On the contextual appropriateness of expression

Renee Timmers

Music, Mind, Machine group, Nijmegen Institute for Cognition and Information,

University of Nijmegen, The Netherlands

Correspondence to: NICI, University of Nijmegen, P.O. Box 9104, NL-6500 HE Nijmegen, E-mail: renee74@xs4all.nl Tel: +31(0)24-3612650, Fax: +31(0)24-3616066.

Abstract

The aim of this study was to show that the quality of an expressive interpretation depends on expressive context. The main hypothesis was that expression is evaluated in relation to preceding expressive variations. Two experiments and a model tested this hypothesis. In the first experiment, 39 listeners rated the quality of the performance of the continuation (second half of the musical stimulus) given the performance of the initiation (first half of the musical stimulus). The results showed a significant effect of continuation on the quality judgements and a significant interaction between continuation and initiation. This interaction was seen as the first confirmation of the hypothesis. In the second experiment, 20 participants rated the quality of the six performances of the initiation and of the continuation separately. The results of this experiment were unable to explain the quality judgements of experiment 1. The low agreement between the judgements was taken as a second confirmation that contextual considerations can overrule general aesthetic preference. A regression model was proposed that predicts the quality rating of experiment 1 from the similarity in rubato extent, key velocity pattern, average articulation, grace note duration and average asynchrony between the two segments. This model was better able to explain the quality judgements of the continuation, providing final confirmation that the quality of the second half was a function of its agreement with the first half.

Introduction

This article aims to take a first step in demonstrating and formalizing the intrinsic constraints on an expressive performance of music. Intrinsic constraints are those that are set by the performer interpreting the music in a certain way and performing it with a certain expression and style. For example, choices at the beginning of a phrase provide constraints and expectations for the rest of the phrase: if the dynamics of a performance increase to the middle of the phrase, this may require a balancing diminuendo in the second part of the phrase. Or if ornaments are chosen to be performed long, this may set the trend for future ornaments.

Previous literature has already reported considerable success in defining performance rules that in general terms constrain the expressive performance of music. For example, expressive variations are shown to relate to the interpretation of musical structure (see e.g. Clarke, 1985; Palmer, 1989; Sundberg, Friberg, & Frydén, 1991a) or are shown to communicate an emotional interpretation (see e.g. Gabrielsson & Juslin, 1996; Juslin, 1997). However, no attempt has been made to formalize and to empirically demonstrate the serial constraints on a performance; having chosen an interpretation and having communicated the interpretation with a certain style of expression, what does this imply for the continuation of the performance?

Overall constraints and intrinsic constraints could be seen as two sides of the same coin or as two separate phenomena operating under different conditions. If the same performance rules apply to the whole performance, it is natural for intrinsic constraints to exist and intrinsic rules would be a consequence of overall rules. If however intrinsic constraints have an independent existence, it is possible for these constraints to overrule the overall constraints. For example, a performance could not be in accordance with general performance rules, but nevertheless be the best option for a given expressive context. Or otherwise, a performance could be well done on itself, but inappropriate for its context. In this way, the constraints on a performance and the quality of it become context dependent. This means that performance rules and the quality of a performance are not always absolute as suggested by e.g. Clynes (1983), Repp (1997) and Sundberg, Friberg, and Frydén (1991b). It also means that the performer may play a considerable role in setting the constraints of the performance.

The hypothesis of intrinsic constraints on a performance is related to Meyer's concepts of tendency and expectation evoked by the structure of the music, which he explored in his influential book on emotion and meaning in music (Meyer, 1956). Performances are one of the aspects that, in the terms of Meyer (1956), evoke a tendency in the listener and give rise to a more or less specific expectation of a consequence. If this expectation is violated, the deviation should sooner or later be resolved. Context dependent norms were further mentioned by Clarke (1995), who suggested the possibility that a way of performing certain figures, such as the long/short interpretation of equal quarter notes, can become the norm from which later performances might deviate. Timmers and Desain (2000) have found musicians referring to the process of setting the norm and deviating from it within the performance of a single piece. Repp (1998) has suggested the existence of expectations on performance variations based on previous variations, but has rejected this hypothesis on the basis of the findings in his own study. He found that the expected timing deviations related to grouping structure were the same irrespective of context.

This study first focuses on demonstrating that the quality of a performance can be context dependent. In other words, it aims to demonstrate that it is possible for intrinsic considerations to overrule overall constraints. Secondly, it focuses on the formalization of the intrinsic constraints. It will stress variations that set the norm that are different from Repp's obligatory expectations, which relate to the processing of musical structure. The dynamically established constraints that are explored in this study concern typical variations of a performance, such as the amount of rubato and the kind of articulation, and consistency constraints that play a role in the well-formedness of an initiated gesture, such as the appropriate closure of an increase in dynamics and the appropriate second occurrence of a grace note.

Method

Two experiments were carried out to show that the quality of performance variations is context dependent. Both experiments were perceptual experiments in which participants were asked to give an aesthetic judgement on the way the music is performed. The musical material was a fragment of the theme of Beethoven's Paisiello Variations for piano solo (G major WoO 70, 1795). In the first experiment, participants rated the quality of the continuation (the second half of the performance) in the context of a certain initiation (the first half of the performance). In the second experiment, participants rated the quality of the initiations and continuations separately. A comparison between the results of the two experiments is intended to show whether the evaluation of expressive variations is context dependent or independent.

