EVALUATING RECORDED PERFORMANCE

An investigation of music criticism through *Gramophone* reviews of Beethoven piano sonata recordings

Elena Alessandri

Submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy
at the
Royal College of Music, London

September 2014

Declaration

I hereby confirm that the entire submission is my own work and has not been submitted for a comparable academic award.

ABSTRACT

Critical review of performance is today one of the most common professional and commercial forms of music written response. Despite the availability of representative material and its impact on musicians' careers, there has been little structured enquiry into the way music critics make sense of their experience of performances, and no studies have to date broached the key question of *how* music performance is reviewed by experts. Adopting an explorative, inductive approach and a novel combination of data reduction and thematic analysis techniques, this thesis presents a systematic investigation of a vast corpus of recorded performance critical reviews.

First, reviews of Beethoven's piano sonata recordings (N = 845) published in the *Gramophone* (1923-2010) were collected and metadata and word-stem patterns were analysed (Chapters 3 and 4) to offer insights on repertoire, pianists and critics involved and to produce a representative selection (n = 100) of reviews suitable for subsequent thematic analyses. Inductive thematic analyses, including a *key-word-in-context* analysis on 'expression' (Chapter 5), were then used to identify performance features (primary and supervenient) and extra-performance elements critics discuss, as well as reasons they use to support their value judgements. This led to a novel descriptive model of critical review of recorded performance (Chapters 6, 7, and 8). The model captures four critical activities – evaluation, descriptive judgement, factual information and meta-criticism – and seven basic evaluation criteria on the aesthetic and achievement-related value of performance reliably used by critics, plus two recording-specific criteria: live-performance impact and collectability.

Critical review emerges as a highly dense form of writing, rich in information and open to diverse analytical approaches. Insights gained throughout the thesis inform current discourses in philosophy of art and open new perspectives for empirical music research. They emphasise the importance of the comparative element in performance evaluation, the complexity and potentially misleading nature of the notion of 'expression' in the musical discourse, and the role of critics as filters of choice in the recording market. Foremost, they further our understanding of the nature of music performance criticism as a form of reasoned evaluation that is complex, contextual and listener specific.

ACKNOWLEDGEMENTS

My first thanks go to my advisor, Aaron Williamon, for the support throughout this work, endless readings of chapter drafts and for the always straightforward and reassuring attitude in difficult moments. Your flexibility and openness made this research journey possible beyond distance and logistical constraints. Thank you for your warm hospitality and friendship, and for treating me as a researcher, and not as a student, from the beginning on.

I owe the utmost gratitude to my second supervisor, friend and mentor Hubert Eiholzer. His contribution impregnates this whole work and extends far beyond it. You saw the path I was supposed to go far before I did, and supported me throughout the way with disarmingly altruism. Thank you for all our absorbing discussions, debating everything from artistic integrity to wine tasting. You nurtured my curiosity, passion for research, and thirst for clarity. Thank you for inspiring me.

I would like to thank my practical consultant, Olivier Senn, for the personal and institutional support to this research, and for trusting me with a research outside your field of expertise, allowing me the freedom necessary to carry on with the project. Your careful, thoughtful readings and almost painstaking obsession with graphs and diagrams significantly improved this work in the final stage.

A special thank you goes then to my 'fourth' *de facto* advisor, Vicky Williamson. During your ten months in Lucerne you have given invaluable inputs to this research and helped me understand what it means to be an academic, and what kind of academic and person I would like to become. I would love to be one day half as good as you are in dealing with colleagues, students, and busy schedules! Thank you for the endless chats on the most disparate research (and off-research) topics, for your patience with my thousands questions and for all the intense hours spent on the review texts.

This project has been partially supported by the Swiss National Science Foundation, the Swiss State Secretariat for Education, Research, and Innovation, the Conservatorio della Svizzera italiana, and, in larger part, from my current institution, the Lucerne University of Applied Sciences and Arts, School of Music.

The whole research would have not been possible without the *Gramophone* permission to access and study their published material so extensively. I am thus most grateful to Luca Da Re and the magazine commission for their collaboration.

I would also like to thank Prof. King and the Institute for Quantitative Social Science at Harvard University for permission to use their text analysis algorithms and for the methodological support in this regard, and Alessandro Cervino for his assistance in collecting data.

Many other persons have contributed to the completion of this work in different ways over these four years. Thank you to Dorottya Fabian, Thüring Bräm, and Daniel Leech-Wilkinson for their valuable comments and inspiring conversations. Thanks to Noola Griffiths and Ivan Hewett for their feedback on the first part of the thesis, and to Rosie Perkins and David Weasley for their advice on methodology and data analysis. I would also like to thank Jürg Huber for taking me on concerts, and letting me experience first-hand the process of critical review writing and the challenges, even ethical ones, that this work may entail.

Finally, I would like to thank the people that have encouraged and supported me with their friendship and love. Thank you Lisa, Geordie, Liliana, Sara, Louise and all the CPS staff for letting me feel part of the group although I was so rarely there. Thank you Ruta for your PhD-fellow advices and for forcing me out of the office on sunny days (and not giving up on me not answering calls and messages).

Most of all, my gratitude and appreciation go to my family. To my mom and dad, who have always been there for me and my brother, and supported all my unconventional choices. To Sim, who is unarguably the best brother one could wish for, and Angela, for her enthusiasm whenever we meet, and her reluctance to let me go every single time. And to my Holger. He dealt with all my ups and downs with infinite patience and loving care, always reminding me of the beauty waiting outside the doorstep. Your love and genuine simplicity make me so much stronger. I love you all dearly.

Elena Alessandri

TABLE OF CONTENTS

LIST OF PUBLICATIONS

The following are published outputs from this thesis.

Alessandri, E., Williamson, V. J., Eiholzer, H., Williamon, A. (2015) Beethoven's recordings reviewed: A systematic method for mapping the content of music performance critique. *Frontiers in Psychology.* doi: 10.3389/fpsyg.2015.00057. See Chapters 4 and 6.

Alessandri, E., Eiholzer, H., Williamon, A. (2014). Reviewing critical practice: An analysis of *Gramophone*'s reviews of Beethoven's piano sonatas, 1923-2010. *Musicae Scientiae, 18*(2), 131-149. doi: 10.1177/1029864913519466. See Chapter 3.

Alessandri, E. (2014). The notion of expression in music criticism. In D. Fabian, R. Timmers & E. Schubert (Eds.), Expressiveness in music performance: Empirical approaches across styles and cultures. Oxford: Oxford University Press, 22-33. See Chapter 5.

Alessandri, E., Eiholzer, H. and Williamon A. (2013). Between producers and consumers: Critics' role in guiding listeners' choices. In A. Williamon and W. Goebl (Eds.) Proceedings of the International Symposium on Performance Science 2013. Association Européenne des Conservatoires AEC, 671-676. See Chapter 9.

Alessandri, E., Eiholzer, H., Cervino, A., Senn, O., & Williamon, A. (2011). Investigating critical practice. In A. Williamon, D. Edwards and L. Bartel (Eds.) Proceedings of the International Symposium on Performance Science 2011. Association Européenne des Conservatoires AEC, 497-502. See Chapters 3 and 4.

LIST OF TABLES AND FIGURES

TABLES

Introduction

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 6

Chapter 7

Chapter 8

Chapter 9

INTRODUCTION

A question that has engaged and fascinated musicians, music lovers, philosophers and scientists for centuries now is what makes a good, bad, or great performance. Why are some performances received as uninteresting and others revered as masterworks? What are the features of a performance that make us thrill or shiver with pleasure or that leave us with a feeling of awe and respect for the great achievement they represent? To what extent is this goodness, or greatness, something 'true', something objective, so that our experience of it can be shared and agreed upon with our peers?

Given the intensity that the aesthetic experience of artworks can achieve, it is not surprising that these questions have drawn much attention. In the 20th century the importance of understanding the processes that underpin the evaluation and appreciation of musical performances has been further emphasised by the development of canons of practices and structures in Western Art Music that place the evaluation and assessment of performance at the core of our musical lives. Performances are constantly made subject to evaluation, from the informal after-concert chat to the verdicts of exam commissions or music competition committees. Besides being a natural part of the listening experience, evaluations – especially those provided by expert listeners – play important roles in shaping musicians' careers, music students' artistic development and listeners' choices about what recording to buy and what concert to listen to.

Considerable effort made in music research in recent decades has offered a partial understanding of the process of music performance assessment and has shed light on the complexity and density intrinsic to this phenomenon. This thesis makes a further contribution to the debate by investigating the way experts come to make their judgements of performances through the examination of a still unexplored source of material: music criticism. Music criticism is a common practice in Western musical traditions, and it probably represents the most complex form of evaluation of compositions and/or performances. Critics are seasoned listeners with rich experience in comparing and evaluating performances and in using words to articulate their impressions of musical products. Yet, this practice has been neglected

in performance evaluation research, such that there is currently no structured enquiry of the phenomena or outcomes involved.

Reasons for this are at least partly due to the complexity – and to some extent confusion – surrounding the terminology used to verbalize musical experiences. In these difficulties, though, resides the potential of a systematic investigation of music criticism that moves the study of performance evaluation into a new territory, free of some of the limitations imposed by the more quantitative studies usually found in the performance evaluation literature and strongly anchored in real world performances. In this perspective, such exploration of critical practice may offer novel insights that complement and inform the findings of the extant literature.

To this purpose, the present thesis reports procedures and findings of an investigation of expert judgements in critical reviews of recordings of Beethoven's piano sonatas published in the British magazine *Gramophone* between 1923 and 2010. Chapter 1 outlines the conceptual framework of the thesis, clarifies the importance of the chosen topic and the appropriateness of the proposed approach, and states the aims and research questions. It explores the current literature on performance evaluation, first looking at findings on inter- and intra-judge reliability, then reviewing studies on the validity of value judgements, and finally focusing on the investigations carried out thus far on music criticism. Chapter 2 is devoted to methodological considerations concerning the different approaches that can be taken in the analysis of unstructured texts. It reviews advantages and limitations of the different perspectives and presents then the methods chosen for the present research. Chapters 3 and 4 finally introduce the corpus of critical review that is the object of analysis in this thesis. In Chapter 3 an overview is offered of the material collected: it presents the critical texts, their structure, length and distribution across decades, as well as the repertoire reviewed and the main people involved – that is, pianists and critics. Chapter 4 gives a preliminary scrutiny of the critical texts that serves to frame the subsequent examinations. A five-step quantitative/qualitative data reduction procedure offers insights into the main objects discussed in reviews and the vocabulary used by critics. Drawing from these findings, a selection of reviews is produced to be used in the subsequent thematic analyses.

From the results of Chapter 4 a challenge emerges related to the notion of 'musical expression' that should be employed in this research. Given the importance

that expression has in the musical discourse, and at the same time the ambiguity that the word 'express' and its correlates can bring to the different notions of expression shared in different contexts (e.g., typical listeners, musicians, researchers), a question concerns how 'expression' should be understood and interpreted in the analysis of critics' writings. To address this, Chapter 5 presents a focused qualitative analysis examining what critics mean – or seem to mean – when they talk of expression in music and how this notion relates to their critical judgements.

Having drawn the conceptual and methodological framework in Chapters 1 and 2, and building on the overview, data reduction procedures, and focused qualitative analysis reported in Chapters 3, 4, and 5, Chapters 6 to 8 present the core in-depth analyses of critical review of recorded performance. First, Chapters 6 and 7 analyse what performance-related properties critics seek out for critical consideration (Chapter 6) and the valence with which these different properties are discussed (Chapter 7). This leads to the development of a visual descriptive model of performance judgements in critical review, and a model of performance evaluation that identifies seven basic criteria reliably used by critics to support their value judgements. Following, in Chapter 8, the focus is enlarged to embrace elements of the end-product recording other than the performance that critics discuss and that seem to enter their final, composite judgement of the recording as artistic product.

The final outcome of Chapters 6 to 8 is a novel model of critical review of recorded performance that for the first time offers a comprehensive analysis of critics' judgements as they are expressed in their published reviews. In conclusion to this work, Chapter 9 summarises research methods and findings for the six studies reported throughout the thesis (Table 9.1, p. 286), and discusses their scope, limitations and possible future applications. The following table offers an overview of the thesis structure.

Table 0.1. Thesis structure.

Chapter		*Content*
Introduction		
1	On the value of music performance	Conceptual framework, research aims and questions, methodological approach.
2	Methodological considerations	
3	Gramophone reviews I: An overview	Overview of the collected material and preliminary analyses.
4	Gramophone reviews II: Turning to the text	
5	A complex notion: Expression in music criticism	
6	Critics' judgements of performance	In-depth, inductive analyses of critics' judgements as they are expressed in their published reviews.
7	Valence of performance judgements	
8	Beyond performance: Reviewing recordings	
9	General discussion and conclusions	Overview of the applied methods and results obtained. Discussion of limitations and wider implications.

1 ON THE VALUE OF MUSIC PERFORMANCE

Evaluation is a natural component of the listening experience and permeates every aspect of the daily musical practice. Music performances are continuously subject to evaluation, from the informal after-concert chat to the verdict of the jury in international competitions. In fact, it can be argued that it is impossible to listen to a performance without instinctively making some kind of assessment, even if just in the form of a vague feeling of liking (or not) what we have heard (Thompson & Williamon, 2003).

Despite its ubiquitous presence, the evaluative dimension of music and of music performance is at its core paradoxical. On the one hand, musical value is usually seen as within the domain of taste and subjectivity *par excellence*: no one is wrong if she prefers listening to Rameau's *Le rappel des oiseaux* over Bach's *Schafe können sicher weiden*. Similarly, it makes no sense to tell someone that she should not prefer Emil Gilels's delicate and sublime rendition of Rameau's piéce de clavecin to Robert Casadesus's more energetic account. Appreciation of works of art is the realm of personal taste, and opinions about taste cannot be objectively right or wrong: *de gustibus non est disputandum*. On the other hand, however, we do treat artworks – and musical works and performances among them – as if they had an objective, measurable value. This is reflected in examination grades dispensed in conservatoires, in competition rankings, in audition verdicts assigned to measure achievement and aptitude, and even in online feedback on recordings quantified through stars or thumbs up/down.

