playing it only once, but repeating it if uncertain), and took notes.[7]

II. RESULTS

A. Use of rating categories

Figure 2 shows the frequency distribution of the four judges' ratings. There were considerable individual differences. BR used the whole range of the scale, with the most

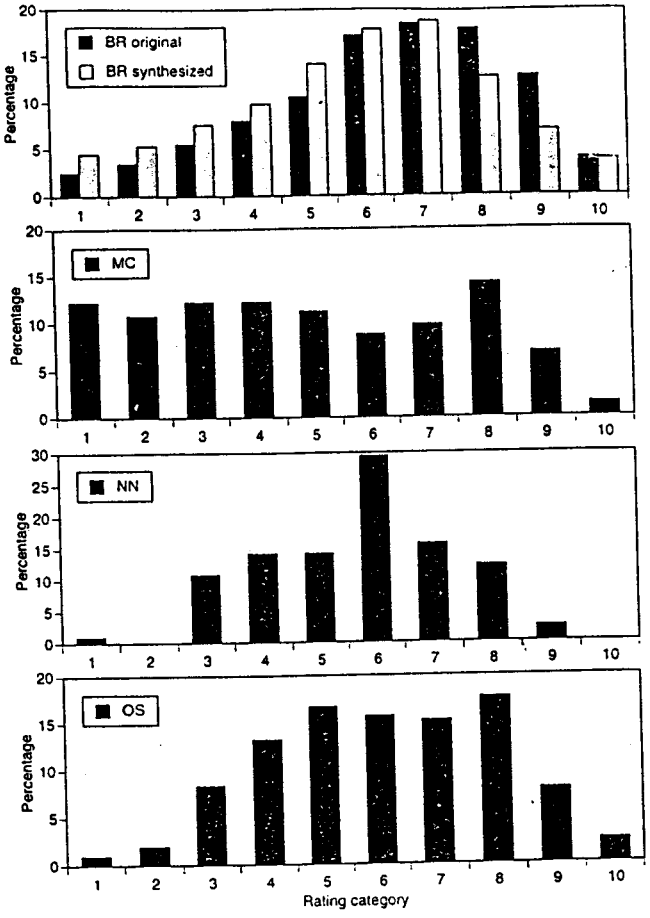

FIG. 2. Frequency histograms of the four judges' ratings.

frequent ratings being in the range of 6–8. Despite a subjective recalibration of the scale, he gave slightly lower ratings to the synthesized than to the original excerpts. (The difference would probably have been larger if original and synthesized performances had been intermixed.) MC was reluctant to use the highest categories (he never gave a straight "10") but frequently gave very low ratings, which resulted in an almost rectangular distribution of ratings. NN, by contrast, basically restricted himself to the middle categories (3–8) and gave many "6" ratings. OS also favored the middle categories, but occasionally gave very high or low ratings. MC, NN, and OS all made occasional use of decimals, whereas BR gave only integer ratings. No transformations were applied to the ratings in the following analyses.

B. Reliabilities and intercorrelations

To determine the reliability of BR's ratings of the original excerpts, the pairwise correlations among his 10 sets of ratings were computed. These 45 correlations ranged from 0.48 ($p<0.001$) to 0.77, with an arithmetic average of 0.64. Although this is not a very high reliability, it does show some consistency in what were rather difficult judgments. Of course, the average ratings were more reliable than those from single sessions. BR's average ratings of sessions 1–5 and 6–10 correlated 0.80, and the average ratings of his odd and even sessions correlated 0.92. The fact that the first correlation is lower than the second indicates some systematic

TABLE I. Intercorrelations among the four judges' average ratings of the original excerpts ($N=102$).

	MC	NN	OS
BR	0.56[a]	0.52[a]	0.30[b]
MC	-----	0.37[a]	0.21[c]
NN	-----	-----	0.29[b]

[a] $p<0.001$.
[b] $p<0.01$.
[c] $p<0.05$.

changes in preferences over time. The reliability of the ratings averaged over all 10 sessions can be estimated from either of these two "split-half" correlations by means of the Brown–Spearman formula $[2r/(1+r)]$; one estimate is 0.89 and the other 0.96. In other words, at least 80%(r^2) of the variance in BR's average ratings was systematic.

For each of the other three judges, only a single correlation between two sets of ratings could be computed. These correlations were 0.76 for MC, 0.50 for NN, and 0.56 for OS (whose two sessions were one month apart). For their average ratings, the Brown–Spearman formula gives reliability estimates of 0.86, 0.67, and 0.72, respectively. The systematic variance thus ranged from 45% to 74%.