Participants

Experiment 1. 39 participants participated in the experiment, 25 of whom were professional musicians. They all played a classical instrument to an advanced level. They were selected on the sole ground of being good performers, and were assumed to have good music listening skills and a clear feeling for rules governing a performance of a classical piece. They included pianists (N = 16) and non-pianists (N = 23).

Experiment 2. A subset of 20 participants from experiment 1 participated in experiment 2, including 13 professional musicians. There were seven pianists and 13 non-pianists.

Stimuli

Six performances of two segments of the theme of Beethoven's Paisiello Variations (see Figure 1) were selected from a database of performances recorded on a Yamaha MIDI Grand in a previous study[1] (Timmers, Ashley, Desain, Honing, & Windsor, 2002).

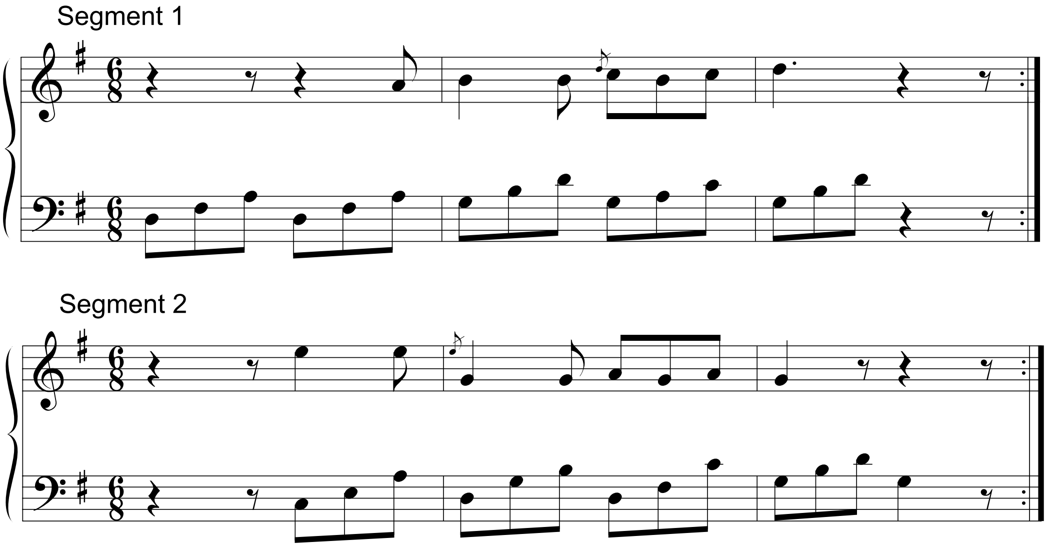


Figure 1: Score of segments 1 and 2. Six performances of these two segments were used as musical material in the experiments.

The six performances of the segments were selected on the basis that they had certain features in common as well as certain salient differences. Table 1 shows an overview of the main characteristics of the selected performances. This characterization plays an important role in the model that will later be proposed and the reader is referred to the section on the model for an explanation of the characterization of the stimuli.

Table 1. Taxonomy of the 12 performances; amount of rubato of the accompaniment inter-onset-intervals (IOI), velocity pattern, articulation of the accompaniment notes, grace note duration and asynchrony between the melody and accompaniment notes (a positive value means that the melody leads).

Pianist	Rubato (std dev)	Velocity (Pattern)	Articulation (Dur/IOI)	Grace IOI (ms)	Asynchrony (ms)
1					
initiation	15.6	Flat	0.978	47	5
continuation	17.3	Flat	1.118	96	3
2					
initiation	18.8	Rise	1.072	116	6
continuation	28.2	Fall	1.076	170	-6
3					
initiation	12.1	Flat	1.163	60	11
continuation	35.5	Flat/fall	1.104	157	-8
4					
initiation	14.2	Rise	0.968	36	16
continuation	35.2	Fall	0.980	114	14
5					
initiation	16.7	Flat	0.916	134	-16
continuation	15.1	Flat	0.959	119	-16
6					
initiation	38.9	Rise	1.268	49	24
continuation	77.3	Fall/flat	1.293	74	27

Stimuli experiment 1. For experiment 1, pairs of performances were combined to form 36 new interpretations of one phrase (see Figure 2). Corrections were made for differences in global tempo and global key velocity between the performances of the two halves. This was done in such a way that the transition between the performances was entirely smooth in tempo and dynamics (see Timmers (2002) for a detailed description).