The tension between these two apparently irreconcilable aspects of value judgements of works of art is not particularly new. One of the most celebrated discussions on this topic can be brought back to the eighteenth century and the philosopher David Hume. In his essay "Of the Standard of Taste" (1757), Hume addresses the problem of the subjectivity and validity of aesthetic judgements claiming that judgements of beauty, despite their being expressions of sentiment, are by no means arbitrary. Beauty, explains Hume, is no objective feature of an object but resides in the sentiment aroused in the person perceiving that object. In this sense

beauty is subjective. The sentiment of beauty, however, is elicited by features that exist in the object, and hence it is not arbitrary, but led by general principles of composition and internal structure of the work of art. Differences of judgements between people, continues Hume, are thus not due to the subjectivity of aesthetic value, rather to the nature of the judges. In fact, even though beauty is subject to universal principles, only few people are able to discern and recognize it appropriately. This ability is grounded on five traits: sensibility towards subtle nuances in the structure and composition of the work; experience in applying this sensitivity to artworks; exposure to a vast set of artworks that allows for comparison between them; freedom from prejudices; and 'good sense' as Hume calls it – that is, the ability to gain a sound understanding of the object of aesthetic contemplation. Those who possess these five traits are what Hume calls 'ideal critics'.

Ideal critics are those able to recognize the true value of artworks, but the status of *ideal* critic can never be fully reached, only approximated. Hence it is the commonly agreed judgement of expert judges (those who come closer to be ideal critics) that is taken as measure of the value of a work. Those expert judges are entrusted to provide a definite verdict on what artworks are valuable and what not: they are those who set the 'standard of taste'. Hume's solution to the problem of taste has been lengthily commented upon and explicated by several aestheticians, and it was recently brought to the fore once more by Levinson (2002, 2010). This is also a solution widely spread in our musical society and mirrored in the reliance on judgements agreed by expert evaluators, endowed with an authority grounded in their experience as musicians and as listeners. Two assumptions underlie the validity bestowed on these judgements (Thompson & Williamon, 2003):

(a) there is something like *the value* of a music performance, and
(b) appropriately informed listeners can perceive this value through attentive listening.

That is to say that the value of a musical performance is a common psychological reality for expert listeners. This hypothesis is critical to the process of evaluation, and given the centrality of evaluation and appreciation in musical practice, research has devoted large effort to its investigation. A conspicuous number

of studies in the last decades addressed the problem inspecting inter-judge reliability (adjudicators' agreement on the value of a performance) and intra-judge consistency (adjudicators' consistency in evaluating the same performance several times), while a different cluster of studies attempted at singling out diverse elements that may enter the experience of a performance and concur at the construction of the final assessment. The first two sections of this chapter address these areas of research and highlight some of the concerns that remain open. The third part of the chapter then proposes a new approach to the problem that exploits a still unexplored set of material and suggests that investigating this material can offer new insights in the way expert listeners formulate evaluations of real world musical performances. Based on this approach, specific research questions are delineated at the end of the chapter.

THE STANDARD OF TASTE

Music performances are constantly made object of evaluation, but not all evaluations are endowed with equivalent authority. When it comes to assessments that play a role in the academic or professional career of musicians or in general in their personal and artistic development, it is typically the opinion of experts that is taken as reference. Experts, understood as people with solid musical knowledge and long-lasting experience in listening to and evaluating performances, are those able to perceive fully what is and is not valuable in the performance and to offer a valid assessment that is a measure of the performance's 'true' value. If that is the case, experts' judgements are likely to converge significantly and to remain stable through time – that is, it is expected that experts agree on the value they perceive and are able to offer consistent opinions when asked to evaluate the same performance more than once.

These hypotheses have been examined in a large set of studies in recent years; however, their findings have offered mixed results and left some questions open. Before turning to these studies and discussing what these questions are, it is instructive to give an overview of the assessment modalities commonly used in empirical research as well as in most domains of everyday musical practice.

Modalities of evaluation

The Reasoning Model

In music as well as in other activities it is possible to distinguish between two types of assessment: the *norm referenced* assessment, in which a performance is assessed through comparison, as being better or worse of another performance, and the *criterion based* assessment, in which a performance is judged in isolation, set against a set of commonly agreed criteria. Norm referenced assessment is typical of music competitions, while a criterion based assessment is usual in academic contexts (McPherson & Schubert, 2004). In research, with few exceptions, performance evaluation is explored through a criterion based assessment procedure. This assumes that in evaluating a series of performances each will be judged and listened to in isolation, as if there were no other performances before or after. An attempt to validate this assumption was carried out by Wapnick, Flowers, Alegant and Jasinskas (1993), whose study showed that listeners' preferences for a given performance were not influenced by preferences for the performance immediately preceding it; however, further research will be required to estimate the extent to which it is possible to evaluate in a non-comparative way, especially when more interpretations of the same piece are proposed. Currently, the common procedure employed to compensate at least partially for possible comparison effects remains counterbalancing the order of reproduction of stimuli across sessions.

In both norm referenced and criterion based assessments, the common assumption is that there are some parameters that guide the evaluation of a performance and that the application of these parameters on the side of the evaluator can lead to a meaningful assessment. This in turn implies that an evaluative judgement of a work of art is not a mere description of one's experience of the work, rather the outcome of a rational act. This idea is encapsulated in what is called the Reasoning Model, of which the most influential contributor and supporter was the American philosopher Monroe Beardsley. According to the model, judgements of value can be supported by means of reasons; and reasons, for being valid ones, must explain why the judgement is true by means of appealing to qualities that are inherent in the artwork (Beardsley, 1968, p. 57). Following this, the Reasoning Model can be expressed in its simplest form through the formula: Performance P is

good/bad or better/worse than Performance W, because it possesses feature F, where F is a feature that resides in the artwork.

An open question related to the model of value judgements as judgements grounded in reason is whether it implicitly states that evaluators perform inferences while listening ('performance P possesses features A, B, and C, therefore it is good') or not. In the latter case value judgements are only connected to reasons and not inferred from them. That is to say that judgements come as immediate, instinctive responses to the performance – using Hume's words, as an expression of the sentiment aroused by the performance – and reasons are then sought out from the evaluator to explain the judgement by offering evidence in its support ('performance P is good. In fact, it possesses features A, B, and C'). The importance of this distinction becomes evident when turning to what are the two most widely used assessment modalities in music schools, competitions, and research: the *holistic* and the *segmented* assessment schemes.

Holistic versus Segmented schemes

Until four decades ago, the main assessment modality used in music was what Mills (1991) called holistic assessment. In the holistic assessment, the listener is required to give a feedback in form of one single grade which reflects the performance's overall quality, with no particular assumptions regarding the process conducting to the final judgement. This assessment type avoids *a priori* prescriptions, and leaves the assessor free to consider the performance as a whole, resulting in an evaluation that is "musically credible" (Mills, 1991, p. 179). Despite its ecological validity, the limitation of this form of assessment is the scarcity of information it offers. In fact, unless it is accompanied by further comment or feedback on the performance, a holistic judgement does not explain where the value of the performance (or lack thereof) lies.

In response to a need for a more explanatory form of assessment that would better serve pedagogical and research purposes, holistic assessment began to be replaced in the 1980s through a form of evaluation in which the final mark is seen as a composite measure – that is, a function of sub-evaluations of features of the performance singled out beforehand (Mills, 1991). This is what is called segmented assessment scheme. Segmented assessment can come in different forms and in the past decades there have been several attempts to construe and validate various

schemes. However, all of them usually include an overall quality mark, which can be a genuine overall mark – if the assessor is given freedom to assign the mark independently from the other criteria – or a factitious one – where the overall grade is computed as mathematical function of the sub-grades (Thompson & Williamon, 2003). The increasing implementation of segmented schemes has been driven by the belief that this modality could offer a higher degree of objectivity and accuracy of assessment (Stanley, Brooker, & Gilbert, 2002). In addition, even if the minimal level of intervention that characterises the holistic approach seems to assure a higher level of ecological validity, it can be argued that the segmented scheme optimizes the post-hoc utility in terms of amount of information offered by the assessment, while keeping a quantified and structured feedback that enables its use for pedagogical or research purposes (Thompson & Williamon, 2003).

Nonetheless, a few concerns accompany the choice of this assessment form. First, at least for those schemes that do not include an overall quality grade or that compute it as average of the grades of the different sub-traits, the segmented scheme assumes a process of evaluation that works inductively, in which the final judgement is inferred by the consideration of the different criteria. But the extent to which this is the case still needs empirical verification. Mills (1991) strongly expressed her worries about the fragmentation of the listening experience in supposedly meaningful sub-parts forced by segmented schemes. To clarify her point she commented:

> As I leave a concert, I have a clear notion of the quality of the performance which I have just heard. If someone asks me to justify my view, I may start to talk about rhythmic drive, or interpretation, or sense of ensemble, for instance. But I move from the whole performance to its components. I do not move from the components to the whole. In particular, I do not think: the notes were right, the rhythm was right, the phrasing was coherent, and so on – therefore I must have enjoyed the performance. (Mills, 1991, p. 175)

The addition of an overall mark does not seem to overcome this problem. Once the listener is led to ponder about different aspects of the performance, the overall evaluation will inevitably be coloured by the considerations of the single components. Indeed, according to Swanwick "the fudge of adding a category called 'overall' only makes things worse" (1996, cited in McPherson & Thompson, 1998, p. 19) by suggesting a holistic consideration of the performance which in fact did not occur.

Second, the use of segmented schemes implies a prescription regarding what quality indicators should be used in the evaluation of the performance. Despite attempts at assembling sets of criteria tailored for given performance contexts (for instance, the instrument specific assessment schemes used by Bergee, 2003) it remains uncertain what might be the correct criteria to apply in specific circumstances. Thompson, Diamond and Balkwill (1998) asked five expert evaluators to assess six performances of a Chopin étude. Prior to this, evaluators were asked to produce an assessment scheme entailing six bipolar constructs that would be used for evaluating the six performances. Despite a certain amount of overlap between the constructs, striking dissimilarities were found, suggesting that evaluators form their judgements based on criteria that are personal and maybe even specific for a given piece (e.g., for one evaluator bar 27 of the étude represented a decisive moment for the assessment of the performance). In the light of this, an external prescription of quality indicators as per segmented assessment schemes seems to bear the risk of marring the validity of the judgement, forcing listeners to focus on aspects of the performance that might not be the ones at which they would instinctively choose to look if allowed to do so freely.

A third important limitation of segmented marking schemes is that they do not account for the relative weighting of the separate criteria. A marking scheme requiring the assignment of a mark up to five for each of the categories of technical flawlessness, stylistic appropriateness, expressiveness and stage presence, with the overall grade defined as the sum of the single marks, assumes that each of the four criteria is equally important for the construction of the performance overall value. A performance that scores five in the first three criteria and zero in stage presence would therefore have the same final mark as a performance that scores five for stage presence but zero for, say, stylistic appropriateness. This problem might be irrelevant for performances that are even in the different components but becomes palpable in performances that display the diverse features less uniformly (Mills, 1991, p. 174).

An in-depth investigation of examiners' perceptions of the use of segmented schemes in music exams was run in 2002 at the Sydney Conservatorium of Music (Stanley, Brooker, & Gilbert, 2002). Here segmented schemes were introduced in the 1990s, shortly after the merging of the Conservatorium with the University of Sydney. In the study, 15 staff members of the Conservatorium, most of them with

more than 20 years of experience in music performance assessment in tertiary music schools, participated in semi-structured interviews aimed at enlightening their perception of the usefulness of criteria-based schemes versus holistic assessments.

Results showed mixed attitudes towards the use of segmented schemes. Some participants felt criteria might be beneficial in a pragmatic perspective, in that they offer a way to "focus on what is assessable" (p. 52), which is helpful especially in doubtful cases. They also are effective when it comes to provide specific feedback to students. Concerns, however, were raised regarding the narrow view of the performance aroused by criteria, with the consequent loss of the big picture. This can be interpreted in two ways: on the one hand focussing on the criteria forces a fragmentation of the experience of the performance that might be unnatural (Mills, 1991). On the other, criteria might be limited or inappropriate to the circumstance, thus not allowing for a comprehensive vision of the performance. These concerns were expressed in the following participant's statements:

> if I meet you as a human being, I don't say to you this and that about your hair or about your eyebrows or about the fact that you wear glasses… I get the total picture of you as a person and then I come out with a general statement that sums up my feeling about you as a person, for me that is all important. (Stanley et al. 2002, p. 52)

> I don't think (allocating grades) can have any connection to the criteria because you can get a kid that plays out of tune and out of time but you are crying because it is so expressive or so wonderful. You can (also) get a kid that plays dead in tune or dead in time and absolutely immaculate dynamics that leaves you totally cold… So that again, I can only assess in a total package after the bread is cooked and not say the flour is off and the yeast did not work and stuff like this. (Stanley et al., 2002, p. 52)

Quantitative holistic and segmented assessments represent the most widely spread modalities of evaluation in the current musical practice and almost the only ones in music research. Even though mixed feelings surround both modalities, they have been used as standard procedures in most studies investigating the process of evaluation hitherto. Notwithstanding the feeling of easiness or appropriateness evaluators may experience while using one or another assessment modality, it can be claimed that the utility of a rating scheme is monotonically related to its content validity. That is, as long as a given assessment method enables a reliable

measurement of the construct that it is supposed to be measuring, then its utility is assured. The content validity of a given assessment scheme has been tested by looking at the reliability of judgements given by expert evaluators.

The agreement of experts

Studies measuring reliability of value judgements have offered mixed results both in terms of degree of agreement that can be expected and of the role musical expertise plays in the enhancement of that degree (Kinney, 2009).