The 10 pairwise correlations among BR's 5 sets of ratings of the synthesized performances ranged from 0.66 to 0.75, with an average of 0.69. The reliability of the average ratings may be estimated to be about 0.90. It is noteworthy that aesthetic quality differences resulting from timing alone could be judged at least as consistently as differences based on a larger number of performance parameters. The probable reason was that there were fewer expressive parameters to attend to and fewer possibilities for changes in their relative weights over time. Of course, practice could also have played a role.

The correlation between BR's average ratings of the original and synthesized excerpts was 0.52 ($p<0.001$). This is substantially lower than the reliability of either set of judgments and thus indicates significant changes in aesthetic preferences from one set to the other.

The intercorrelations among the four judges' average ratings of the original excerpts are shown in Table I. Although significant, they are not high and bear witness to large individual differences in aesthetic judgment. As expected, OS stands out as the one most different from the other three judges. The complete average ratings of all judges are listed in the Appendix.

C. Correlations of judges' ratings with performance characteristics

These correlations are shown in Table II. The columns represent the four judges' average ratings of the original excerpts, BR's ratings of the synthesized excerpts (BRSYN), and the difference between these ratings and BR's ratings of the original excerpts (BRDIFF). The rows represent 16 independent variables derived from the earlier performance analyses: the basic tempo estimated from the median IOI duration (TEMPO); the duration of the initial upbeat relative to the average IOI duration (UPBEAT); the within-performance coefficient of variation of the IOIs (relative modulation depth of timing, RELMOD); the performance's loading on the first unrotated principal component for timing (TUPC1) or, equivalently, its correlation with the grand average timing profile—an index of its typicality; its loadings on the four Varimax-rotated principal components for timing (the "timing strategies," TPC1–TPC4); the standard deviation of the melodic peak sound levels (MELSD); the difference between the average peak sound levels of melody and accompaniment (MELACC); the loading on the first unrotated principal component for dynamics (DUPC1)—the typically index of the dynamic profile; and the loadings on the five Varimax-rotated principal components for dynamics (the "dynamic strategies," DPC1–DPC5). No correlations are shown between BRSYN and the dynamic performance variables because there were no dynamic differences among the synthesized excerpts. However, it did make sense to compute correlations between BRDIFF and the dynamic variables, because these changes in ratings may reflect the effect of *removing* different aspects of expressive dynamics.

TABLE II. Linear correlations between judges' average ratings and objectively measured performance characteristics ($N=117$ for BR, $N=102$ for MC, NN, and OS). See text for explanation of abbreviations.

	BR	MC	NN	OS	BRSYN	BRDIFF
TEMPO	−0.34[a]	−0.42[a]	−0.22[c]	−0.03	−0.59[a]	−0.30[b]
UPBEAT	0.17	0.11	0.10	0.29[b]	0.24[c]	0.08
RELMOD	−0.05	−0.17	−0.12	−0.24[c]	−0.34[a]	−0.31[a]
TUPC1	0.21[c]	0.16	0.31[b]	0.19	0.48[a]	0.31[a]
TPC1	0.06	0.16	0.18	0.10	0.35[a]	0.31[a]
TPC2	0.13	0.07	0.12	0.00	−0.10	−0.23[c]
TPC3	0.05	−0.15	0.07	0.04	0.08	0.04
TPC4	−0.06	0.03	−0.18	0.04	0.02	0.07
MELSD	0.07	0.13	0.11	0.07		−0.09
MELACC	0.02	0.03	−0.03	−0.10		0.01
DUPC1	−0.05	−0.06	−0.06	0.07		0.15
DPC1	−0.04	−0.05	−0.00	0.07		0.23[c]
DPC2	0.12	−0.05	−0.01	−0.17		−0.15
DPC3	0.06	0.08	0.09	0.17		−0.02
DPC4	−0.16	0.06	−0.13	0.10		0.17
DPC5	0.02	−0.13	0.02	−0.12		−0.12

[a] $p<0.001$.
[b] $p<0.01$.
[c] $p<0.05$.

Surprisingly few of the many correlations reached significance, and none were very high. The ratings of three of the judges (BR, MC, NN) were negatively correlated with TEMPO, indicating a preference for slower over faster performances.[8] This preference was even more pronounced in BR's ratings of the synthesized excerpts. Since there was reason to expect a curvilinear (convex) relationship between tempo and preference, with the most preferred tempo somewhere in the middle, second-order polynomials were also fitted to the tempo data. This indeed improved the fit for three of the judges; only MC showed no quadratic trend at all.[9] The "quadratic correlations" (square roots of the variance accounted for by the quadratic function) were 0.50 for BR, 0.48 for NN, and 0.34 for OS (all $p<0.001$). The quadratic functions for these three judges were similar in shape and had their maxima in the vicinity of 30 quarter-note (or 60 eighth-note) beats per second, which is very close to the

average median tempo of the performances (Repp, 1998: Fig. 2). BRSYN and BRDIFF did not show a significant quadratic trend.