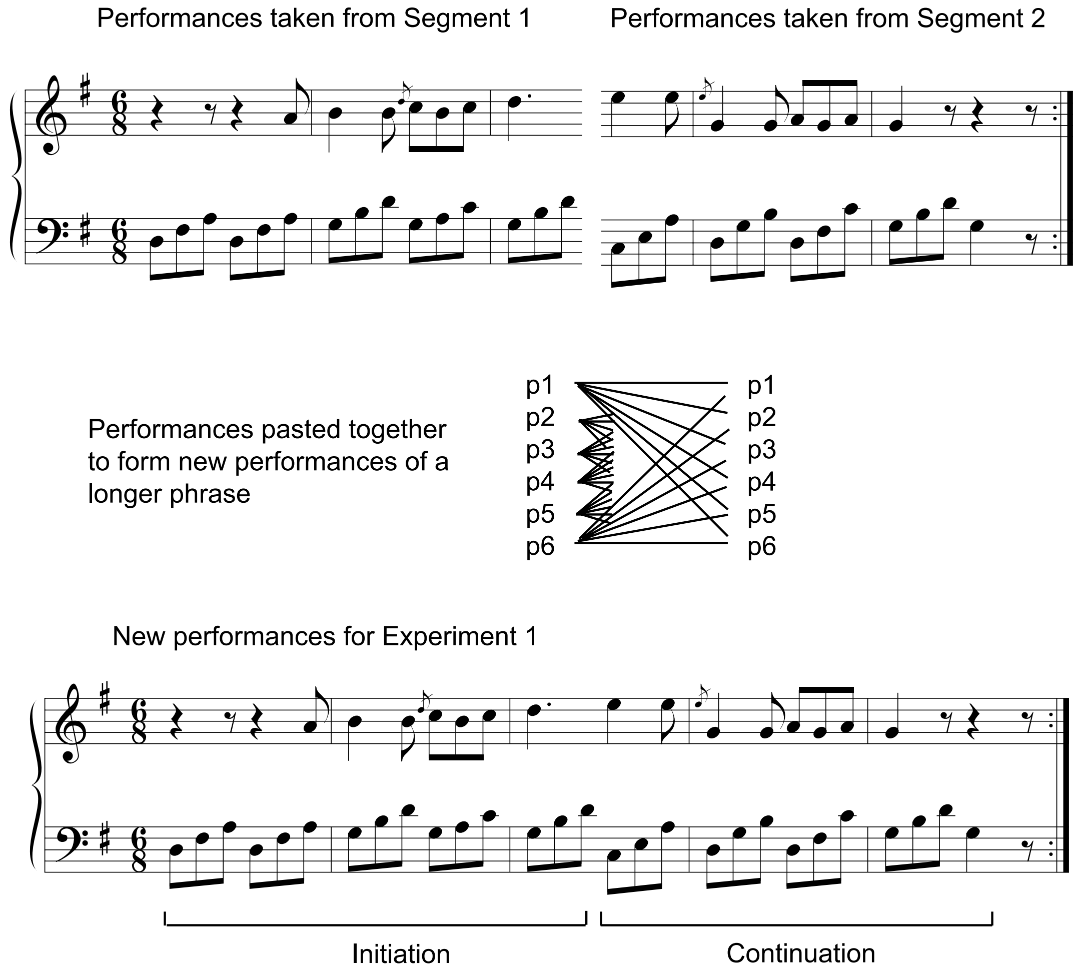

Performances taken from Segment 1
Performances taken from Segment 2
Performances pasted together to form new performances of a longer phrase
p1
p2
p3
p4
p5
p6
p1
p2
p3
p4
p5
p6
New performances for Experiment 1
Initiation
Continuation

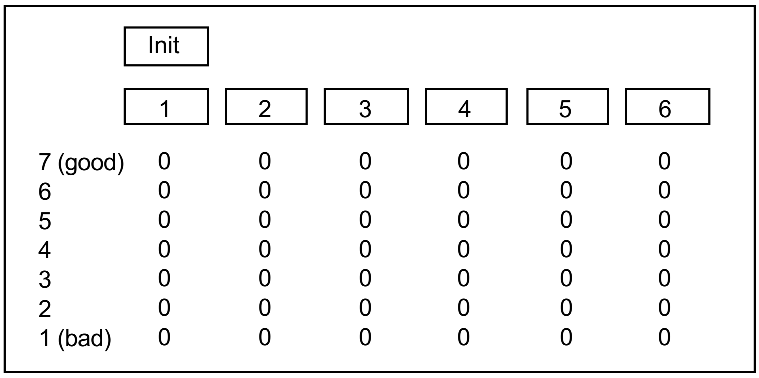
Figure 2: Combination of performances: a performance of segment 1 is concatenated to a performance of segment 2. The combinations (p1 with p1, p1 with p2, etc.) make up new performances of a longer fragment; the phrase used in experiment 1.

Stimuli experiment 2. The stimuli of experiment 2 were the six performances of the first half and the six performances of the second half presented separately in blocks that contained first all initiations and then all continuations.

Procedure

Experiment 1. Each subject was seated behind a portable Macintosh computer and read the instruction from a text file. S/he heard the sound over headphones. The instruction indicated that the participant would hear performances of a fragment of the theme from Beethoven's Paisiello Variations and that s/he was asked to give an aesthetic judgement of the continuation with respect to the initiation. For a block of six performances, the first half of the phrase would each time be performed in the same way, while the second half of the phrase would each time be performed differently. The first half was presented separately in order to familiarize the participant with the standard.

Participants alternately listened to the first half only and to a complete performance, and rated the performances on a scale from one to seven (where one meant a bad continuation given the initiation, or a badly fitting second half; while seven meant a good continuation given the initiation or a successfully fitting second half). The rating was done by clicking on the appropriate radio button (see Figure 3).



Init

	1	2	3	4	5	6
7 (good)	0	0	0	0	0	0
6	0	0	0	0	0	0
5	0	0	0	0	0	0
4	0	0	0	0	0	0
3	0	0	0	0	0	0
2	0	0	0	0	0	0
1 (bad)	0	0	0	0	0	0

Figure 3: User interface for experiment 1. At the top, the play buttons for the performances (one start only and six entire phrases). Below, the radio buttons for giving a rating of quality for each continuation.