Judges' consistency

In one of the first studies on evaluators' internal consistency, Fiske (1977) asked 33 recent graduates of a music education programme to assess 20 different performances of one solo trumpet on five performance traits: intonation, rhythm, technique, phrasing, and overall quality, with the overall mark defined as a separate trait (and not as an average of the other four). Applying a test-retest procedure, Fiske let participants listen to and rate each performance twice and computed correlation coefficients for each judge. These coefficients were then used as a dependent variable to control for the effect of practical music expertise and theoretical music knowledge. Average level of internal consistency was moderate (0.60) and quite variable, ranging from 0.32 to 0.82. Overall marks showed a higher level of reliability on average (0.71). To test the influence of practical music expertise and theoretical music knowledge on judgement reliability, a trait intercorrelation matrix was calculated using data on the students' grades in applied music, music history and music theory during their undergraduate studies. Unexpectedly, different levels of practical music expertise did not affect the consistency degree, while theoretical music knowledge was in a statistically significant inverse relationship with level of consistency. This suggests, according to Fiske, that disciplines like music theory or history offer little aid to the development of performance assessment skills, or may even become detrimental to it. The reason for this could lie in the dissimilarity between abilities required in one and the other activity: providing absolute responses based on factual knowledge for the former; develop subjective decisions based on a continuous comparison process for the latter.

A lack of effect of musical experience on the internal consistency of evaluations was also found by Wapnick et al. (1993). They investigated consistency of value judgements of piano performances, but differently from Fiske, they employed an indirect procedure to compute consistency. Two sets of seven interpretations of the same piece each (one slow excerpt and one fast excerpt of Liszt's *Totentanz*) were used to build two groups of 21 trials, in which the seven interpretations were presented in all possible two-combinations. Eighty pianists were randomly assigned to one of the two sets of 21 trials and were asked to choose, for each pair, the performance they preferred. Consistency was computed by deriving triad combinations from the 21 trials and testing for consistency within each triad. If an evaluator preferred performance A to B in the first trial, and then performance B to C in the next trial, it would be expected that in the trial proposing the combination A-C, s/he would prefer A. Controlling in this way for each possible combination of three performances, Wapnick et al. (1993) derived for each participant an 'overall consistent triad score' ranging from zero to 35. This score was used to test effects of expertise, dividing participants into four groups:

- Undergraduate piano majors who had been undergraduates for less than 2 years,
- Undergraduate piano majors who had been undergraduates for more than 2 years,
- Experienced pianists who were former piano majors, and
- Experienced pianists who were university faculty.

Beside expertise level, Wapnick and his colleagues were interested in testing the usefulness of two wide spread practices in academic environment: the use of notation while listening to the performance and the use of segmented assessment schemes. Therefore, participants were randomly assigned to one of four conditions: preference only (P), preference plus score (PS), preference plus rating scales (PR) and preference plus score plus rating scales (PSR).

Internal consistency was found to be independent from level of musical experience, although consistent triad scores for faculty were nominally (but not significantly) higher. Use of score and of segmented scheme did not enhance

consistency either. The only near-significant difference ($p = 0.06$) found was between the two music excerpts. The authors felt that these results indicated that not just criteria but also the possibility of providing reliable judgements depends upon the specific music piece heard. Further investigation, however, is required to test this hypothesis.

Judges' reliability

Similarly to consistency level, studies on reliability between judges often found low to moderate degree of agreement among evaluators. Thompson and Williamon (2003) asked three expert adjudicators, a pianist, a cellist and a clarinettist to assess sixty-one performances of music major students using a 14-item assessment scheme which included an independent overall quality mark. Computing correlation coefficients for each of the six combinations of the three judges, Thompson and Williamon found an average reliability of 0.50, range 0.33 – 0.65.

It was also noted a high level of multicollinearity between the different components of the rating scale. Thompson and Williamon proposed four possible explanations for this:

- Judges used the rating scales carelessly – that is, after having assigned an overall mark, they simply skimmed through the other traits without appropriately reflecting upon them;
- Judges were unable to discriminate between different traits. This may lie either in a lack of discriminating abilities or in the chosen assessment criteria being inapt to reflect the real nature of the performance;
- Judges were able to discriminate between categories, but categories were correlated with each other;
- All performances in the sample were even in the different parameters.

Thompson and Williamon's aim was to investigate if music performance evaluation performed with the currently available procedures could be used as a reliable research tool. Given the low degree of reliability and the apparent lack of discrimination displayed in the assessment scheme, they concluded that this was not necessarily the case.

Other studies have gathered more positive results. Bergee (1997) in an investigation of the relationships between faculty, peers, and self-evaluations of musical performance, found peers' as well as faculty panels' inter-judge reliability to be acceptable to good. The assessment scheme used was segmented, but tailored for each group of instrument. Collinearity between categories was not explored. Again, higher level of musical expertise did not enhance reliability, and peers' reliability coefficients, ranging from 0.83 to 0.89, were more stable than faculty members', who spread from 0.23 to 0.93 (alpha coefficients). The wide range obtained by faculty, however, was due to one specific percussion panel that performed with very low agreement. Two of the three jurors in this panel were assistant teachers; therefore, Bergee felt that the results suggested experience and expertise actually mattered. This conclusion, however, seems to collide with the rest of the findings and, in particular, it does not explain why all student panels performed with high consistency.

These results were supported in a later study (Bergee, 2003). Here as well specific segmented schemes were used for different instrument families (brass, percussion, woodwinds, voice, piano, strings). Categories for the rating scales were derived from previous studies that validated segmented schemes through factor analysis (fact-factorial approach). For piano no such scale was available, therefore categories were developed through literature study and discussion with piano faculty (Bergee, 2003, pp. 140-141). Evaluators were recruited among faculty members and teaching assistants who agreed in compiling the rating scale for a selection of performances they attended during end-of-semester exams. In addition to completing the sub-item rating scale examiners were asked to assign an overall letter grade (A+ to F). Panel size ranged from two to five hence reliability was computed through coefficient of concordance (Kendall's W). Bergee found that reliability of the letter grade for each family of instrument was good to high independent of panel size. Sub-item ratings also displayed a good correlation for all instruments except percussion (and with the exception of the category 'Suitability' for *voice*). An examination of differences between levels of experience among examiners indicated that expertise did not improve reliability.

The role of expertise

That musical expertise does not heighten reliability of judgements as suggested by the studies discussed so far seems to be counterintuitive. It is important to interpret these results considering the nature of participants in these tests. Fiske (1977) explored the role of expertise comparing grades in applied music by undergraduates of a music education programme; Wapnick et al. (1993) and Bergee (1997, 2003) distinguished between music major undergraduates, graduates and faculty. All these groups are representative of people with at least a solid basic musical understanding and who had a conspicuous amount of musical exposure. Even if the lack of significant difference between those groups is surprising (indeed faculty and not peers in schools are entrusted with the authority of expert evaluators), it is still possible that exploring a wider range of musical expertise may offer different results.

This hypothesis was investigated by Kinney (2009). In his study, 63 undergraduate non-music majors and 42 undergraduate music majors were compared. Non-music majors were further divided in those who did not have any previous formal training in music (so called non-participants, n = 28) and those who had at least two years formal study in high school music ensembles (participants, n = 35). All participants listened to ten different performances of three songs (total = 30 excerpts) and assessed them for 'accuracy' and 'expression'. Performances were prepared using MIDI software that allowed for subtle changes in different musical parameters. For each song five performances were repeated twice, to allow for test of consistency. Computing internal consistency with Pearson's product-moment correlation Kinney found significant differences between different groups of expertise, with music majors displaying the strongest internal consistency in both accuracy and expression ratings (Table 1.1).

Table 1.1. Mean Pearson's product-moment coefficients for non-participants, participants, and music majors as reported by Kinney (2009, pp. 329-331).

	Accuracy	*Expression*
Non-participants	0.10	0.18
Participants	0.35	0.41
Music majors	0.62	0.64

Scheffé post-hoc comparisons showed a significant difference also between non-participants and participants – that is, already after a mere two years training in high school music ensembles. These results combined with those of the studies discussed above seem to suggest that music exposure, besides music training or expertise, strengthens judgements' reliability and consistency.

Open concerns

Concerns remain on the possibility of relying on the agreement of experts in pursuing valid performance assessments. Studies displayed moderate degrees of reliability and the implementation of segmented rating scales did not improve the performance substantially. Bergee (1997, 2003) obtained higher levels of agreement implementing instrument tailored assessment schemes; however, comparing results might be difficult given the diverse procedures used to compute reliability.

A more serious concern, however, regards the nature of the information that studies on reliability offer. Even where high level of agreement in assessing the value of a performance is found, this alone does not imply that the performance value is a common psychological reality for listeners, since no information is given on the evaluation process that brought to the final assessment. One and the same overall quality rating could be reached using different criteria, or applying the same criteria in diverse ways, for instance weighting them differently. High level of assessment reliability does not rule out the presence of biases either, nor can it be taken as evidence of the content validity of the assessment schemes. It is true that judges having a common perception and understanding of the performance heard, applying the same criteria in the same way, being free from prejudices, having the same level of experience and musical exposure that allows for similar comparisons among performances and making use of the assessment scheme in similar way as a tool to summarize in one rating their judgement of the performance will inevitably come to a perfectly agreeing assessment. However, the same agreement can also be obtained, for instance, by a judge perceiving the performance as stylistically highly appropriate even though not very expressive and with a few minor technical uncertainties and one who finds it extremely expressive but with important technical issues, and who is biased by the attractiveness of the performer. In similar ways, judges whose opinions

converge in terms of musical value of the performance might come to different ratings through a different use of the scale as a quantification of their impression.

The use of segmented schemes apparently offers a solution to sort out at least a few of the possible alternative explanations to a lack (or not) of agreement. However, this assessment system leaves too many open questions – not least those related to the high level of intercollinearity among sub-parameters found by Thompson and Williamon (2003) – for the sub-traits ratings to be taken as evidence of more accurate information on the evaluation process. Hence, assessment reliability as it is tested in empirical research cannot be taken as a measure of listeners' agreement in Hume's sense of the word. What is of interest for both artistic and pedagogical purposes is to gain understanding of *how* experts construe their evaluations, what features of performances they seek out for critical consideration, how they perceive the interplay between different features, and what are exogenous elements that might create biases even in experts' perception of the performance. The second part of this chapter discusses how music research has addressed these questions so far.

THE PROCESS OF PERFORMANCE EVALUATION

Besides examining experts' reliability and consistency, a large corpus of research has sought understanding of the phenomena of performance evaluation and appreciation by investigating the influence of selected factors in the construction of the final assessment. Results of these studies offer glimpses of what are some of the elements that enter our experience of a performance and may account for the diversity of judgement among listeners.

The body of research dealing with the validity of value judgements of performances is vast, and various factors were examined in those studies. A first attempt to organize these into a model of performance evaluation was offered by McPherson and Thompson (1998). This model took into account contextual, musical and non-musical elements, as well as characteristics of the performer and of the evaluator. The model not only offered an overview of the studies on performance evaluation run at that time, but it attempted to bring to the fore a long series of unexplored performance elements that may have an influence in the final assessment, calling for further research to investigate those elements. Since 1998, several studies addressed some of the issues highlighted by McPherson and Thompson. Six years

later, McPherson and Schubert (2004) proposed a simplified version of the performance assessment model, aimed at suggesting what potential areas of interventions are available to a performer to enhance her performance assessment. The model represents a pragmatic tool for musicians to reflect upon the complexity underpinning the assessment of their performances, and to distinguish between those elements they can have an influence on and those over which they cannot hold any control. Drawing from this second model, in what follows an overview is offered of some of these elements. The overview does not aim to be comprehensive, but it attempts to illuminate mechanisms underpinning performance evaluation through a review of state-of-the art empirical and theoretical studies in music research.

McPherson and Schubert's model of performance assessment

Based on the assumptions stated at the beginning of the chapter, that the value of musical performances is a common psychological reality that expert listeners can perceive through attentive listening, the process of evaluation can be summarised by the schematic diagram in Figure 1.1. Accordingly, the process of performance evaluation involves assessing the performance by weighing it against a set of commonly agreed musical criteria. The value is given by the extent to which the performance meets or does not meet those criteria.

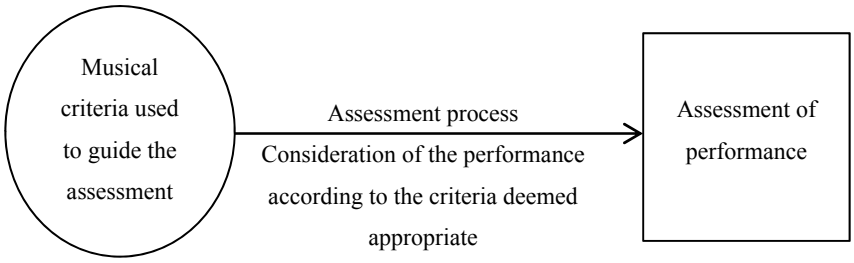


Figure 1.1. The assumed performance assessment process as depicted by McPherson and Schubert (2004).

However, as McPherson and Schubert explain, in the everyday musical practice the assumed assessment process described in Figure 1.1 does not hold. Even if we retain the assumption that there are commonly agreed parameters that can be used as valid reasons to justify a value judgement, we also ought to acknowledge that

there are other parameters or factors that do not represent valid reasons for evaluating a music performance, but which are there and affect the final judgement. A better description of what music performance evaluations involve needs thus to take into account extra-musical and non-musical factors that interfere in the assessment process, together with measurement error, which comes naturally together with any attempt to measure something and which is due – at least partly – to the susceptibility to error by any evaluator. Hence, McPherson and Schubert argue that the process illustrated in Figure 1.1 should be expanded to include also these other factors (see Figure 1.2). The distinction between musical, extra-musical and non-musical factors is far from neat. Particularly extra-musical factors, as McPherson and Schubert explain, are an "unclearly defined, fuzzy set" of elements, whose location is subjective and partly dependent on the performance circumstances. The distinction between these categories in the model was guided by the criterion that regarding extra-musical factors performers "might be able to use some knowledge about the factor to systematically enhance their performance assessment" (p. 65). Some examples of non-musical, extra-musical and musical factors are offered in the following sections.