Weak positive relationships were obtained between the relative duration of the initial upbeat (UPBEAT) and the ratings of OS and BRSYN, indicating a slight preference for longer upbeats. Hints of a convex quadratic relationship, indicating a preference for intermediate upbeat durations, were found for BR ($r=0.22$, $p<0.05$) and NN ($r=0.26$, $p<0.01$).

Relative modulation depth of timing (RELMOD) was negatively correlated with the ratings of OS, BRSYN, and BRDIFF, which indicates a preference for weaker over stronger timing modulation, especially in the synthesized performances. Since here, too, a curvilinear relationship could be expected, quadratic functions were again fitted to the data. Marginally significant relationships of the expected convex form emerged for BR ($r=0.23$, $p<0.05$) and NN ($r=0.21$, $p<0.05$).

The relative typicality of the timing profiles (TUPC1) had some modest positive relationships with BR's and OS's ratings, indicating a slight preference for more typical or conventional timing. However, there was a much stronger correlation of this variable with BRSYN. This is consistent with the results of an earlier study using a smaller set of synthesized performances of the same excerpt and a group of pianist judges (Repp, 1997). Evidently, unconventional timing is more difficult to appreciate in synthesized than in original performances. There was also a positive correlation with BRDIFF.

Remarkably, there were no significant relationships between the judges' ratings of the original excerpts and any of the timing strategies or any of the dynamic variables. BRSYN and BRDIFF, however, showed a positive correlation with TPC1 loadings. Since this was the most common timing pattern and similar to TUPC1, this basically reflects the same preference for relatively typical timing patterns in synthesized performances. BRDIFF also correlated negatively with TPC2, indicating that this rather unusual timing pattern (called PC-II or Type II in Repp, 1998) was difficult to appreciate in the synthesized versions. BRDIFF also showed a small positive correlation with a dynamic variable, DPC1 (called PC-I or Type I in Repp, 1999), indicating that removal of this dynamic pattern tended to increase ratings.[10]

To determine the total amount of rating variance accounted for by the performance variables, stepwise multiple regression analyses with an inclusion criterion of $p<0.05$ were conducted on each set of ratings. Since this kind of analysis takes into account any intercorrelations among the performance variables, there was a possibility that additional significant relationships would emerge in the form of partial correlations. (Most of the intercorrelations were quite low, however, and principal component patterns are by definition orthogonal.) The results are shown in Table III.

The most noteworthy outcome is that performance variables accounted for 53% of the variance (R^2) in the BRSYN ratings, but only between 9% and 18% in the judge's ratings of the original excerpts, even though there were twice as many relevant performance variables. The estimated systematic variance in these average ratings ranged from 45% to 80%. The large amount of unexplained variance must be due to performance aspects that were not captured by the objective measurements. Even in the BRSYN ratings, a substantial portion of the systematic variance (estimated at about 80%) remains unexplained, which suggests that there are aesthetically relevant aspects of expressive timing that were not captured by the eight timing variables.

TABLE III. Results of stepwise multiple regression analyses on the ratings.

Ratings	Performance variable	(partial) r	R
BR	TEMPO	−0.36	
	DPC4	−0.20	0.39
MC	TEMPO	−0.42	0.42
NN	TUPC1	0.31	0.31
OS	UPBEAT	0.30	
	RELMOD	−0.25	0.38
BRSYN	TEMPO	−0.46	
	TUPC1	0.41	
	RELMOD	−0.21	0.73
BRDIFF	RELMOD	−0.30	
	TUPC1	0.27	
	TPC2	−0.26	
	TEMPO	−0.17	0.54

The variables that made a significant contribution to the regression equations were largely the ones that showed significant simple correlations with the ratings (Table II), with the following exceptions: For BR, TUPC1 dropped out and instead DPC4 made a significant (negative) contribution, which is a rare instance of a dynamic variable making an appearance. For NN, TEMPO dropped out. For BRSYN, UPBEAT and TPC1 dropped out, and for BRDIFF, TPC1 dropped out.

D. Correlations with sociocultural variables

The sociocultural variables were the year of recording, the artist's birth date, age at the time of recording, gender, and nationality (country of birth). Only OS's ratings showed a positive relationship with the year of recording ($r=0.38$, $p<0.001$). When a quadratic trend was included, the correlation increased to 0.45 and revealed that the increase in ratings occurred especially since about 1970. This is interesting because OS (the youngest of the judges, born in the late 1960s) grew up with these recordings. The relevant factor may have been either recorded sound quality (although OS, like the other judges, had been instructed to disregard this aspect as much as possible) or some other as yet unidentified quality associated with more recent recordings. (Although more recent recordings tend to be more typical, OS's ratings showed no significant correlation with TUPC1 or DUPC1, the typicality indices.) OS's ratings also showed a positive relationship with artists' birth dates ($r=0.23$, $p<0.05$), which of course were strongly correlated ($r=0.79$) with year of recording. There were no significant linear or curvilinear relationships of any judge's ratings with the artists' age at the time of recording, so that there is no evidence from these data that more mature artists produce better performances (of

a single phrase). Furthermore, there were no significant effects of artists' gender or nationality, although the latter analysis included only French, Polish, and Russian pianists, of whom there were sufficient numbers in the sample.