In total, the experiment consisted of six blocks of six performances, with each block based on a different initiation. The order of the blocks was randomised. The order of the continuations within blocks was also randomised. The user-interface and playback system were made in POCO (Honing, 1990).

Experiment 2. Each participant was seated behind a Macintosh PowerBook and read the instructions from the screen. These explained that the performances of the previous experiment were made by combining six performances of the first half with six performances of the second half. They would now hear the performances of the first and second halves separately. They were asked to rate the aesthetic quality of the performances on a scale from one to seven. One would mean a bad performance, while seven would mean a good performance. The interface was the same as for experiment 1, though it only contained the six buttons for the six initiations or six continuations. The participants first rated all initiations and then all continuations. The order of the performances within blocks was randomised.

Results

Experiment 1

The agreement between the participants was moderate to low with some positive and some negative exceptions for both experiments. The average of the pair-wise correlations between the ratings of the participants was 0.32 for experiment 1 and 0.28 for experiment 2.

For experiment 1, a repeated measures ANOVA tested the main effects of initiation, continuation and the interaction between initiation and continuation on the quality rating of the continuation. There was a significant effect of continuation (F (5, 34) = 34.5, $p < 0.001$) and a significant interaction between initiation and continuation (F (25, 14) = 3.6, $p = 0.008$). There was no significant effect of initiation, which indicates that the initiations did not bias the rating. Instead, the participants generally used the entire scale for each initiation (see Figure 5). The main effect of continuation, and the interaction between initiation and continuation, remained significant when corrections were made for violations of sphericity according to the

Greenhouse-Geisser epsilon (F (3.8, 143) = 35.6, p < .001 and F (12, 456) = 13.2, p < .001, respectively). Tests of simple effects showed that there was a significant effect of initiation on the ratings of continuations 1, 4, 5 & 6 ($p < 0.001$)[2].

Figure 4 plots the average rating for each continuation. It shows that continuations 1 and 4 were given, on average, a higher rating than continuations 3 and 6. Figure 5 plots the average rating of the continuations split per initiation. It shows the interaction between the rating of the continuation and its context. Continuations 1 and 4 were rated high in most contexts except in the contexts of initiations 5 and 6. Continuations 5 and 6 were rated high in their own context and low in all other contexts.

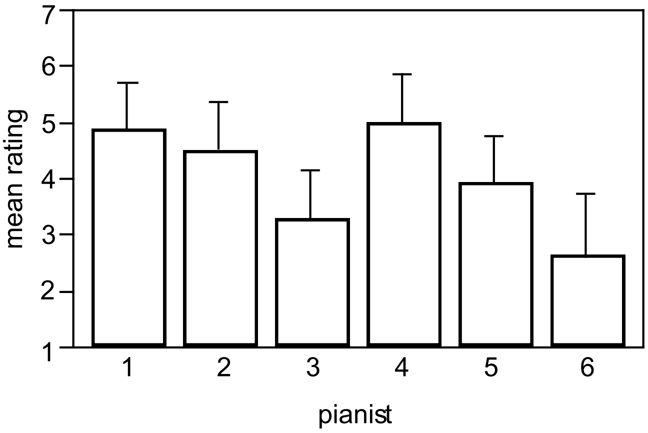


Figure 4: Mean quality ratings for each continuation across all subjects and all initiations. Capped bars indicate 1 standard deviation.

In other words, for each initiation, there were continuations that were considered to fit much better or worse than others. The continuations that were rated high were often the same ones in different conditions (continuations 1 and 4), as were the continuations that were rated low (continuations 3 and 6). There was, however, also a strong interaction between the rating of the continuation and the context in which it occurred.

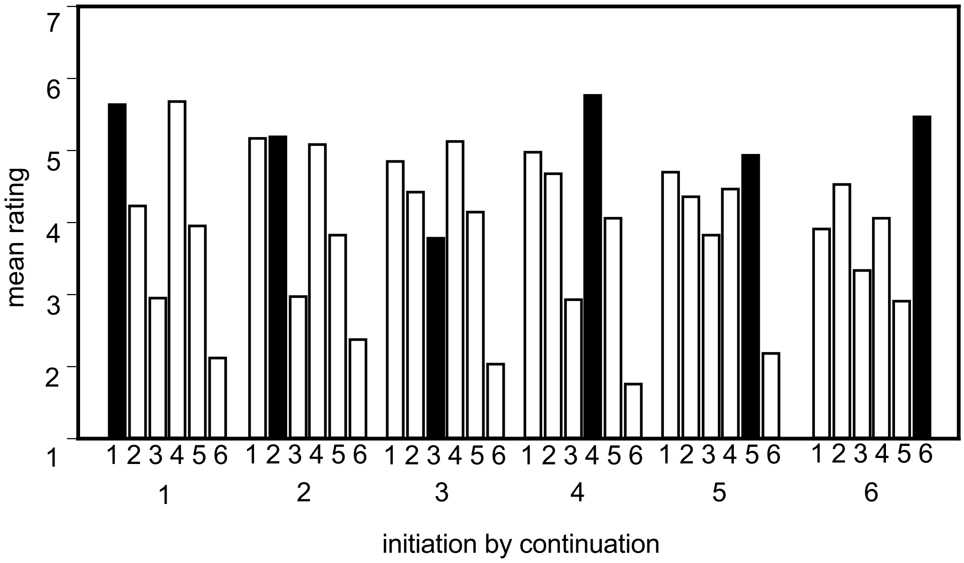


Figure 5: Mean quality ratings for each continuation of each initiation across all subjects.