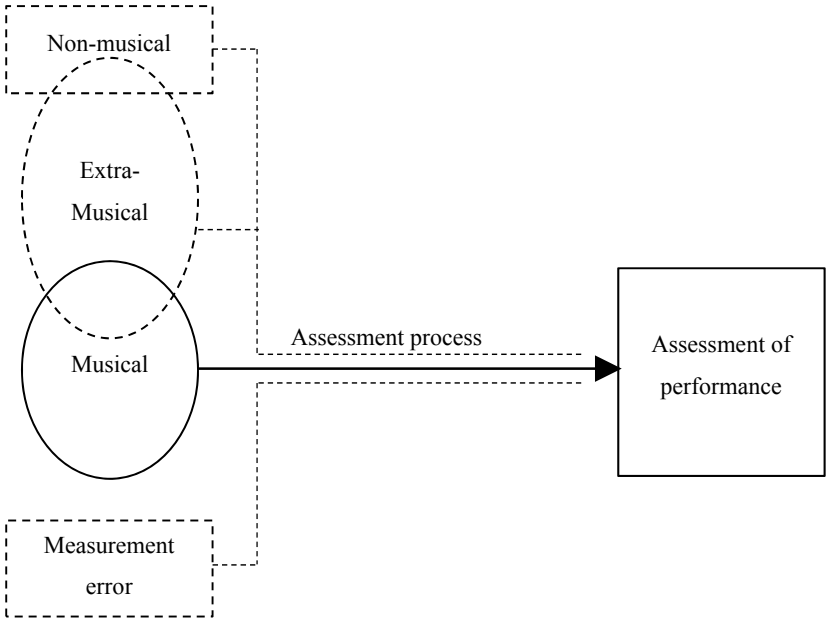


Figure 1.2. An expanded model of the performance assessment process (McPherson & Schubert, 2004).

Non-musical factors

Non-musical factors are elements that are not supposed to enter the evaluation of a performance; when they do, they affect the validity of the judgement creating unfair biases (McPherson & Schubert, 2004, p. 73). Evaluators are usually unaware of those biases and performers cannot have control over them. Two clear examples of this kind of factors are cultural preconceptions, like race or gender biases, and contingent factors like – in the case of a series of performances as in a music competition or school exams – what has been listened to prior to the given performance.

Gender and race

Given the history and cultural tradition of Western Art Music, it is possible that gender and race bias could influence performance evaluation. Davidson and Edgar (2003) reported that among more than 200 undergraduates who read music at Sheffield University between 1989 and 1999 only two were non-Caucasian. Also, O'Neill and Boulton (1996; cited in Davidson & Edgar, 2003, p. 170) stressed the persistence of gender-instrument prejudices among peers, that place flute among 'female' instruments and drums among 'male' ones, with cello and piano occupying gender-neutrality.

A ground-breaking study in music performance assessment addressing gender and race biases was run by Elliott (1995/6). Elliott filmed eight musicians – four flautists and four trumpeters – performing the same piece. For each instrument performers were chosen to represent the following four categories: White-man, White-woman, Black-man and Black-woman. One flute and one trumpet audio recording was then dubbed over all videos. Participants – 88 graduates and undergraduates with major in music education – were asked to evaluate the different performances without being aware that they were in fact evaluating different videos but always the same audio file. Results showed a general race bias, so that Black performers were rated lower than White performers. A gender-instrument bias also emerged, with female flautists rated higher than female trumpeters, thus supporting the thesis of cultural stereotypes concerning what instrument is appropriate for men and for women.

A more recent study by Davidson and Edgar (2003), which compared ratings given to Afro-Caribbean and European pianists, did not find support for the race bias,

but provided further evidence for gender bias, with female pianists rated consistently higher than male pianists. The lack of race bias is encouraging, even though as the authors highlight, this should be read in relation to the nature of the jurors who took part to the study and the places and cultures in which they live: 36 young musicians, many of whom had attended "multi-racial schools in particularly vigorous and innovative anti-racist education schemes" (p. 180).

Order of performances

Another factor of which many musicians are aware, but mostly have no control over, is the order in which performances occur. There is evidence that when a series of objects are observed and evaluated assessments are influenced by the order in which the objects are presented. In 1956 Filipello found evidence of biases in wine tasting that favour the first sample tested (Filipello, 1956) while gymnastic jury evaluations were found to be influenced by the within team order in which the athletes competed (Scheer & Ansorge, 1975). Plessner (1999) found that this bias occurs both in the encoding phase – the period in which the assessor perceives elements and errors of the performance – and in the evaluation phase. That suggests that not just the final evaluation is influenced, but rather that the experience of the performance is affected as well.

In music performance, the most famous study in this direction was carried out by Flores and Ginsburgh (1996). They studied the results of the Queen Elisabeth competition contests run between 1951 and 1993 (21 contests: 10 for violinists and 11 for pianists) and investigated how the order of appearance in the third stage of the competition influenced the final ranking. The Queen Elisabeth competition is one of the most important and best-known events for piano and violin. It is divided into three stages, and only 12 performers are allowed to the third and final stage. Here, players have to perform among other pieces one 'unknown' concerto composed for the occasion and given to the musicians seven days previous to the public performance. Candidates perform at a rate of two per day and the order of appearance is randomly assigned. Flores and Ginsburgh analysed two 3 x 6 contingency tables, containing six columns for the six days of appearance and three rows for the final placement of the performer (rankings where collapsed into the groups: 1^{st} to 4^{th}, 5^{th} to 8^{th}, and 9^{th} to 12^{th}). The results showed that candidates performing on the first day had fewer chances of getting a high rank than candidates

who performed on the last three days. In particular, performing on the fifth day turned out to be the best predictor for achieving a high placement. Based on these results, the researchers suggested that the growing familiarity with the 'unknown' piece due to repetitive listening may affect the jury appreciation of that piece thus favouring those players performing in later days. A different interpretation is that jury members have higher expectations and thus more strict rules at the beginning of the competition. In subsequent days, these expectations may be adjusted to the reality of the actual performances. Irrespective of the explanation, the fact remains that final evaluations of performances were not independent from the order in which they occurred. Results of this study were also supported five years later through a different investigation of the same data by Glejser and Heyndels (2001).

Extra-musical factors

In McPherson and Schubert's (2004) model the category of extra-musical factors is large and varied and entails elements that might easily be seen as having the right to be considered in the assessment of a performance – like communication with the ensemble or use of expressive variations – together with factors that seem rather to belong to the non-musical group – like performers' attractiveness and flair. Given the variety and quantity of factors that fall into this group, McPherson and Schubert propose a sub-distinction between performer-related, context-related and adjudicator-related aspects.

Performer-related aspects

Physical appearance and stage behaviour

A large number of studies in personality and occupational psychology over the last four decades has given evidence to a positive correlation between physical attractiveness and different outcomes like helping behaviour, teachers' judgements of student intelligence, perceived job qualification, predicted job success, and hiring decisions in job interviews (for a meta-analysis of experimental studies in this domain see Hosoda, Stone-Romer, & Coats, 2003). In music as well different studies suggest that physical appearance may play a role in the perception and evaluation of performance.

Davidson and Coimbra (2001) analysing jurors' discussions of second- and third-year vocal student exams at the Guildhall School found that singers' appearance was a major factor in the evaluation of the performance. This could perhaps be explained by the importance that stage presence and bodily language have for singers' performances. However, a series of studies run by Wapnick and colleagues at McGill University suggested the presence of an attractiveness bias also for violinists' and young pianists' performances (Wapnick, Darrow, & Dalrymple, 1997; Wapnick, Mazza, & Darrow, 1998, 2000). Comparing audio only and audiovisual ratings of singers (1997) and violinists (1998), the researchers found that more attractive performers were rated higher in the audiovisual treatment than in the audio only. However, in both studies, more attractive female performers were also rated higher in the audio only condition. This led researchers to suggest that an attractiveness bias may occur already in the training stage, so that more attractive children obtain more attention and encouragement from early on. This hypothesis was supported in their third study (2000); looking again at both audio and audiovisual versions of performances, they found that more attractive children at their third year of piano study were rated higher than less attractive ones.

Further studies suggest that attractiveness bias may function in a complex way, for example affecting different groups of performers differently depending on their level of skill or gender. Ryan and Costa-Giomi (2004) in an investigation of evaluation of novice pianists' performances found that attractiveness bias favoured the more attractive pianists among the female performers, the more attractive pianists among the best players and the less attractive pianists among the male performers. The first study on attractiveness bias among top-level performers was run two years later by Ryan, Wapnick, Lacaille and Darrow (2006), and it showed results that to some extent contradicted those of previous studies. Among high-level professional performers (participants at the Van Cliburn piano competition), Ryan et al. found that attractiveness did not influence pianists' ratings, or if it did, it favoured the less attractive performers and those showing a lower level of stage behaviour. Care should be taken, however, in interpreting these results. Wapnick and colleagues in discussing their findings point out that attractiveness differences between performers were in fact – according to the perception of the researchers – not so striking. All performers were felt by the authors to be attractive. Hence, in line with the results of

their previous studies, they suggested that an attractiveness bias at training period may have contributed to a selection among performers, eliminating early on the non-attractive ones.

Body movement and gestures

A recurrent finding in studies on attractiveness and stage behaviour has been a general increase in average ratings from audio to audiovisual conditions (Ryan et al., 2006, p. 568). More research will be required to interpret this properly; however, as Ryan et al. suggest, it seems that in evaluating the audio or the audiovisual version of the performance judges are actually assessing two different objects: the playing in the first case and the performance as a whole in the second. Research in recent years has offered plenty of evidence of the importance of the visual component on the perception of sound and of music.

One landmark discovery in this direction was the serendipitous detection of the McGurk effect in speech perception (McGurk & MacDonald, 1976). McGurk and MacDonald noticed that contrasting stimuli between lip movements and produced sound induced listeners to perceive a stimulus different from the one actually played. Dubbing for instance the sound 'ba' over the lip movement for 'ga' listeners reported to perceive 'da', while reversing the dubbing process listeners perceived a combination of the two syllables ('bagba' or 'gaba'). Similar effects of cross-modal sensory interaction were found on the perception of loudness in hand clapping (Rosenblum & Fowler, 1991), on the timbre identification of pluck and bow sounds in cello playing (Saldaña & Rosenblum, 1993) and on the perception of sound length in marimba playing (Schutz & Lipscomb, 2007). These studies suggest that the brain's attempt to make sense of what we experience leads to a fusion or combination of stimuli that might elicit aural illusions. These do not necessarily have negative effects. For example, as Schutz and Lipscomb suggest following their findings, marimba players may take advantage of this sensory interaction to induce a perception of different sound lengths in an instrument that would otherwise not allow it, by applying wider or narrower gestures to the sound production.

The influence of the visual component of performance is not limited to listeners' perception of sound but enters the cognitive dimension of music as well. Davidson (1993) asked four violinists to perform a piece of their choice 'without

expression', 'with normal expression' and 'with exaggerated expression'. Twenty-one music undergraduates rated the expressivity of each performance in audio only, video only and audiovisual conditions. The aim of the study was to examine in which condition participants would be more accurate in identifying the performer's expressive intention. As results showed, this was the case for the video-only condition, followed by the audiovisual one. Responses to the audio only condition displayed the least degree of discrimination between levels of expressivity. Davidson concluded that visual stimuli carry important information for the listeners' perception of expressivity in music. More recently, Vines, Krumhansl, Wanderley, and Levitin (2006) investigated the role of visual information in the perception of tension and musical structure. Through an analysis of performances by two clarinettists for which 30 musically trained listeners gave continuous judgements of tension and phrasing, the researchers found that musicians' movements served to both augment and diminish the perceived tension at diverse points throughout the pieces and to shape the structural understanding of the music by highlighting the beginning of new phrases or changes in emotional content.

Context-related aspects

Instrument and acoustics

Among the contingent factors that may influence the outcome and evaluation of a performance the ones musicians are most familiar with and aware of are instrument quality (for pianists and other musicians who cannot afford to carry their own instrument with them) and room acoustics. These are factors players can at least partially control, for example by making sure to have rehearsal time to test those conditions and to adjust the interpretation accordingly.

Purpose of the performance

The purpose of the performance must also be taken into consideration. Musical performances can serve different purposes, and these may call for different interpretive and performative choices. The interpretation appropriate to a performance in an international music competition may not be the same suitable to a school exam, gala concert, rehearsal, or recording. Originality for example may be a property desirable in a concert but less so in a school exam. Similarly, a recording may require different interpretive choices than a live performance given the

repeatability of recorded sounds and their lack of visual element (excluding video recordings). A strong *rubato* or an emphasised *subito forte* for example may add expressivity and sense of surprise in concert, but in a repeated listening they may over time become annoying. Likewise, as Rostropovich suggests, pauses should be tightened when performing for a recording, since the lack of bodily communication does not allow for tension to be held over a long silence (Katz, 2004, p. 27). Also within the same medium and type of performance, interpretation may need to vary to fit the specific circumstances. As Levinson (1987) affirms, what is the most suitable interpretation of Haydn's Symphony no. 80 for a performance done in the course of the musicology conference 'Haydn: The Music of the Future' would probably not be the same of the one prepared for the conference 'Haydn: The Rococo Roots', and the performance of the same Beethoven's sonata may have a different purpose (and therefore should be evaluated differently) if it is meant to be an isolated performance of that one sonata or instead part of the performance of Beethoven's sonata cycle.