The BRSYN ratings showed small positive correlations with year of recording ($r=0.25$, $p<0.01$) and with artists' birth dates ($r=0.25$), while the BRDIFF ratings correlated with recording date only, again positively ($r=0.35$, $p<0.001$). Thus it was *not* the case that older, scratchy recordings benefitted from being partially reincarnated as synthesized performances; on the contrary, they suffered more than did more recent recordings. The presumable reason for this was that they represented more individual timing profiles that, as we have seen, were more difficult to appreciate in the context of the digital piano sound. These results also suggest that BR successfully disregarded recorded sound quality when making his evaluations of the originals.

III. GENERAL DISCUSSION

Timing and dynamics are often considered the two most important parameters of expression in piano performance. Psychological research on music performance has focused predominantly on timing and to a much lesser extent on dynamics, but the importance of dynamics has surely been recognized. A third important parameter, even less investigated but likewise appreciated, is articulation—the degree of overlap and separation of successive tones. In the present Chopin Etude excerpt, however, variation in articulation hardly played a role because *legato* was employed throughout and the damper pedal was used almost continuously. Thus timing and dynamics were the primary expressive dimensions, and it was startling to find that they accounted for only a small amount of variance in the judges' ratings.

This result was not due to low reliability of the ratings: The systematic variance in judges' average ratings (the reliability coefficient squared) ranged from about 45% to 80%, whereas the variance accounted for in the regression analyses ranged from 9% to 18%. It was also not due to significant nonlinear relationships between the ratings and the independent variables, although some small quadratic trends were noted. Thus the variance accounted for was perhaps underestimated by a few percent, but the basic result remains. Furthermore, it cannot be attributed to insensitivity or idiosyncrasy of the judges who were all experienced musicians and/or listeners. The fact that they employed somewhat different criteria in making their evaluations only increases the generality of the results.

The judges' exact criteria were not known, of course, and it would have been too cumbersome to attempt to determine them precisely, given the large number of judgments to be made. Also, appropriate methods for eliciting this information are only now beginning to be developed (see Thompson *et al.*, 1998). Even experienced judges of musical performances often do not know how they arrive at their overall assessments; this is in large part an intuitive, not a cognitive process, especially at the upper end of the scale. For example, the author (BR) feels that he rated the (to him) most excellent performances (ratings of 8 and above) on the basis of the degree to which they "engaged his body" and made him "move inwardly" with the music, which reveals little about the specific performance properties that elicited this kind of biophysical (perhaps merely metaphorical) resonance. In performances that he gave ratings of 5 or below, however, BR usually detected specific flaws that he mentioned in his comments.

One conclusion to be drawn from the results is that the different patterns of expressive timing and dynamics uncovered by the principal components analyses of the 115 performances are all aesthetically viable, as are different tempi and different degrees of expressive modulation, within broad limits. This makes sense because nearly all of the recorded pianists are distinguished concert artists who play with great expertise and conviction. Certainly there is no single optimal way of playing the Chopin phrase, even within the individual minds of the four judges: Each judge gave equally high ratings to performances of very different character. A second conclusion is that, in this particular musical excerpt, the performance aspects of primary aesthetic importance are not the ones measured in the objective performance analyses (Repp, 1998, 1999).

What then are these important but as yet unquantified aspects? Here only speculations can be offered, based in part on the judges' comments and mainly on the author's intuitions. First, there may be additional aspects of timing and dynamics that were not captured by the performance analyses. Second, there may be additional expressive dimensions.

One clue that there may be aesthetically relevant aspects of timing that were not captured by the performance analyses is offered by the finding that the timing variables accounted for only 53% of the variance of BR's ratings of the synthesized performances, whereas the systematic variance was in the vicinity of 80%. Since these performances differed in timing only, some aspect of timing must have been responsible for the unexplained 27% of the variance. One likely candidate is smoothness or continuity. Comments such as "jerky," "mannered," and "distorted" kept recurring in BR's notes, usually with reference to the same performances. This aspect is very difficult to quantify. Although Todd (1995) has proposed a principle by which the smoothness of timing might be gauged, it is difficult to apply to short melodic-rhythmic segments within a phrase. Performances considered "jerky" by BR generally had high degrees of relative modulation depth, so that the impression may have derived in part from exaggerated *rubato*. Performances with high relative modulation depth that were rated favorably by BR often had merely a very long initial downbeat. If the initial IOI had been omitted, perhaps the RELMOD variable would have explained more variance in BR's ratings than it did. However, this can hardly be the whole story.