Experiment 2

For experiment 2, a repeated measures ANOVA tested the effect of pianist, segment and the interaction between pianist and segment on the rating of aesthetic quality of the performances. There was a significant effect of pianist (F (5, 15) = 10.4, p < 0.001) and a significant interaction between pianist and segment (F (5, 15) = 6.5, p = 0.002). There was no effect of segment and sphericity was not violated. Tests of simple effects showed that there was a significant effect of segment on the rating of pianists 3 and 5 only (p < 0.001)[3].

Figure 6 plots the average quality rating of each segment for each pianist. It shows that for pianist 3, the initiation was rated high, but the continuation was rated low. For pianist 5, the opposite was true. Both segments of pianist 6 were rated high, while the segments of pianists 1, 2 and 4 were given intermediate ratings.

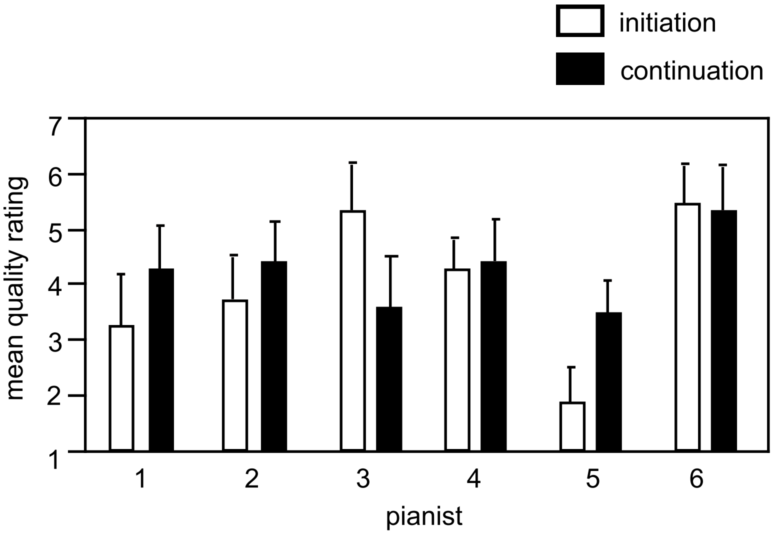


Figure 6: Mean quality ratings for each segment of each pianist across all subjects. Capped bars indicate 1 standard deviation.

Discussion experiments 1 and 2

As mentioned before, a comparison between the results of the two experiments is intended to show whether the evaluation of expressive variations was dependent or independent of context. A regression analysis tested whether the average rating of the initiations and the continuations of experiment 2 could explain the average ratings of experiment 1. This model did not reach significance (F (2, 33) = 0.45, p = 0.64) and had an R^2 of only 0.03. This low R^2 is not surprising, if we compare Figures 4, 5 and 6. Clearly, the quality ratings of the separate performances cannot explain the quality ratings of the combined performances – particularly the generally low rating of continuation 6 and the generally high rating of continuations 4 and 1. Neither can it explain the changes in the rating of the continuations in the context of different initiations.

Aesthetic preference might have played a larger role in the judgements of individual participants. Therefore, the same regression analysis was done for the participants who participated in both experiments. The aesthetic ratings that each participant gave in experiment 2 were used to explain the quality judgements that he or she gave in experiment 1. For these 20 participants, the regression analysis reached significance in only 6 cases. On average, the $\underline{R}^2$ was 0.15 and ranged between 0.00 and 0.45. For the significant models, it was mostly the continuation that significantly contributed to the explanation. The contribution of the initiation reached significance for only two participants.

To summarize, the two experiments showed a strong preference for context dependent ratings above context independent ratings. This was firstly demonstrated by a significant interaction between the rating of the continuations and their initiations. It was secondly demonstrated by a low predictability of the ratings of the combined performances of experiment 1 on the basis of the ratings of the separate performances of experiment 2. Part of the main effect of continuation in experiment 1, such as the high average rating of continuation 1 and the low average rating of continuation 6, was not due to general aesthetic preference, but rather was due to context. The following model is an attempt to formalize this finding.

Model

The final question of this study is to formulate the grounds on which a performance was considered to be a good continuation of the initiation and define those aspects of the initiation that were expected in the continuation. This formalization leads to a prediction of the quality of the continuations in the context of the initiations, which can be compared to the observed ratings given in experiment 1.

The proposed model assumes that the quality judgement of the combined performances of experiment 1 related directly to the similarity (or lack of difference) in expression of the second segment to the first segment. In order to calculate the similarity in expression, the two segments were first characterized along five variables that were easily transferable from the first segment to the second segment and that captured salient interpretive choices of the pianists. The decision was made to characterize the interpretation of the accompaniment notes, the interpretation of the grace notes, and the phrasing of the melody. This was done by measuring a) the amount of rubato in the accompaniment notes, b) the average articulation of the accompaniment notes, c) the duration of the grace note, d) the average asynchrony between the melody and accompaniment notes, and e) the pattern of key velocities of the melody and the accompaniment notes (see Table 1). The amount of rubato was measured as the standard deviation of the accompaniment IOI's. The articulation was measures as the onset-to-offset duration of the notes divided by the inter-onset-interval (IOI) between subsequent notes. The grace note was defined as the time interval between the onset of the grace note and the onset of the following main melody note. The asynchrony between the melody and the accompaniment note was generally positive, which indicates that the melody was generally ahead of the accompaniment. It became negative if the grace note was performed on the beat and the melody main note was considerably delayed.