Dress code

Purpose and context of the performance extends beyond interpretive and performative matters to include social factors like dress and stage behaviour etiquette. Wapnick et al. (1998) while investigating the evaluation of performances by six female and six male violinists found that in evaluating performance quality players rated higher for appropriateness of dress and stage behaviour were favoured in videoed recordings.

Recently, dress biases were isolated and addressed specifically by Griffiths (2008, 2010). Griffiths recorded four female violinists performing a classical, jazz, and a folk music piece in three different outfits: concert dress, jeans, and a nightclubbing dress. A further 'point-light' condition was added that allowed jurors to follow body movements without recognizing dress or appearance of the performer. Each combination of performer, piece and dress was presented to listeners twice: once with the performer's own interpretation and once with a master track interpretation provided by a fifth (male) violinist dubbed over the video. Thirty musicians were recruited among students at Sheffield University and members of the Sheffield Philharmonic Orchestra to evaluate the technical proficiency and musicality of the different recordings, as well as performer's attractiveness and

appropriateness of dress. The findings supported the hypothesis that expectations of dress affected performance evaluations. Listeners seemed to have a clear idea about what dress was appropriate for which musical style, with concert dress rated highly suitable to classical repertoire but not so for jazz or folk music, and casual dress fitting jazz and folk pieces more than the classical one. The different outfits however also related to evaluations of musical features: across all three pieces, technical proficiency was rated significantly higher for performers wearing concert dress than for those wearing jeans and nightclubbing dress; nightclubbing dress also scored the lowest mean rating for musicality.

The results suggest a dress code bias in the evaluation of performances. However, as Griffiths proposes, different factors should be accounted for in the interpretation of the findings. On the one hand, drawing from Citron's mind/body split (1993, in Griffiths, 2010, p. 171), it could be argued that the nightclubbing dress, attracting attention to the body, prompted an idea of woman strongly associated to her physicality that marred listeners' focus on the performer's musical abilities. On the other hand, as the author noted, the nightclubbing dress might have actually bounded performers' movements, both through the physical impediment given by the tight outfit and because of the uneasiness performers may have felt in presenting themselves in such revealing clothes. And listeners may have perceived this lack of freedom and spontaneity in the violinists' gestures. In any case, the study supports the idea of performance evaluation as a process influenced by visual as well as aural stimuli and coloured by a rich net of thoughts and expectations partly related to the listeners' cultural and social background.

Evaluator-related aspects

Mood and attention

It is reasonable to think that being alert or tired, or in a good or bad mood, may influence the listening experience. Using an example by Levinson (2004), if a man listened to a good piece of music right after his wife has died, he would expectedly not take pleasure in that experience. Similarly, it is reasonable to assume that the attention level while listening can increase and diminish and that a trough in this level may momentarily affect listeners' capacity to appreciate the performance (McPherson & Schubert, 2004; Thompson, Williamon, & Valentine, 2007). There is no research yet addressing the influence of mood, neither on music perception nor on

performance evaluation; however, a survey run among musicians and music lovers by Thompson (2007) showed that being relaxed and in a good mood are important determinants of listeners' self-reported enjoyment of a performance. During the past three decades a large set of studies in personality and social psychology has also shown that value judgements of diverse objects are formed not only on the basis of content information but also on the basis of feelings, like being in good or bad mood or having a positive or negative attitude toward an object (Greifeneder, Bless, & Pham, 2011, p. 107).

Preferences for the work

Beside the dress code bias, results of the study by Griffiths (2010) point also to a different preconception in performance assessment: one linked to the work being performed. In performance evaluation, it is reasonably expected that listeners – especially expert evaluators – will be able to differentiate between the quality of the piece performed and the quality of the performance. No one would find it acceptable if, in a chamber music competition, the performance of Brahms's A minor Trio Op. 114 would get the highest score because of the jurors' preference of Brahms's music over another piece of canon repertoire, say, Rimsky-Korsakov's Quintet in B-flat major for piano and winds. Yet in reality, this distinction may not be so neat. Preconceptions linked to different music styles and compositions could conceivably lead to different attitudes toward a performance. In Griffiths' study, performances of the folk music piece were rated higher in technical proficiency than those of the classical or jazz music across different dress conditions. Another instance of possible piece bias was found by Glejser and Heyndel (2001) in their investigation of rankings in the Queen Elisabeth Music Competition. Analysing rankings of finalists in all contests of the competition run since 1956 they found that performers who played a more recently composed concerto obtained a higher rank while those who performed a popular concerto – especially among violinists – were penalized, thus suggesting that juries appreciated innovation over popularity.

The problem of piece bias is compounded when we take into account the personal associations with past experiences that a piece may hold for the listener. The issue is delicate since music affects listeners on an emotional level, and

emotional reactions may in fact be far from conscious control and difficult to scrutinise deliberately.

Music's capacity to arouse emotional states has long been investigated, either through self-report on intense emotional experiences (Goldstein, 1980; Sloboda, 1991) or through measurement of physiological responses (Krumhansl, 1997; Lundqvist, Carlsson, Hilmersson, & Juslin, 2009; Rickard, 2004). In social, clinical and personality psychology, music is regularly used to induce emotional states to explore mood influences on diverse behaviours (see review of studies employing the "Musical Mood Induction Procedure" in Västfjäll, 2001-2002). Gorn, Pham and Sin (2001), for example, investigating the interplay between valence and arousal components of affective states in the evaluation of advertisements, found that different types of music (with positive, negative or ambiguous valence and with low or high arousal level) affected both arousal and valence component (happiness and sadness) of participants' mood. This in turn, coloured participants' evaluation of the ads. If that is the case, it could be argued that level of arousal and positive or negative valence of the repertoire played may influence the way the performance is perceived (McPherson & Schubert, 2004).

Familiarity with the work and expertise

It is straightforward that musically informed listeners at different levels of expertise will use different criteria in assessing a performance, in that their listening will be informed by a series of concepts and notions unknown to the untrained listener. From this it does not follow that expert listeners will be able to offer better, more reliable or more consistent evaluations as it has been shown by studies on inter- and intra-rater consistency (see Kinney, 2009, pp. 323-324), but merely that they will listen to the music differently and apply to their listening experience a different set of thoughts than a non-expert.

This concept is explored in Levinson's (1987) discussion of what he calls the "Perspective Relativity of Evaluation of Performance" (PREP). Levinson distinguishes between different kinds of expertise, differentiating between average listeners', performers', and composers' (including analysts, historians, musicologists, etc.) perspectives. A performance that casts light on the compositional process or on Schenkerian underlying form (p. 80) might be good for composers or music theorists

without being particularly rewarding for non-composing listeners.

He also stresses that we have to distinguish at least four types of audiences according to the familiarity of the listener with the work being performed: first-time listener, familiarized listener, jaded listener and the one-time listener. To make his point Levinson discusses as example the performance of the first movement of Schubert B-flat minor opus post. piano sonata. For the first-time listener, he argues, it might be advisable to choose a brisker tempo, which would probably make it easier:

> to sense the overall progression and span without losing interest. It might also be said that the continuity and flow of the various sections, evident to a practiced listener at a moderate tempo, are more readily grasped by the neophyte auditor if the basic pulse is somewhat accelerated. (p.78)

On the other hand, a jaded listener, the one "who knows the work so well that all its musical implications and realizations, as Leonard Meyer puts it, have been fully absorbed and internalized" would find a "standard performance almost sleep-inducing" (p. 78).

Likewise, interpretive details like stretching to the maximum permissible distension the half note relative to the two quarter notes in the ½ ¼ ¼ rhythm of the Andante with variations movement of the *Death and the Maiden* Quartet thus "imparting to the figure more of a pulsing or surging quality than it carries in more conventional readings" might seem intriguing to the jaded listener, annoying to the familiarized one and even confusing or misleading to the first-time listener (p. 78). It is true, continues Levinson, that the perspective of the familiarized listener is often taken as a central one. It may be that the familiarized listener has a privileged role in determining what other legitimate perspectives are; nonetheless, his is not the *only* legitimate one, and it should not be taken as paramount status for performance assessment.

A particular type of musical expertise is the music making skill – that is, the ability to play a certain instrument. Williamon and Thompson (2003) in a study on inter-rate reliability of performance evaluations, found evidence of a possible instrument bias: three expert evaluators – a cellist, a pianist and a clarinettist – were asked to listen to and evaluate sixty-one performances by students of the Royal College of Music. The researchers found that among students string players were

given lower grades than other players by the cellist evaluator, thus suggesting that the higher competence level led the evaluator to assess string players more severely.

Expectations and beliefs

A last reflection on the evaluator-related factors should be made in regard to the expectations of the listener. We have seen that prejudices and cultural stereotypes can enter the experience of the performance: also beliefs and expectations concerning the quality of the performance one is going to listen to can affect this experience. Duerksen (1972) asked two different commissions to evaluate the same recording. One commission was told the performance was an amateur recording by a music student, the other was told it was a professional recording produced by a high-level performer. Results showed that the presumed student's recording was given lower ratings. Duerksen conceptualises these results suggesting that thinking of the recording as a high-level professional recording might induce a feeling of awe and reduce self-confidence in criticising features we do not like of the performance.

However, recent studies in a different domain, that of food perception, seem to suggest that quality expectations influence not only the evaluation of a product, but also the physical experience of that product. Plassmann, O'Doherty, Shiv, and Rangel (2008) recruited 20 participants to taste and evaluate five different sample wines identified with their retail price: $5, $10, $35, $45, and $90. While tasting, participants' brains were scanned using fMRI. Unbeknownst to the subjects, there were only three different wines in the experiment, with two wines being proposed twice with different price tags (e.g., wine 2 was presented once with the tag $90 - its real retail price – and once with $10). Researchers found that more expensive wines were given higher ratings. More interestingly, they also found that more expensive wines actually triggered an increased activity in the medial orbitofrontal cortex (mOFC), a part of the brain whose activation has been correlated with behavioural pleasantness ratings for odours, tastes, and music (Plassmann et al., 2008, p. 1052). Even if it is not possible from these results to conclude that participants actually experienced more pleasure, the potential implications of the findings are intriguing: beliefs about the quality of the experience we are going to have do not only influence the evaluation of the experience, but actually change the experience itself. This

suggestion was also supported by a later study, again on wine tasting (Siegrist & Cousin, 2009).

Musical factors

Having reviewed non-musical and extra-musical elements that may influence the assessment of a performance, this section offers an overview of musical factors that enter performance evaluation. As discussed in the first part of this chapter, in recent decades segmented assessment schemes have increasingly substituted holistic assessment in music schools. McPherson and Schubert (2004) surveyed the published literature on evaluation criteria applied in those schemes and grouped the criteria into four competence domains: technique, interpretation, expression and communication. Table 1.2 shows their groupings.

Three main considerations are done based on these parameters. These concern the distinction between explanatory and non-explanatory reasons in the evaluation of works of art, the existence of general principle of musical value, and the relativity of the notion of value of musical performances.

Explanatory and non-explanatory reasons

A first observation that emerges from the list of criteria offered by McPherson and Schubert (2004) is that not all parameters listed relate to the aural dimension of the performance. Features within the domain of technique like *posture*, *bodily coordination* and *physical endurance* are properties of the performer, while in the case of *communication among the members of the ensemble* it is probably a combination of aural and visual features that are the object of observation. These non-aural properties relate genetically to the sounds of the performance, for it may be assumed that through a correct posture, secure coordination and a good level of endurance and stamina the player will produce a better performance.

Table 1.2. Musical factors in performance assessment as reported by McPherson & Schubert (2004, pp. 63-64).

Technique	Physiological	- Breathing - Posture - Relaxation-tension - Balance - Coordination
	Physical	- Sound: production, projection, and control of instrument/voice and consistency, clarity, and focus of tone across all registers and dynamic levels - Range - Intonation - Physical control (e.g. stamina, endurance)
	Instrumental	- Bodily coordination - Accuracy, assuredness, and facility of rhythm, pitch, articulations, dynamics, timing, as well as the degree to which errors undermine and detract from the overall quality of the performance - Pacing of performance - Sensitivity to intonation, both individual and ensemble
Interpretation		- Authenticity: understanding of the style/genre and established performance practice (e.g. use of a reliable edition) - Accuracy: based on faithful reading and/or memorization of the score, and realization and exploration of the composer's intentions - Musical coherence: perceptive choice of tempo, phrase shaping, dynamic shadings, sense of line, and understanding of overall structure
Expression		- Understanding of the emotional character of the work - Projection of mood and character of the work - Communication of the structural high points and turning points of the work - Sensitivity to the relationship between parts within a texture - Appropriate use of tone and color, light and shade, and / or drama
Communication		- Among the members of the ensemble (e.g. listening and leadership) - Confidence, as demonstrated in performances that are both convincing and purposeful - Interest, in terms of the degree to which the performer holds the audience's attention, maintains a sense of direction, creates a sense of occasion, and ends the work convincingly - Projection of expressive, interpretative, and structural features of the composition performed

However, it could be asked what weight these features should have in the evaluation of the performance. Correct posture is a feature that *mostly* enhances performance value by facilitating good sound production, but there may be players with terrible posture who perform greatly (representative examples among pianists being Glenn Gould and Keith Jarret) or, on the contrary, players who display an impeccable posture whose sound quality is poor. Beardsley (1968) in his classification of critical reasons distinguishes between two main types of reasons on which evaluative judgements can be grounded: those that *explain* why a work of art is good (or poor) and those that offer logical *support for* believing that the work of art is good (or poor) (Beardsley, 1968, p. 56). Beardsley's discussion of the validity of art criticism was focused on the evaluation of works of art – and not on the evaluation of performances of those works. Nonetheless, the distinction between explanatory and non-explanatory reasons may apply also to musical performances. Following Beardsley, explanations of why the performance is good could be quality of the sound produced, technical flawlessness, dynamic range, etc. On the other hand, claiming that a performer has a correct posture or good physical coordination and endurance could be reasons in support to the belief that the performance is good, since knowing about them would make us expect the performer to be technically sound and thus the performance to be technically flawless, the sound well controlled, etc. According to Beardsley, even though reasons of the second type are commonly used, they are not as relevant to the evaluative judgement of the work of art as are explanatory reasons, for they do not explain directly why the work is good.