BR also noted a few performances in which the basic tempo seemed to change in the course of the phrase, to their detriment. This seemed to rest mainly on a change in the speed of the accompaniment passages between melodic seg-

ments, which could be quantified more easily but probably would account for only little variance. The observation raises some interesting questions, however, about the perception of basic tempo constancy in highly modulated contexts (cf. Repp, 1994).

There are few clues to aspects of (horizontal) dynamics that may have been missed. However, occasional comments from judges concerned exaggerated or misplaced accents. In one performance, for example, a single bass note was far too loud, apparently a slip of some sort. This probably diminished the aesthetic impact of the performance, but the overall effect of such local flaws on the judges' ratings cannot have been very large.

It is possible that interactions between timing and dynamics had some aesthetic importance. Although Repp (1999) found no significant correlations between timing and dynamics in the original performances, this does not rule out the possibility that certain combinations of timing and dynamic profiles are aesthetically more appealing than others. If so, that may also explain why BR's judgments of the synthesized performances, which had a fixed and fairly typical dynamic profile, were biased in favor of more typical timing profiles (see also Repp, 1997). Research is planned to specifically investigate the aesthetic interdependence of timing and dynamics; at present, there is no direct evidence for such an interaction. Interactions may also occur very locally. For example, one of BR's most favored performances had a hesitation on the first note of the final melodic gesture that he perceived as especially poignant and moving. However, the same hesitation did not sound poignant at all in the synthesized version of this timing profile, which received a much lower rating.

Among the performance dimensions that were not properly measured and quantified, due to methodological limitations, are vertical timing and dynamics. Vertical timing refers to asynchronies among tone onsets that are nominally simultaneous in the score. Large asynchronies, usually due to the bass leading the melody, occurred in a number of performances, mostly of older vintage. While these asynchronies were occasionally annoying, they also could make a positive aesthetic contribution. Therefore, this was probably not a major factor. However, the role of smaller asynchronies that are imperceptible as such remains uncertain. They may have contributed importantly to the highlighting of the melody (see Palmer, 1996) and to the sound texture.

Vertical dynamics refers to the relative intensities of simultaneous tones, also referred to as texture. This is likely to be an important factor in the present excerpt, where the melody needs to stand out from the accompaniment (cf. the epigraph at the beginning of this article) and the accompaniment itself represents different voices with different functions. However, the dynamic contrast between melody and accompaniment was partially captured by the MELACC variable in the regression analysis, which surprisingly explained no variance at all. Since this measure was based only on the accompaniment during sustained melody tones, however, it was perhaps not representative of the texture at melody tone onsets, and certainly not of the texture within the multi-voiced accompaniment.

Another factor not considered because it is impossible to measure in acoustic recordings is pedaling. Only MC occasionally complained of "bad pedaling." However, there may have been effects of pedaling on texture and sound quality that were not directly perceived as being due to pedaling. Two other factors that may have played a role are instrument quality and room acoustics. While the judges seemed to have been successful in disregarding differences in recorded sound quality (such as crackles, noise, or frequency band limitations), the quality of the piano and of the room in which it was recorded were difficult to separate from the performance itself. Needless to say, they are also almost impossible to measure in an objective way.

The author suspects that the aesthetically most important aspect of the performances was the elusive quality often referred to as "tone" or "touch" (see again the epigraph). This remains the most poorly understood aspect of pianistic skill, but quite possibly the most crucial one. It lends the melody a singing quality that listeners resonate to because it invokes processes in their own body, related to voice production and movement. It is probably a combination of several or all the factors mentioned above, especially vertical dynamics, small asynchronies, smoothness of timing, pedaling, and instrument quality. Perhaps the elusiveness of this quality rests on the fact that it involves nearly all aspects of pianistic skill, in which case it really may not have much explanatory value. Tone and touch may simply be synonymous with complete mastery of the instrument.

Beyond such mastery, the quality of the instrument itself is important. This is demonstrated by BR's ratings of the synthesized performances which confirmed Repp's (1997) observation that digital piano renditions elicit a preference for typical timing patterns. Although there are several possible explanations for this finding (discussed by Repp, 1997, and above), the most plausible one is considered to be the quality of the digital piano sound in combination with the fact that the original performance onto which the timing profiles had been grafted was not of the highest standard. These factors prevented the melody from soaring (i.e., from having the kind of aesthetic impact described metaphorically in the epigraph), and this in turn reduced the listener's tolerance for atypical and original timing patterns. True originality and individuality in performance timing, which often may reflect spontaneity of expression, seem to require a beautiful tone or touch—in other words, supreme skill *and* a fine piano.[11]

The present results then may serve as a reminder that there is still a large gap between objective performance analysis and our understanding of the aesthetic impact of a performance. It is hoped, however, that further research will slowly close this gap.