The key velocity pattern could be either flat, rising, or falling, which separates the performances with a clear dynamic phrasing from the performances without such clear phrasing. The performances with clear phrasing have a rise in dynamics towards the middle of the phrase and a fall in dynamics towards the end of the phrase (see e.g., Todd, 1992). The differentiation between these envelopes was made on the basis of regression analyses. For each half of each performance and for each voice, a line was fitted to the key velocity of notes with increasing score time. If this line fit showed a significant increase or decrease of key velocity with increasing score time, the key velocity pattern was assigned a rising or a falling pattern, respectively. If there was no significant fit, the pattern was regarded as flat. The rising patterns only occurred in the first half, and the falling patterns only in the second half[4]. A second order line fit would have yielded the same results as this linear line fit did. The performances without a significant fit showed no trend in the overall key velocity level. Their key velocity pattern would be better described as a saw-tooth pattern.

The similarity between the expressive characteristics of each performance combination was calculated by first taking the absolute difference between the rubato, articulation, grace note, and asynchrony values of the segments as shown in Formulas 1-4. The variable D stands for the difference measure of rubato (rub), articulation (art), grace note duration (grace) and asynchrony (asyn), respectively.

$$D_{rub} = \left| std(IOI_1) - std(IOI_2) \right| \tag{1}$$

$$D_{art} = \left| \overline{dur_1 / IOI_1} - \overline{dur_2 / IOI_2} \right| \tag{2}$$

$$D_{grace} = \left| graceIOI_1 - graceIOI_2 \right| \tag{3}$$

$$D_{asyn} = \left| \overline{asyn_1} - \overline{asyn_2} \right| \tag{4}$$

These difference values were then inverted to similarity measures (capital S in formula 5) in a range between 0 (maximal difference) and 1 (equality) for a certain variable x.

$$S_x = 1 - \frac{D_x}{\max(D_x)} \tag{5}$$

The similarity between key velocity patterns was measured in a different way. The combinations of patterns were assigned to a hierarchy. There were four possible combinations: rise-fall, rise-flat, flat-fall, and flat-flat. Of these, the combinations rise-fall and flat-flat were considered to be consistent interpretations, or combinations with similar key velocity treatment in both halves. In addition, the combinations with a falling key velocity at the end had the benefit of having a clear phrase-ending, which has been shown to be a preferred characteristic of a performance (see e.g. Todd, 1992; 1995; Friberg, Sundberg, & Frydén, 1994; Clarke and Windsor, 2000).

The ordering of pattern combinations was therefore: The pattern rising-falling was best (S = 1.00). The pattern flat-falling was second best (S = 0.67). The pattern flat-flat was second worst (S = 0.33). And, finally, the pattern rising-flat was the worst combination (S = 0.00). The intermediate cases in which the fall was only present in one voice were given an intermediate value. The effect of context was the penalty for a flat second half following a rise in key velocity. This is in contrast to the higher rating of a flat continuation that follows a flat initiation.

Validation

A regression model with the similarity in expressive variables as continuous factors was fitted to the judgement data of experiment 1 and optimal weights were calculated for the different factors (see Formula 6). The model was highly significant ($p < 0.001$)

and explained the data fairly well; it had an $\underline{R}^2$ of 0.22 with all between participant variance taken into account. All factors significantly contributed to the explanation, except grace note duration, as shown in Table 2.

$$Rating = i + a * S_{rub} + b * S_{art} + c * S_{grace} + d * S_{asyn} + e * S_{vel} \qquad (6)$$

Table 2. Weights and significance values of the parameters of the consistency model fitted on the judgment data of experiment 1.

Parameter Similarity in	Weight	$\underline{p}$ value
Rubato	2.09	< 0.001
Velocity	0.76	< 0.001
Articulation	0.94	< 0.001
Grace note duration	0.40	0.051
Asynchrony	1.23	< 0.001

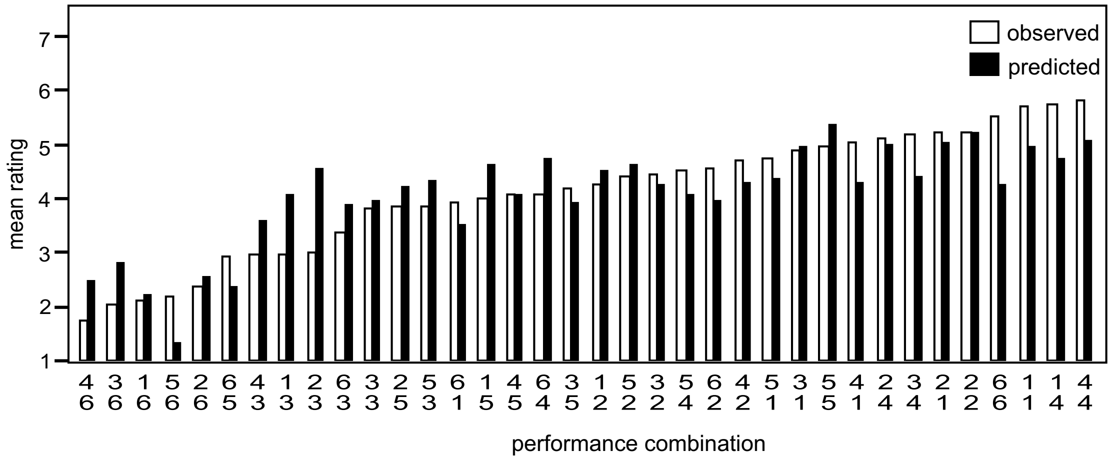


Figure 7: Predicted rating and mean observed rating of experiment 1 for each performance combination.