General principles of musical value

A second observation based on the list in Table 1.2 is that some factors have a proper criterion-like nature, while others represent areas of competence or skill on which the evaluator should focus. For instance, *accuracy* and *musical coherence* are criteria, so that an increase in one of them, all the rest being equal, will produce an increase in the value of the performance. Elements like *intonation* and *range* could possibly be reduced to criteria by assuming an underlying principle of the like: 'the more precise the intonation, the more valuable the performance' or 'the wider the (e.g., dynamic or tessitura) range, the more valuable the performance'. On the other hand, it is not easy to see what are the principles corresponding to elements like *breathing*, *sound production* or the *pacing of performance*. These are elements that the evaluator is

suggested to focus on in her evaluation process. Each of these elements can be judged as being good or not, but it is up to the listener to decide what criteria to use to evaluate them – that is, to decide what it means for breathing to be good or for a performance to have a good pacing. It could be that, for instance, good breathing is breathing that is accurate and non-obtrusive, done so as to facilitate a clear and expressive communication of the structure and expressive character of the work. Good pacing on the other hand could be pacing that is appropriate to the style and character of the work and that allows for important details to emerge while maintaining a sense of direction and holding the audience's attention. There seems to be an overlap between criteria underlying different competence areas, and this in turn could imply the possibility of reducing the list to a small number of criteria which may apply to different elements or areas of the performance.

In the field of philosophy of art, the evaluation of musical performances has received little attention so far (one important exception previously discussed in this section is Levinson's argument on the relativity of performance evaluation, see Levinson, 1987). On the other hand, extensive work has been devoted to the discussion of criticism of works of art in general, including musical compositions. A long debate within this discussion concerns the existence (or not) of general principles of aesthetic value, so that with 'F' being a value adding feature for a work of art, any work of art that possesses 'F' will be more valuable than a work that does not, all the rest being equal. The most authoritative theory in defence of the existence of general principles of aesthetic value was proposed by Beardsley in 1962 and then re-proposed and revised in following years (1968; 1982). According to Beardsley's theory there are three and only three properties that are directly relevant to the evaluation of a work of art. These are unity, complexity and intensity. These properties are primary criteria in the sense that "the addition of any one of them or an increase in it, without a decrease in any of the others, will always make the work a better one" (1962, p. 485). Any other property of a work of art can be a valid reason to support a value judgement only to the extent to which it tends to increase one of these three primary properties. This however does not imply that these three properties are sufficient conditions of goodness, but only that they contribute to the goodness of the artwork.

A main critique advanced to Beardsley's theory is that of context dependency (Dickie, 1987; Goldman, 2005). Works of art are so varied that it is not possible to find any feature that will always be value adding for every artwork. And this is true even for the general principles of unity, complexity and intensity. So for instance, complexity may often be a merit, but a work could also be praised for its simplicity while an excessive degree of complexity could result in the work being chaotic. Beardsley answers this critique by construing the properties so that an excessive increase in one will determine a diminishing in another. In this view, simplicity is no longer seen as a lack of complexity but rather as a high level of unity. In a similar way a monotone work of art does not have an excessive degree of unity, rather a lack of complexity, and when a work is criticised for being chaotic what is meant is not that it has too much complexity rather too little unity. As Goldman (2005) points out, this tactic to avoid counterexamples removes content from the theory. Beardsley, continues Goldman, seems to be "correcting the critical practice instead of reflecting it, as his theory explicitly set out to do" (2005, p. 186).

A different answer to the context-dependency problem comes from Sibley (see Dickie, 1987). Sibley claims for it to be unnecessary to seek properties that are context-independent. The real distinction, according to him, is not about general and context-dependent criteria, rather between features that are aesthetically positively charged and those that are negatively charged. So, continues Sibley, properties like elegance and humour are always intrinsically positive, while garishness and sentimentality are always intrinsically negative, notwithstanding the fact that in the context of a specific work of art these properties may interact with other features to produce value or disvalue. Sibley's answer to the context-dependency problem is then that of leaving out from the definition of general criteria any reference to the context of the work of art:

> A property is a primary positive criterion of aesthetic value if it is a property of a work of art and if *in isolation from other properties* it is valuable. (in Dickie, 1987, p. 232, emphasis added)

Neither Beardsley nor Sibley seem to offer a definite solution to the problem of context-dependency of aesthetic properties, which remains one of the strongest arguments against the existence of general value principles, and in turn against the

possibility of delivering judgements grounded in reasons, which is at the core of the Reasoning Model itself. If each artwork is unique and incomparable with others, how is it possible to generate judgements? And similarly, if there are no general principles of value in art, so that all the rest being equal, property F is always a value-adding feature of an artwork, how is it possible to evaluate at all?

An answer to this question is formulated and defended in a recent contribution to the topic by Carroll (2009) that will be discussed in detail in the third part of this chapter. Carroll argues that it is not necessary to have rules of art in order to evaluate it. Indeed, says Carroll, what the critic evaluates is her experience of the artwork, and not if the artwork followed certain rules or not: similarly to food experience, the critic is interested in how the pudding tastes, and not in the degree to which the pudding was prepared in adherence with the recipe (Carroll, 2009, p. 26). Mills' (1991) considerations on the musical validity of holistic judgements seem to resonate in Carroll's claim. But Carroll continues: that artworks are unique is nothing more than a "Romantic and then Modernist fantasy" (p. 27). Artworks fall into categories; they belong to genres and styles. And within precise categories of artworks it is feasible to find principles that function as assessment criteria for all works in that category. The importance of classification – that is, of approaching an artwork as an instance of a certain type or category of artworks – for understanding and appreciating the work was already exemplarily portrayed and discussed in a seminal paper by Walton (1970; see also 1988). Carroll grounds on this notion his defence of the feasibility of objective evaluation. For, even if no general principles of aesthetic value can be found, it is possible to have principles specific for given categories of artworks which are general enough – within those categories – for the adjudicator to rely on in construing and supporting his/her judgement.

Value(s) of musical performances

Carroll's answer to the critique of the uniqueness of artworks defends the notion of objective evaluation. On the other hand, however, it points to the necessity of different assessment criteria that may suit different musical styles or genres. It is interesting to see how the desire for a theory of value reducible to a few general principles is not exclusive of the arts. One of the most studied theories of value in environmental psychology is the model proposed by Kaplan and Kaplan (1989, cited in Van den Berg, Vlek, & Coeterier, 1998) which reduces the value of the experience

of an environment to the four principles of mystery, complexity, legibility and coherence. In a study on the aesthetic value of landscapes, Van den Berg et al. (1998) used three of these four criteria – complexity, coherence and mystery – as predictors of beauty ratings. The closeness of this triad with Beardsley's theory of complexity, unity, and intensity is evident. Van den Berg and her colleagues found that these criteria were good predictors for beauty ratings; however, strong differences were found between different groups of users (farmers, residents and tourists) suggesting that aesthetic judgements were biased by usability implications – that is, thoughts (on a conscious or non-conscious level) concerning the purpose or possible use of the landscape.

These results hint at a further issue concerning the validity of value judgements: that of the value being assessed. It could be argued that in Van den Berg et al. (1998) different groups of people were in fact evaluating different values of the landscape (e.g., aesthetic the tourists, economic the farmers). If that is the case it is clear that different observers also used different criteria in their evaluation, for instance, fertility might have been an important criterion for farmers but not for tourists. The problem of value, not just in relation to musical performances but regarding artworks in general, has been lengthily discussed in philosophy of art, and one of the most authoritative accounts of it currently available is the one offered by Budd (1995). Budd stresses that in talking about the value of a work of art it is necessary to distinguish *which* value we are talking about. Each work of art can possess different kinds of values, like cognitive, social, educational, sentimental, religious, economic or therapeutic value. So for instance, Mozart's sonata K448 had been said to possess therapeutic value (Jenkins, 2001), Bruckner's symphonies or Mozart's *Requiem* may be religiously significant; Duchamp's urinal, Warhol's *Brillo boxes* as well as Cage's *4'33''* might be praised for their art-historical value, Beethoven's Eroica for its social value, and so on. All these kinds of value are *instrumental* values, in that they are determined by, or they lie in, the benefit (or harm) that the experience of the artwork offers to the perceiver.

There is, however, one type of value which is not instrumental and which is peculiar and distinctive of a work of art *as* work of art. This is the *artistic* or aesthetic value. Artistic value is not better or worse than any other values mentioned above, but it has a privileged position among them in that it responds to what (we assume) is

the main aim of an artist's activity: to produce an artwork – that is, to produce a product which possesses artistic value. Arguably, this value is what musicians would wish to be assessed for when performing. And similarly, when someone talks about the value of music performance without further explanations, it can be assumed that what it is meant is more often than not artistic value. Nonetheless, artistic (or aesthetic) value remains but one among several values which a performance may potentially possess and be evaluated for.

The distinction between different kinds of value, and the privileged role of the artistic value among them, might at first seem straightforward. However, this distinction is far from neat: different kinds of value can obviously coexist, and even overlap, and the question to what extent features that we would easily relate to the art-historical or cognitive sphere ought to enter our appreciation of the work as artwork is not easily answerable. For instance, one's appreciation for Leon Fleisher's performance of Bach's Cantata BWV 208 *Sheep may safely graze* in the album *Two Hands* may be enhanced by the knowledge that this recording signalled the pianist's return to performing after 35 years of fighting against *focal dystonia*, given the social and moral value that will add to the artistic one. It could be argued that should this performance be assessed in the context of a music competition, this enhanced appreciation linked to the social and moral components would be inappropriate. This statement however might not hold when substituting Fleisher's recording with Joyce Hatto's. Knowing that what is listened to is not the product of Joyce Hatto's performance, rather the result of a well done copy-and-paste engineering job will diminish (or nullify) the appreciation of the performance *as* Hatto's performance (Dutton, 2007). And in this case, again thinking of the context of a music competition, it could probably be argued that this contextual information is relevant to the evaluation of the musical product, and as such should be taken into account in the evaluation process.

The question to what extent contextual information – information other than the ones obtainable through sensual perception – should enter the appreciation of a work of art has long occupied philosophers of art in what has been known as the debate between *empiricism* (or formalism) – according to which any appreciation and understanding of a work of art should be based solely upon what it is possible to perceive directly during our encounter with the given object – and *intentionalism* (or

contextualism) – which claims that sometimes in order to appreciate and understand an artwork properly we need a certain amount of information external from what we can perceive through the pure experience of the artwork (Beardsley, 1988; Currie, 1989; Davies, 2006; Graham, 2006).

Back to the list of criteria summarised by McPherson and Schubert (2004), it appears that those parameters are tailored to the assessment of musical performances in an academic context. In a different context, as in the evaluation of a professional performance in a music festival, elements like posture or bodily coordination may receive only marginal attention, while within the evaluative domain of interpretation authenticity and musical coherence may be sided by, for instance, originality. As Gabrielsson (2003) highlights in his review of the state-of-the-art research in music psychology, empirical studies have focused so far mainly on the academic environment, thus it is not surprising to find academic-oriented criteria in McPherson and Schubert's (2004) model. And in the academic context the purpose of the evaluation may be to assess the student's achievement – and be able to offer a detailed feedback – more than judging the musical value of the performance. It could be asked however on the one hand, in what proportions musical and academic value should be assessed, and accordingly what relative weight the different criteria should have. On the other hand, if it is the academic achievement that is evaluated, to what extent contextual information relative to the academic history of the student should enter (or not) the assessment.

Performance as event

In the light of the studies and reflections mentioned so far, the process of music performance evaluation appears as a complex phenomenon, in which several factors linked to the different components of the performance can play a role. These studies offered evidence of what is a difficulty in delineating the object of assessment in music performance. To listen to a performance of Brahms's A minor Trio Op.114, unavoidably means to listen to that Trio performed by musicians A, B, and C in a given venue at a given time. The distinction between performer, performance and work being performed can only be partially done on a perceptual level, and features of the composition and of the performer will enter listeners' experience of the performance, interact with their beliefs, previous experiences, expectations and

prejudices and consequently colour their evaluation and appreciation of the music heard. The complexity underpinning this phenomenon poses the question what is meant by 'performance' when discussing 'performance evaluation'. The notion of performance that emerges from the discussion of the reviewed literature can be best understood appealing to Godlovitch's (1998) account of the nature of musical performances. According to Godlovitch, a performance is *an event*. As such, it may entail different elements, but there are four components that are always present and that are necessary for a performance to occur. These are:

(a) the sounds;
(b) the agent(s) – that is, those who produce the sounds;
(c) the work(s) being performed (in the tradition of Western classical music); and
(d) the audience, who are not passive receivers but rather active components of the event-performance, constitutive of it.

These elements move within a specific context, that is, in its minimal form, in a given venue and at a certain time.

The term 'performance' is often used also with a narrow meaning to indicate just one element of the event-performance: *the sounds* of it. To this shift towards a narrower, acousmatic notion of 'performance' at least partly contributed the enormous increase and dissemination of recorded music, which has offered an experience of music deprived from its agent and context (Clarke, 2007). In fact, as Clarke suggests, an attempt to develop awareness of the context in which the production of sound for a recording occurred might be not just difficult or merely impossible for the listener, but it may also become detrimental to the enjoyment of the listening experience. The use of 'performance' as pure sounds is widely spread in music parlance. But embracing the notion of performance-as-event allows a better appraisal of the complexity underpinning the processes of appreciation and evaluation.