ACKNOWLEDGMENTS

This research was supported by NIH Grant No. MH-51230. Thanks are due to Manfred Clynes, Nigel Nettheim, and Omer Shomrony for lending their expertise to this project. Address correspondence to Bruno H. Repp, Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695; e-mail: repp@haskins.yale.edu

APPENDIX: THE JUDGES' RATINGS OF THE RECORDINGS

BR's ratings represent the average of 10 sessions, those of MC, NN, and OS the average of two sessions each. Some performances were rated by BR only. The entries under BRSYN represent BR's ratings of the synthesized excerpts, averaged over five sessions.

Pianist	BR	MC	NN	OS	BRSYN	Pianist	BR	MC	NN	OS	BRSYN
Aide	6.5	5.5	6.8	7.8	7.6	Koczalski	7.2	6.3	4.8	6.5	3.6
(duplicate)	6.2	6.8	6.8	7.0	7.8	Koyama	7.0	6.0	5.8	6.3	7.2
Anda	7.2	7.9	4.3	3.3	2.8	Kyriakou	6.1	7.0	4.5	4.5	4.8
Anievas	8.0	7.0	4.0	7.0	6.4	Larrocha	5.3	1.5	3.5	5.5	4.6
Arrau—1930	9.2	8.1	6.0	6.3	6.4	Levant	6.6	6.3	4.3	5.0	4.4
Arrau—1956	6.7	6.0	8.0	7.0	7.0	Liberace	4.9	2.8	3.0	4.3	5.4
Ashkenazy—1959	6.0	6.5	5.8	5.8	5.6	Licad	5.8	6.5	4.5	7.3	6.2
Ashkenazy—1974	7.4				7.0	Lopes	8.5	4.3	7.3	5.3	9.0
Backhaus	8.2				7.0	Lortat	6.8	5.0	6.8	6.5	5.2
Badura-Skoda	5.2	3.5	5.8	4.0	5.0	Lortie	6.5				7.6
Berezovsky	8.4	4.8	6.0	8.8	6.8	Magaloff	6.3	3.8	5.3	8.3	6.0
Bingham	6.1	7.8	5.3	9.5	6.8	Magin	9.4	6.8	6.8	4.5	10.0
Binns	8.1	7.5	6.8	8.8	6.6	Malcuzinsky	7.9	4.0	4.5	7.0	6.4
Biret	6.5				4.6	Mamikonian	5.3	4.0	5.3	4.5	5.6
Brailowsky	6.9	4.0	6.8	5.5	8.2	Manz	7.0	3.0	5.3	7.5	6.4
Browning	8.2				5.8	Murdoch	4.8	4.3	5.5	5.0	5.6
Cherkassky	9.6	8.1	7.5	5.0	7.2	Niedzielski	7.9	8.5	5.5	1.5	4.6
Ciani	9.2	9.5	6.5	5.0	6.8	Novaes	6.4	3.0	4.5	6.0	5.0
Ciccolini	6.9	2.0	3.0	4.0	8.4	Paderewski	1.8	2.0	2.8	4.0	2.2
Cliburn	5.4				8.0	Pennario	3.6	1.0	3.0	4.3	3.2
Coop	6.6	2.0	6.3	9.0	7.4	Penneys	7.1	5.5	5.5	8.5	6.4
Cortot—1933	8.3				4.0	Perahia	9.0	8.0	7.5	8.3	6.0
Cortot—1942	7.1	4.0	7.3	5.5	1.4	Perlemuter	6.1				4.0
Costa	6.5	2.8	5.3	4.0	6.6	Pokorna	3.7	2.0	3.3	3.0	5.0