From the regression analysis, it was possible to predict the quality ratings of experiment 1. Figure 7 shows both the observed and predicted quality ratings. The predictions approached the observed values very closely ($\underline{r} = 0.83$). It was, however, too optimistic for the combinations that included continuations 3, and too pessimistic for certain combinations that included continuations 1 and 4. The low evaluation of continuation 3 may have been due to the participants' dislike of this performance. The high evaluation of continuations 1 and 4 may have been due to a lack of better options; they were the best combination, given the set of performances.

The same regression analysis was fitted for each participant separately, since different factor weightings for each participant could lead to a better prediction of the observed ratings. Appendix 1 shows the $\underline{R}^2$ obtained in these analyses as well as the $\underline{p}$ values of the whole model and the factors that contributed significantly to the fit. For most participants, the model explained the observed ratings well. For 15 participants (out of 39), more than 50% of the variance was explained. For 6 other participants, however, the model did not reach significance. On average, the $\underline{R}^2$ was 0.43, which is considerably higher than the average $\underline{R}^2$ of 0.15 obtained in the regression between individual results of experiments 1 and 2.

Note that in most cases only one factor contributed significantly to the explanation, which suggests that the participants generally valued only one of the suggested variables. There were only three participants for whom three factors reached significance. Similarity in rubato was the factor that most often reached significance (for 22 participants), followed by similarity in asynchrony (for nine participants), grace note duration (for six participants), articulation and key velocity (both for five participants).

To summarize, the quality of the continuations of experiment 1 can be fairly well predicted by a model that takes the similarity in rubato extent, key velocity pattern, articulation, grace note duration, and asynchrony into account between the continuation and its preceding initiation. A similar rubato, articulation, grace note duration or asynchrony means a similar amount of these variables in both halves. A similar key velocity pattern means that a rise is followed by a fall. It also means that a flat key velocity followed by a flat key velocity is judged to be better than a rise followed by a flat profile. The variable that was most influential was the similarity in rubato, followed by the similarity in asynchrony, though not all participants valued the variables equally strongly.

General discussion

The conclusion is that for the tasks in this study it was possible for intrinsic considerations to overrule general performance preferences. It may therefore be concluded that intrinsic constraints on the expressive performance of music exist independently of general performance rules.

Before making this into a firm and general conclusion, the reliability and the generality of this finding should first be checked, since this study has dealt with limited musical material and with specific tasks for the participants. The first question of reliability concerns the design of the experiments and the participants' tasks. What strategy did the participants use in the two experiments and how could this strategy have influenced the results? In the first experiment, the participants were asked to rate the quality of the continuation given the initiation. Asking this is something between asking for an aesthetic judgement and a judgement of goodness of fit. As long as the participants did not explicitly give a similarity judgement, the rating is what has been aimed at. The participants reported that they found it difficult to give an analytic evaluation of the second half with respect to the first half and instead judged the quality of the continuation more intuitively, which could mean more aesthetically. Nevertheless, the large difference in the results of the first and the second experiment strongly suggests a different strategy for the first and the second experiment, or considerable inconsistency in giving an aesthetic judgement. Inconsistency seems unlikely for experiment 1 given the high predictability of the results by the model. Differences in strategy are more likely and are actually an aimed result. Assuming that they were not giving a similarity judgement, they must have given a (more or less intuitive) judgement of goodness of fit. Experiment 1 and the model have therefore given insight into contextual constraints on expression.

The generality of the results is restricted by the brevity and simplicity of the musical material, the choice of material that did not contain a full start or ending of a piece, and the experimental conditions in which the performances were recorded including some limitations on the level of expertise of the pianists. These experimental recording conditions resulted in considerable confusion between the pianists, which was beneficial for the experiment, because it led to graded judgements of the quality of the continuation and not to a polarised judgement (either a good or a bad continuation). In a more realistic situation, the specific identity of each pianist would have probably been clearer. Secondly, short and simple music was chosen to optimise the probability of a correct answer; they were expected to facilitate the judgements. It is easier to attend to the performance if the music is simple and a preference for consistency is more likely if expression is compared within a phrase. This means that for more complex music and for longer pieces the effect of consistency might have been less pronounced. The influence of the pianist on the character of the music might have been greater, but the definition of consistent expression might have been more ambiguous, and the demand for consistency might have been less strong or might even have been replaced by a demand for change. What would remain is a context-effect: the framework is set by initial variations and following expressive variations are interpreted accordingly; they are interpreted as consistent or as deviating depending on their relation to the previously established norm. And this study has demonstrated that expressive variations can indeed set a norm.