A consequence of the notion of performance-as-event is that the construct of value of a musical performance as something fixed and objective whose perception can be nonetheless biased or obscured by different non-musical and extra-musical

factors does not hold any longer. In the light of the distinctions done up until now, it seems that talking of the value of a performance is of little use unless we specify in which context and for whom. This suggests in turn that there might be a number of different perspectives from which a musical performance can be legitimately evaluated. In Levinson's (1987) words:

> The question, "is performance P of work W a good one, and if so, how good?" can generally receive no *single* answer, but only a *series* of answers, for specifications of the question for various musically legitimate individuals, positions, contexts, and purposes (Levinson, 1987, pp. 87-88).

The variety of elements classified as 'extra-musical' in McPherson and Schubert's (2004) model illustrate the difficulty in determining where borders lie between different equally legitimate evaluations and non-valid assessments. The efforts done in music research to gain an understanding of the evaluation process of music performance have offered important insights and deepened the comprehension of the density of this process. There are nonetheless two issues that remain open and call for further investigation.

First, as mentioned, the majority of studies on performance evaluation run so far have focused on the academic environment, both in terms of musicians and listeners involved and of performance context and purpose (with consequent evaluation parameters). As Gabrielsson (2003) suggests, it might be time to free music performance evaluation studies from the academic context. Investigating evaluation of real world performances would, according to Gabrielsson, allow a narrower focus on aesthetic properties of the performance. On the other hand it may bring to the fore other, non-aesthetic properties that might have escaped examination so far.

Second, studies on the validity of aesthetic judgements mostly employed quantitative feedback, looking at the relationship between selected elements of the performance and listeners' (holistic or segmented) assessments of it. However, concerns on the strength of these assessments as a measure of listeners' evaluative opinion – as discussed in the first part of the chapter – compounded with the complexity highlighted by findings up until now questions the appropriateness of this form of feedback as a tool to gain a deeper understanding of the evaluation process.

Despite the indisputable advantage of quantifiable, comparable measurements and controlled design, it is possible that a more explorative approach will lead to novel insights, complementing and informing the findings of current literature.

One such approach – explorative in nature and which moves the investigation outside of the academic boundaries – is proposed in the third part of this chapter.

MUSIC CRITICISM

Seeking further understanding

Evaluation and appreciation of musical performances are essential components of music listening and permeate every aspect of our daily musical practice; however there is still only a partial understanding of those processes. The assumption that seems to permeate our musical practice, that the commonly agreed opinion of experts is a measure of the value of the performance, has not been confirmed through empirical research. Results of studies on inter- and intra-judge reliability displayed contradictory findings and often showed a low to moderate level of reliability even within expert evaluators, while studies on validity of judgements showed how the experience of an artwork is a cognitive act in which stimuli coming from the different elements that constitute the performance-event (sound or visual) are constantly set against thoughts of different sort which in turn colour and shape the experience of the performance.

Quantitative assessment versus Verbal feedback

One challenge to a deeper understanding of the process of performance evaluation is the kind and amount of information acquirable through listeners' feedback. Two of the main types of assessment used in empirical research are *holistic* and *segmented*, and both of them offer researchers very limited information regarding that to which the quantified evaluation refers. The segmented assessment is apparently more informative; however, studies have shown that this form of feedback does not enhance inter-rater reliability and there seems to be very limited scope for evaluators to differentiate between features of the performance being assessed. It is possible that this lack of differentiation is due to the nature of the assessment criteria, which are given top down by the researcher and may not reflect the quality indicators that the

single listener would actually use if let free to choose. Without further data, studies on reliability could be interpreted as a measure of the strength of the assessment scheme, of the homogeneity of the construct observed or of the evenness of participants' attitude towards this construct. Moreover, even if strength of the assessment scheme could be assured, the relationship between the final assessment and the evaluation process that led to it would still need clarification.

A different path to tackle this problem could be to require from listeners a feedback in form of unstructured text that describes and evaluates the performance. This would offer a far deeper level of information, as well as maintaining a high degree of ecological validity letting the single listener completely free to decide what to listen to and what to seek out for critical examination. This approach has been undertaken much less frequently in the research literature due to the methodological difficulties obviously implied in a qualitative inquiry and the limitations such material would offer in terms of comparability.

Mills (1991) moved a step towards a similar approach when she investigated the elements of a holistic assessment which can be verbalized. Mills interviewed eleven student teachers after they listened to five performances using a triangulation technique (Kelly, 1955). Choosing randomly three of the performances Mills asked students to describe one characteristic possessed by two of the performances but which was absent in the third. The principle behind this was that people might not be able to describe directly their own construct system, but once asked to describe another persons' they may unwarily seek out among the possible properties the one which corresponds to their personal system. Repeating this exercise with different groups of three performances and across the eleven teachers Mills was able to extract a set of twelve recurring bipolar constructs (see Table 1.3).

Mills asked two groups of listeners (i.e., music teachers and teachers of subjects other than music) to assess ten performances holistically and according to these 12 constructs. Applying regression analysis, she found that these constructs accounted for 70% of the overall mark on average and 73% for the music teachers alone. Mills concluded that there might be no advantage in using a segmented assessment scheme, since this does not appropriately mirror the process of forming a final judgement.

Table 1.3. Bipolar constructs extracted by Mills through triangulation procedure (Mills, 1991, p. 178)

C1	The performer was CONFIDENT/NERVOUS
C2	The performer DID ENJOY/DID NOT ENJOY playing
C3	The performer WAS FAMILIAR WITH/HARDLY KNEW the piece
C4	The performer MADE SENSE/DID NOT MAKE SENSE of the piece as a whole
C5	The performer's use of dynamics was APPROPRIATE/INAPPROPRIATE
C6	The performer's use of tempi was APPROPRIATE/INAPPROPRIATE
C7	The performer's use of phrasings was APPROPRIATE/INAPPROPRIATE
C8	The performer's technical problems were HARDLY NOTICEABLE/DISTRACTING
C9	The performance was FLUENT/HESITANT
C10	The performance was SENSITIVE/INSENSITIVE
C11	The performance was CLEAN/MUDDY
C12	I found this performance INTERESTING/DULL

Despite the negative conclusion, Mills's procedure points towards what may be a diverse approach to the performance evaluation process. Analysing the way listeners describe and compare performances might offer valuable insights on the processes of performance evaluation and appreciation and the way listeners make sense of them. One of the possible shortcomings in her study might have been the relatively small number of participants and performances (i.e., only eleven interviews comparing five performances). It is possible that the examination of a larger set of responses might have led to a more comprehensive picture. It is also important to recognize what an approach like this does not attempt to do: build, through the elicited responses, a construct system that may reflect the performance experience in a *comprehensive* way. Investigation of listeners' accounts of performances might offer awareness of critical aspects of the listening experience and its conceptualization, but the aim of such inquiry cannot be that of producing a set of all-encompassing evaluation criteria.

Verbalization of music perception

Contemplating the possibility of employing listeners' textual accounts to investigate the phenomenology of music, it is essential to ponder carefully the choice of potential participants. Speaking and writing about music can be a complex matter. As for other forms of perception, there are strong limitations to the average listener's

ability to describe music in words. Musical parlance relies widely on vocabulary drown from other semantic fields and applied to music by means of metaphorical and suggestive language. The extent to which listeners are actually able to use language to describe their musical experience accurately and efficiently is unclear, and recently concerns have also being raised about the impact that this kind of descriptive task may have on listeners' perceptions.

Schooler and Engstler-Schooler (1990) published a seminal paper that reported results of six experiments on face and colour recognition. Against the common assumption that verbal processing improves memory performance, these six experiments indicated that verbalizing visual stimuli impaired the subsequent recognition performance. They termed this phenomenon verbal overshadowing (VO). Since 1990, a large number of experiments have displayed VO effects in different fields such as recognition of wine taste (Melcher & Schooler, 1996), visual forms (Brandimonte, Schooler, & Gabbino, 1997) and fencing movements (Ait-Said, Maquestiaux, Didierjean, 2014). In these experiments the ability to recall and discriminate between different perceptual stimuli was weakened by the attempt to describe verbally the target experience. This impairment seemed to be common to diverse perceptual tasks, in particular to those tasks which are difficult to put into words (Schooler & Engstler-Schooler, 1990). The impairment however did not affect all participants. As Melcher and Schooler (1996) suggested based on their results, VO seems to occur only in those with a gap between their perceptual expertise and their ability to communicate verbally their experiences (Melcher & Schooler, 1996, p. 239), what they called the *clash* between verbal and perceptual expertise.

Verbal overshadowing has been shown to affect activities other than recognition. In a recent study on golf playing abilities, Flegal and Anderson (2008) found that asking highly skilled golfers to describe a golf-putting task temporarily affected their motor-skills, marring their ability to reproduce the task physically after having described it. This effect however was not found in low level players, hence confirming Melcher and Schooler's expertise *clash* hypothesis: those who have a low or high level of *both* practical and linguistic expertise are not affected by VO.

Wilson and Schooler (1991) tested the impact of verbal overshadowing on the quality of preferences and decisions. Forty-nine psychology undergraduates were asked to taste and rate five brands of strawberry jam. Subjects were randomly

assigned to one of two conditions: one group was instructed to write down "why you feel the way you do about each jam" (p. 183) before evaluating them, while the control group was not given any additional instruction. Participants' ratings of the jam were compared with the experts' ranking published in the *Consumer Reports* magazine. The comparison between subjects' and experts' rankings revealed a clear difference between evaluations given by participants in the reasoning and in the control group, with control subjects' preferences corresponding well with the experts' ranking, while preferences of those who were asked to think about their liking or disliking moving away from experts' choices.

In a second experiment (Wilson et al., 1993), forty-three female undergraduates were asked to evaluate five posters. Again subjects were assigned to two conditions: a reasoning condition in which they were asked to write why they liked or disliked each of the posters before evaluating them, and a control condition, where they were asked to fill a form with background information before handling back the questionnaire with the evaluations. At the end of the test, all participants were given the possibility to choose one of the posters, independently from how they had evaluated them, and bring it home. Wilson et al. found that being assigned to the reasoning or the control condition significantly influenced participants' evaluations and choice of the painting. A few months later, different researchers (unaware of the condition to which participants were assigned) called all participants by phone and investigated through a short survey how satisfied they were with the poster they had brought home. Interestingly, people assigned to the control condition displayed a greater satisfaction with their poster than people in the reasoning condition. Wilson et al. suggested that introspecting about one's own preferences, like when requested to give reasons for evaluative judgements, interferes with the way people feel about a given object. When asked to give reasons for their preferences, participants begun to focus on those features of the jam or the poster that it was possible for them to verbalize, thus temporarily changing their set of priorities and quality indicators for that object (Wilson & Schooler, 1991, p. 185). As for the recognition task, also in this case there is evidence that expertise mediates the overshadowing effect: Hodges and Wilson (1993) found that being knowledgeable about the task or attitude object moderates the effect of analysing reasons.

The effect of reasoning on music appreciation and assessment has not yet been explored. Mitchell and MacDonald (2011) verified the occurrence of VO in musicians' ability to recognize singers' voices. They found that verbalizing their perception of the singers' voice reduced participants' likelihood to recognize the target voice. From the analysis of musicians' voice descriptions, Mitchell and MacDonald also highlighted that listeners seemed to lack the necessary vocabulary to verbalize subtle timbral differences. They hypothesized that the inadequacy of vocabulary to capture impressions of singers' voices in turn limited listeners' memories of the unique sound they heard (Mitchell & MacDonald, 2011, p. 79). That is to say that the verbally ineffable component of the perceptual experience, that *something* that cannot be articulated through linguistic means, was lost through the attempt of conceptualization. Mitchell and MacDonald felt that these results have profound implications for the way we perceive, process, and describe listening experiences, and they called for further studies to investigate the extent to which different levels of perceptual and semantic knowledge may moderate this VO effect.

Focus on music critics

Studies on VO call for reflections on the way verbalization and reasoning tasks (which include also the use of segmented schemes for assessment) colour listeners' experience of the performance and impact their ability to assign valid and reliable evaluations. Musicians possess highly refined perceptual skills, trained and nourished often since early childhood, but they are often not equally trained in describing and explaining their perceptual experience through verbal means. They might therefore be particularly sensitive to the obscuring effects of reasoning when asked to count suddenly on their verbal, instead of perceptual, resources. Hence, an exploration of the process of performance evaluation that relies on participants' verbal accounts should seek out listeners who are highly knowledgeable about music, with a vast amount of musical exposure, and who also possess solid linguistic skills and have long-lasting experience in using those skills to describe their musical impressions.

A category of listeners who respond to all these requirements are music critics. Music critics, at least those with several years of experience and who regularly write for specialized music magazines, are usually musically competent, possess solid semantic and linguistic skills and are trained in using those skills to discuss music. They are seasoned listeners, the only category of listeners indeed which is regularly

paid to attend concerts or listen to recordings, and they are presumably not affected by predetermined quality indicators like those discussed in recent decades in the academic environment. Music critics therefore could potentially offer important insights into the process of listening to and appreciating music.

It was also mentioned earlier when discussing Mills (1991) that it may be important to have more than a few interviewed listeners and discussed performances to gain adequate insights from this material. In this regard, music critics seem to offer an appealing solution, that of investigating texts already written and published over several decades.

Music critics' writings represent a vast source of material and one which is highly ecologically valid. They also give the opportunity to explore in depth the phenomenon of music appreciation and description in a setting other than the academic one, thus answering Gabrielsson's call for a change of context in performance studies (Gabrielsson, 2003) that may enable an enhanced focus on aesthetic issues. The extent to which critics' writings can actually be used to explore the phenomenon of performance evaluation however depends upon its being engaged with the description and evaluation of musical performances. The following section will hence discuss the relationship between criticism and evaluation.

Criticism as evaluation

Music criticism can be broadly defined as "the intellectual activity of formulating judgements on the value and degree of excellence of individual works of music, or whole groups or genres" (Bujić, n.d., 'criticism of music', The Oxford Companion to Music). In this broad sense criticism can manifest in very different forms, from music teaching to conversation about music to diverse kinds of writings on music. However, in a narrower understanding, music criticism is seen as "a genre of professional writing, typically created for prompt publication, evaluating aspects of music and musical life" (Maus et al., n.d., 'criticism', Grove Music Online). Both these definitions expose what is the activity that characterises criticism and distinguishes it from other forms of musical parlance: evaluation.