Crown	6.1	2.8	4.8	3.5	5.8	Pollini	7.5				3.8
Cziffra—1954	2.4	2.0	6.8	4.8	2.0	Ranki	6.6	2.3	6.0	6.5	6.8
Cziffra—1981	2.0	1.0	3.3	5.0	2.8	Renard	4.2	5.8	4.5	3.5	2.4
Darré	5.7	4.5	8.3	3.5	4.6	Richter	8.0				7.4
Donohoe	7.8	3.3	6.3	6.8	8.2	Saperton	8.2	8.8	4.0	5.0	4.2
Drzewiecki	7.0	5.3	6.3	7.5	8.0	Sasaki	6.4	4.5	6.0	7.0	5.2
Duchâble	6.1	8.3	6.0	8.5	6.8	Sauer	7.0	8.0	5.0	5.0	3.4
Egorov—1978	7.6				4.8	Schein	7.5	8.0	6.3	6.5	6.8
Egorov—1979	6.4	5.8	6.3	6.8	5.6	Shebonova	7.7	4.5	6.0	8.0	7.4
Ellegaard	7.8	6.4	7.3	6.8	4.8	Simon	7.8	2.8	5.5	8.3	6.4
Entremont	8.6				6.2	Skavronsky	6.1	4.5	6.5	8.5	6.4
Farrell	6.8	7.9	6.5	5.5	6.6	Slenczynska—1956	6.7	2.0	6.5	4.0	6.2
Fou Ts'ong	2.1	1.0	2.5	7.0	1.2	Slenczynska—1975	5.3	1.5	7.5	6.8	6.2
François	4.5	1.0	3.0	5.0	1.6	Slobodyanik	3.7	2.0	4.0	2.0	7.6
Goldenweiser	4.8	1.8	6.0	3.5	7.2	Smith	6.9	6.3	7.0	4.3	3.2
Goldsand	6.7	6.5	6.0	6.0	8.0	Sofronitzky	6.4	2.0	3.8	5.8	3.8
Goodman	6.0	3.0	4.8	7.8	6.4	Solomon	6.7	5.6	7.0	4.5	9.0
Haas, M.	5.9	2.5	5.0	8.0	4.8	Székely	4.3	4.8	6.0	8.5	6.2
Haas, W.	5.3	3.5	5.0	6.5	5.2	Timofeyeva	4.9	3.5	4.8	4.0	4.8
Haase	1.6	1.0	2.5	5.0	2.0	Uninsky	4.1	2.0	7.3	5.8	3.2
Harasiewicz	6.4				5.4	Varsi	1.2	1.0	1.0	3.0	1.2
Hesse-Bukowska	8.3	7.3	7.0	6.8	7.0	Vásáry	6.6				5.4
Hobson	7.9	3.5	7.0	5.0	7.4	Vered	7.5	8.0	7.5	5.5	8.8
Horowitz—1951	7.9				5.2	Virsaladze	7.2	6.8	6.5	6.8	9.2
Horowitz-1972	3.9	6.8	5.3	4.8	2.2	Volondat	4.6	7.5	5.0	8.3	9.0
Iturbi	6.0	4.5	4.0	7.3	6.8	Weissenberg	7.0	4.5	5.0	6.0	8.2
Janis	5.2	2.3	3.5	4.0	2.8	Wild	6.5	7.0	7.3	6.8	6.8
Johannesen	8.0	2.5	6.3	5.0	7.4	Woodward	5.4	7.5	6.3	6.3	6.8
Joyce	6.3	2.5	5.8	4.5	1.8	Woytowicz	8.6	5.5	6.0	5.3	8.8
Kahn	8.8	2.8	6.0	6.5	6.0	Yamazaki	2.7	1.5	4.3	5.3	5.0
Karolyi	7.1	4.5	4.5	7.8	6.4	Yokoyama	7.3	4.3	6.8	9.0	7.8
Katz	8.7	9.0	3.3	6.0	7.0	Zarankin	3.6	6.3	5.0	6.0	6.2
Kentner	8.2	6.1	4.8	7.3	5.0	(duplicate)	4.9	6.5	4.5	4.5	5.8
Kersenbaum	8.0	7.8	6.5	7.0	7.6	Zayas	4.9	1.0	6.0	5.0	4.0
Kilényi	6.8	7.0	6.8	8.0	7.2						