Acknowledgements

This work was undertaken as part of the ‘Music, Mind Machine’ project at the Nijmegen Institute for Cognition and Information (NICI) which is funded through a Pionier grant from the Netherlands Organization for Scientific Research. Part of the research was done at Northwestern University with financial support of the University of Nijmegen. I would like to especially thank Ric Ashley, Henkjan Honing and Peter Desain for support and help. I also like to thank Laia Viladot, Makiko Sadakata, Paul Trilsbeek and Yvonne Schouten for their aid with the experiments. I got very helpful comments from Eric Clarke, John Michon, Dirkjan Povel, Charles de Weert and three anonymous reviewers. Finally, I am thankful to all who participated in the pilot studies and the experiments.

References

Clarke, E. F. (1985). Structure and Expression. In P. Howell, I. Cross, and R. West, Musical Structure and cognition (pp. 209-36). London: Academic Press..

Clarke, E. F. 1995. Expression in performance: generativity, perception and semiosis. In J. Rink (ed.), The Practice of Performance: Studies in Musical Interpretation. Cambridge: CUP, pp. 21-54.

Clarke, E., F. & Windsor, L. W. (2000). Real and Simulated Expression: A listening study. Music Perception, 17, 277-314.

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

Friberg, A., Sundberg, J. & Frydén, L. (1994). Recent musical performance research at KTH. In J. Sundberg (ed.), Proceedings of the Aarhus symposium on Generative grammars for music performance1994, 7-12.

Gabrielsson, A., & Juslin, P. N. (1996). Emotional expression in music performance: Between the performer's intention and the listener's experience. Psychology of Music, 24 (1), 68-91.

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

Honing, H. (1990). POCO, An Environment for Analysing, Modifying and Generating Expression in Music. In Proceedings of the 1990 International Computer Music Association (pp. 364-368). San Francisco: CMA.

Palmer, C. (1989). Mapping Musical thought to musical performance. Journal of Experimental Psychology, 15 (12), 331-346.

Repp, B. H. (1997). The Aesthetic Quality of a Quantitative Average Music Performance: Two Preliminary Experiments. Music Perception, 14 (4), 419-444.

Repp, B. H. (1998). Obligatory 'expectations' of expressive timing induced by perception of musical structure. Psychological Research, 61 (1), 33-43.

Sundberg, J., Friberg A. & Frydén, L. (1991a). Common Secrets of Musicians and Listeners: An analysis-by-synthesis Study of Musical Performance. In P. Howell, R. West, & I. Cross (ed.), Representing Musical Structure (pp. 161-197). London: Academic Press.

Sundberg, J., Friberg, A., & Frydén, L. (1991b). Threshold and preference Quantities of Rules for Music Performance. Music Perception, 9 (1), 71-92.

Timmers, R., & Desain, P. (2000). Vibrato: Questions and Answers from Musicians and Science. Proceedings of the Sixth International Conference on Music Perception and Cognition. Keele, UK: Keele University, Department of Psychology.

Timmers, R., Ashley, R., Desain, P., Honing, H., & Windsor, L. W. (2002). Timing of ornaments in the theme of Beethoven's Paisiello Variations: Empirical Data and a Model. Music Perception, 20 (1).

Timmers, R. (2002). Freedom and constraints in timing and ornamentation: investigations of music performance (pp. 85-109) Maastricht: Shaker Publishing.

Todd, N. P. M. (1992). The dynamics of dynamics: a model of musical expression. Journal of the Acoustical Society of America, 91 (6), 3540-3550.

Todd, N. P. M. (1995). The kinematics of musical expression. Journal of the Acoustical Society of America, 91, 1940-1949.

Appendix 1

subject	$\underline{R}^2$	significant parameters	$\underline{p}$ value
1	0.44	art	0.003
2	0.36	asyn	0.014
3	0.41	rub	0.006
4	0.43	rub, vel	0.003
5	0.40	rub	0.007
6	0.30	asyn	0.045
7	0.48	rub, asyn	0.001
8	0.64	grace, asyn	0.000
9	0.52	rub	0.000
10	0.12		0.538
11	0.50	asyn	0.001
12	0.18		0.295
13	0.27		0.084
14	0.40	rub	0.008
15	0.48	rub, art, asyn	0.001
16	0.61	rub	0.000
17	0.39	rub	0.008
18	0.42	rub	0.005
19	0.58	rub, art, grace	0.000
20	0.64	rub	0.000
21	0.53	rub, vel	0.000
22	0.47	rub	0.001

23	0.61	rub, asyn	0.000
24	0.30	grace	0.046
25	0.52	rub, vel	0.000
26	0.50	vel	0.000
27	0.31	asyn	0.041
28	0.55	rub, art	0.000
29	0.24		0.129
30	0.56	rub, grace	0.000
31	0.30		0.049
32	0.45	rub,	0.002
33	0.53	rub, grace	0.000
34	0.29		0.060
35	0.21		0.187
36	0.39	art, grace,asyn	0.008
37	0.51	rub	0.000
38	0.66	rub, vel	0.000
39	0.29		0.056

Results of regression analyses with similarity in rubato, key velocity, articulation, grace note duration, and asynchrony as independent variables and rating of experiment 1 as dependent variable.

[1] The performances can be found on www.nici.kun.nl/mmm/.

[2] Significant at the $\alpha = 0.008$ level.

[3] Significant at the $\alpha = 0.025$ level.

[4] Surprisingly, only the key velocity profiles could be easily characterized. The IOI profiles showed generally much less clear patterns.