[1]The metronome indication in Fig. 1 is much too fast and must be a misprint. The music was copied from this old German edition because it is the only one that fits the first five bars into a single system.

[2]It should be emphasized that this phrase was evaluated here as an isolated excerpt, not as part of the whole composition. It is conceivable that a beautiful first phrase somehow does not fit with what follows or that a plain first phrase gains in meaning as the music continues. However, any such contextual effects are irrelevant to the present study, as long as generalization of the results to performances of the whole Etude is eschewed.

[3]This was a copy of the digital master tape the author had obtained from the International Piano Archives at the University of Maryland. Two of the performances turned out to be duplicates, but they were included in all analyses.

[4]These excerpts had been retained only as digitized sound files, some of which had a lower sampling rate and thus were not of optimal sound quality. BR felt that this did not influence his evaluations.

[5]In sessions 1–5, BR listened to digitized sound files stored in the computer, which were of slightly lower fidelity than those on the digital master tape. In sessions 6–10, he listened to a digital copy of the master tape (with randomly changed track numbers, so as to disrupt any memory for earlier judgments associated with particular serial numbers), except for the 15 additional excerpts, to which he still listened as computer files, going back and forth between tape recorder and computer, as necessary.

[6]Small differences from the timing of the originals may have arisen from two sources. First, the IOI measurements of the original excerpts (Repp, 1998) contained some human error, although it was believed to be well below the perceptual discrimination threshold. If there were any larger errors, they would have occurred in the timing of the accompaniment notes (which are more difficult to measure because of their low intensity) and would have been perceptible as a momentary irregularity, but BR heard very few such instances in the synthesized versions and believes they had little impact on his aesthetic judgments. Second, due to a peculiarity of the software used for playing the synthesized versions (MAX), the tempo of the synthesized excerpts was 2.4% faster than that of the originals. However, this affected all excerpts equally and thus should have had little effect on their relative aesthetic impressions. Another possible concern was that the synthesized excerpts ended with a prolonged chord, so that the final IOI of an imported original timing profile was not contextually appropriate (a longer IOI would be expected before a final chord). However, this, too, is unlikely to have had any effect on the overall aesthetic impression, which mainly derived from the timing of the passage as a whole.

[7]It would have been nice to construct an analogous set of synthesized performances varying in expressive dynamics only. However, this turned out to be problematic, due to the fact that only horizontal dynamics (the succession of overall peak sound levels) had been measured in the original excerpts. To construct synthesized performances from these measurements, the relative intensities of the individual notes in simultaneities (vertical dynamics) would have had to be estimated somehow, and it seemed unlikely that this could be done very accurately. Therefore, this possibility was not pursued further.

[8]The fastest performance in the sample, by Dinorah Varsi, was a true outlier in the tempo distribution (see Repp, 1998: Fig. 2) and received very low ratings from all four judges. When this performance was omitted, the correlations were slightly smaller, and that for NN fell short of significance.

[9]No separate tests were conducted to determine the significance of the improvement.

[10]In the terminology of Repp (1999), the constant dynamic profile of the synthesized performances was of Type II–III. (Its dynamic PC-II and PC-III loadings were 0.50 and 0.49, respectively, and its correlation with the average of PC-II and PC-III was 0.70.) Thus original performances with a Type I dynamic profile indeed "lost" this profile in their synthesized incarnations.

[11]An interesting prediction following from these considerations is that artists who play with very individual expression on a concert grand may adopt a more conventional style on a poor instrument, such as a digital piano. The quality of the instrument may affect artistic originality and spontaneity.

Berry, W. (1989). *Musical Structure and Performance* (Yale University Press, New Haven, CT).

Clarke, E. F. (1985). "Structure and expression in rhythmic performance," in *Musical Structure and Cognition*, edited by P. Howell, I. Cross, and R. West (Academic, London), pp. 209–236.

Cook, N. (1990). *Music, Imagination, and Culture* (Clarendon, Oxford, U.K.).

Levinson, J. (1990). "Evaluating musical performance," in *Music, Art, and Metaphysics* (Cornell University Press, Ithaca, NY), pp. 376–392.

Levinson, J. (1997). *Music in the Moment* (Cornell University Press, Ithaca, NY).

Palmer, C. (1989). "Mapping musical thought to musical performance," J. Exp. Psychol. HPP **15**, 331–346.

Palmer, C. (1996). "On the assignment of structure in music performance," Music Percept. **14**, 23–56.

Repp, B. H. (1994). "On determining the basic tempo of an expressive music performance," Psychol. Music **22**, 157–167.

Repp, B. H. (1997). "The aesthetic quality of a quantitatively average music performance: Two preliminary experiments," Music Percept. **14**, 419–444.

Repp, B. H. (1998). "A microcosm of musical expression: I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major," J. Acoust. Soc. Am. **104**, 1085–1100.

Repp, B. H. (1999). "A microcosm of musical expression: II. Quantitative analysis of pianists' dynamics in the initial measures of Chopin's Etude in E major," J. Acoust. Soc. Am. **105**, 1972–1988.

Scruton, R. (1997). *The Aesthetics of Music* (Oxford University Press, Oxford, U.K.).

Shaffer, L. H. (1995). "Musical performance as interpretation," Psychol. Music **23**, 17–38.

Sherman, R. (1997). *Piano Pieces* (North Point Press, New York).

Sloboda, J. (1998). "Does music mean anything?" Musicae Scientiae **2**, 21–31.

Thompson, W. F., Diamond, C. T. P., and Balkwill, L.-L. (1998). "The adjudication of six performances of a Chopin Etude: A study of expert knowledge," Psychol. Music **26**, 154–174.

Todd, N. P. McA. (1995). "The kinematics of musical expression," J. Acoust. Soc. Am. **97**, 1940–1949.

Wapnick, J., Flowers, P., Alegant, M., and Jasinskas, L. (1993). "Consistency in piano performance evaluations," J. Res. Music Educ. **41**, 282–292.