Historical grounds

The understanding of art criticism as a form of reasoned evaluation is grounded in "long standing historical trends in the practice of criticism" (Carroll, 2009, p. 16). For instance, this was the assumption underpinning the two seminal essays on musical criticism by Calvocoressi (1923) and Newman (1925) and also, forty years later, Walker's *An Anatomy of Musical Criticism*, in which this idea is stated explicitly at the opening (1968, p. xi):

> The practice of criticism boils down to one thing: making value judgements. The theory of criticism, therefore, boils down to one thing also: explaining them. If you formulate a theory of criticism, it is not enough to know that one work is a masterpiece and another is a mediocrity. You must also explain why they are different.

Despite the different views on criticism portrayed in these works – the search for a rigorous theoretical basis for music criticism by Calvocoressi and Newman; the critique to the idea of 'objective' criticism by Walker – they all have in common the understanding of criticism as a practice characterised by the generation of value judgements supported by explanations of why the judgement be the case.

Reviewer versus Critic

More recently, this notion of criticism was defended and discussed by the composer, pianist and music theorist Edward Cone (1981). In his essay on the authority of music criticism, Cone discusses the difference between the figure of the reviewer, whose aim is that of guiding the reader's choice in terms of what to buy and listen to, and that of the critic, whose aim is to broaden and deepen the reader's appreciation of music. The reviewer can at times also be a critic and vice versa; this overlapping however does not abate the distinction. What is common to the activity of both figures is the evaluative component of their writings. In order to achieve their aim reviewers need to offer (a) a description of the musical product (how did the music sound?) and (b) a judgement of it (is it worth listening to it?). This is the basis also for the critic's job. In addition to this however the critic, drawing on her knowledge as well as historical, technical and experiential understanding of music, has to offer an interpretation of the musical work or performance and a reasoned evaluation of it, one that "draws correct inferences from verifiable facts" (p. 6). The difference

between reviewer and critic then – and the overlapping point between the two – is signposted through two facts: when the reviewer's description becomes interpretation of the performance and his or her judgement evolves into a reasoned evaluation of the artistic product, the reviewer becomes a critic.

Cone's terminological distinction between reviewer and critic will not be employed in this thesis, however, this differentiation is important in that it points to the fact that the term 'criticism' can be used to embrace a wide range of meanings and activities. Music 'critics' working for newspapers and magazines can be asked to cover different roles, from that of the proper critic, to the reviewer (in Cone's understanding of it), to the news reporter. And even though criticism, properly so called, can enter the activity of the reviewer, and even the reporter, this cannot be taken for granted. In any case, following Cone, the assessment of the artistic product is essential to both reviewing and criticism properly called, even though we might not expect to find the same intellectual depth and reasoning in both forms of writing.

Against evaluative criticism

Despite its solid grounding, the notion of criticism as evaluation found opposition among critics themselves in the last decades. As Elkins (2003) discusses in his book *What Happened to Art Criticism?*, the second half of the twentieth century witnessed a denaturalization of criticism into a non-judgemental, descriptive and evocative exercise. This change of focus in critical practice is, according to Elkins, "an amazing reversal, as astonishing as if physicists had declared they would no longer try to understand the universe, but just appreciate it" (p. 10) and this is what drew criticism ultimately to a "worldwide crisis" (p.1). Bearing evidence for this statement Elkins reports the results of a study conducted in 2002 by the National Arts Journalism Program at Columbia University (Szántó et al., 2002). In this study 160 visual arts critics recruited among 260 daily and weekly American newspapers and 9 newsmagazines were interviewed on different aspects of their practice. Asked to evaluate the amount of emphasis given in criticism to five different dimensions (describing, contextualizing, theorizing, evaluating, and creating a valuable piece of writing), only 27% affirmed to give a great deal of emphasis to the 'rendering of a personal judgement or opinion about the work being reviewed', thus making the evaluative component the least important of the five dimensions accounted for in the survey. A follow-up of this study run in 2005 (Conrad et al.) and focused on music

critics resulted in a less extreme picture; still, less than half of the 181 critics interviewed (45%) stated the rendering of judgement to be of great importance in reviewing. Evaluation was not the least important of the six dimensions accounted for in the survey, but still it was given a place clearly secondary to the aim of portraying the work or performance being reviewed as well as the critic's aural experience of it as vividly as possible.

In line with Elkins's claims, a loss of focus on judgement in critical practice seems to be attested by these results. However, following the path hinted at by Cone (1981), it may be instructive to look closer at the kind of activities in which the critics in these surveys were engaged. In the first study, when asked to report about what kind of writings they publish, only 16% of participants claimed to publish purely criticism (in the survey labelled 'evaluative review'). The residual 84% of critics admitted to regularly combine critical writing with some kind of reporting, and for more than one third of them critical writing represented less than half of their activity. This is a problematic issue related to the employment conditions of critics; as consequence of this, it could be argued that the seemingly low importance given by critics to the evaluative component of criticism might be linked to the nature of the writings they are used and required to deliver at their institutions. The picture is similar for the second survey, on music critics. Here 53% of the critics interviewed stated that more than half of their writings were not evaluative reviews, but rather diverse stories like "profiles of musicians, composers and musical figures" (p. 16). In line with these results 41% of critics defined themselves not as critic, rather as "arts reporter", "staff music writer who splits a part-time critic position with another beat", "program annotator", "general assignment critic", or "entertainment writer" (p. 12).

It is thus difficult to estimate the extent to which these survey results actually witness a substantial change in critical practice. In 2006, Rubinstein published a collection of essays on criticism by thirteen art critics and professors. Most of them explicitly discussed the importance of judgement in art criticism, defending the notion of criticism rooted in evaluation. Among the few writers maintaining the opposite position – that judgement is not an essential feature of art criticism – the most authoritative voice is that of the American art critic and philosopher Arthur Danto. Danto was a visual art critic, and his main argument is that evaluating works of art is an activity that has been removed from the responsibilities of the critic and

now lies in that of art curators in museums and galleries. In the selection of what works to exhibit lies the implied judgement: the curator, who decides about what to show for public appraisal and what not, is the one required to express a judgement on what are good works of art. The critic on the other hand, comes into play after this selection has been done, to describe, analyse and contextualise the art work, so to render it accessible to the audience.

Noël Carroll's account of evaluative criticism

An answer to Danto's argument as well as to other reasons commonly offered to reject the essentiality of evaluation for criticism is given by Carroll (2009) in what is the most influential among recent contributions to the philosophy of criticism, or meta-criticism. Carroll proposes that evaluation is essential for criticism in that it is what differentiates criticism from other forms of discourse about art, like art history or cultural studies. This idea is supported through a series of discussions and rejections of some arguments commonly raised against the view of criticism as evaluation grounded in reasons. To Danto's argument Carroll replies that there may be reasons other than the value of the artwork for a curator or gallery owner to select a given work, for instance "it may be that the art on display is work of questionable value by an influential artist or patron which must be shown for economic or even political reasons" (Carroll, 2009, p. 23). Moreover, continues Carroll, one of Danto's critical principles informed by his philosophy of art is that to qualify as such a work of art must always be about something, and it is up to the critic to determine – by means of analysis, contextualization, interpretation, etc. – to what extent the form of the artwork is suitable or appropriate to whatever the work is about. But determining the appropriateness of the form of the work to its content is an act of evaluation itself. And in fact, Carroll continues, evaluative terms can be found to "pepper" pieces of criticism by Danto or others who claim criticism is not about evaluation (Carroll, 2009, chapter 1).

As already discussed in the section on evaluation criteria, Carroll also addresses the problem of uniqueness of artworks defending the notion of objective criticism by means of stressing the importance of classifying artworks. It is part of the critic's job to contextualise the artwork and identify the category it belongs to, and upon this classification a reasoned judgement can be built.

Carroll's thesis, sharply defended in his book, is that criticism, properly so called, is essentially a matter of evaluation supported by reasons. Evaluation is not the only activity entailed in criticism, other activities like description, elucidation, classification, contextualization, interpretation and/or analysis are part of criticism as well (Carroll, 2009, pp. 13-14). However, among those activities evaluation is *primus inter pares*, in that all other activities are aimed at offering reasons for the evaluation. The distinction between the different activities should not suggest that evaluation must be stated explicitly and separated from description, elucidation, classification, etc. In fact, often evaluation is given implicitly in criticism, through the choice of value laden terms used for describing, elucidating, classifying, contextualising, interpreting and/or analysing the work.

Summary

In summary, the notion of criticism as reasoned evaluation is grounded in the history of critical practice, and it has been defended also recently by authoritative authors like the pianist, composer and music theorist Cone (1981), most of the critics and professors who contributed to Rubinstein (2006) and, from the perspective of philosophy of criticism, Carroll (2009). Reasons in support of value judgements in criticism are grounded in the activity of describing, classifying, contextualizing interpreting and analysing the work being reviewed. These activities are in turn informed by the historical, technical and experiential understanding of the critic. Discussing criticism as evaluation we need to be aware of the typology of activities that might be generally positioned under the 'criticism' label: a form of critical writing widely spread in the Western art tradition of the twentieth and twenty-first centuries is the art critical review, but it is an open question if this form of criticism currently reflects the canon of evaluative criticism grounded in reasons. However it might be expected that even where evaluation is not the aim of critical writing, an assessment of the object reviewed will probably emerge no matter what through the use of value laden terms in the description, contextualization, analysis and interpretation of the artistic product.

This notion of music criticism supports the hypothesis that critics' writings represent a rich source of insight on the way listeners, usually seasoned ones, make sense of their musical experience and reflect on evaluative issues related to music. But to what extent has music research tapped into this source of information so far?

Studies on criticism

Criticism of musical performances

A distinction is necessary at this point concerning criticism and the interest of this thesis. Critical writings can come in various forms and can be about different objects. A piece of criticism can discuss general aspects of musical life or stylistic tendencies in composition or performance practice (for instance Joachim Kaiser's "Music and Recordings", 1 June 1973, Süddeutsche Zeitung, in Haskell, 1996) or it can be about specific music genres or compositions (like Edward H. Krehbiel, "The Salome of Wilde and Strauss", 23 January 1907, New York Tribune, also in Haskell, 1996). Or again it can be criticism about criticism (meta-criticism) like Newman's and Walker's essays discussed above or like Henderson's "The function of music criticism" (1915). Despite the importance and richness of these kinds of criticism, the contribution these writings can offer to an investigation of music performance evaluation is often marginal. But there is a different type of criticism, which is of direct interest to this investigation, and this is what we might call *performance* criticism – that is, criticism of musical performances, either live or recorded ones, whose main focus is the *realisation* of the work being performed and not the work itself.

In the Western tradition of classical music, music criticism is a well-established practice, whose origins can be brought back to the late 17th century. Criticism of musical *performance* on the other hand, considered as a serious practice, is a quite recent affair (Monelle, 2002). Most stories of criticism wrote up to the beginning of twentieth century were instances of the first three types described above, with a large presence of stories about music compositions (often new ones) while criticism of performance was not seen as an important, nor valuable, activity. Still in 1915 the *Musical Quarterly* critic William J. Henderson affirmed:

> We are confronted by the demand of the interpretative artist. Of this any one who places the function of criticism upon a high plane would wish to say very little. The consideration of the performer is the least important office of real criticism (1915, cited in Monelle, 2002, p. 213).

Henderson could not imagine that this situation was going to change drastically in the following decades. The developments in the recording technology and the

decrease of performances of new compositions accompanied by the establishing of a canon of classical music repertoire played important roles in this change, contributing to the elevation of the figure of the performer from the status of executor to that of *interpreter*. Critics had suddenly fewer new compositions to discuss, but a new challenge with which to cope, that of discussing and comparing different interpretations of the same piece by different performers. Performance criticism spread and entered newspapers as well as specialist magazines, like *The Gramophone* (now *Gramophone*[1]), founded in 1923, bound to become one of the most authoritative voices for criticism of classical music performance in the twentieth century.

Music criticism in musicology and philosophy of art

Performance criticism is a phenomenon of the twentieth and twentieth-first century, and still a quite unexplored one. In fact, despite that criticism has been largely dealt with in musicology, these studies focused mainly on criticism from its origins to its flourish in 19th century. Inquiries may focus on a specific geographical area (McColl, 1996), repertoire (Cowart, 1981; Morrow, 1997; Wallace, 1986), institution (Ellis, 1995; Flynn, 1997; Morgan, 2010) or author (Reid, 1984). They discuss the institution of music criticism in its cultural and historical context and with a non-systematic approach, addressing a wide palette of themes emerging from critics' writings, like changes in musical taste and in the role of critics, the relationship between music, music criticism and society, and changes in the ways of listening to music (Morgan, 2010). Given the complexity and in some cases vastness of the relevant material, these studies are often focused on very short periods of time, like Morrow's analysis of German music criticism in the late eighteenth century or McColl's study of music criticism in Vienna between 1896 and 1897. A special case in this panorama is the ambitious anthological work by Haskell (1996), who under the title "Three Centuries of Music Criticism" collected and discussed a selection of 100 pieces of music criticism, each by a different critic from different regions worldwide, spread from the beginning of the 18th century up to the end of the 20th.

But the form of criticism taken by all these studies is almost purely criticism of musical compositions (or compositional genres, styles, tendencies) and meta-

[1] The name was changed in 1969 (Pollard, 1998, p. 109).

criticism. Two particular cases however should be mentioned. The first is Morrow's (1990) analysis of concerts criticism in the 19th century Vienna. Morrow focused on reviews of performances stressing how these might offer a different kind of insights than reviews of compositions based on written notation. At first, this might be thought to be a study on performance criticism, but it is not. The reviews analysed are reviews published in Vienna between 1800 and 1810. As such, even though they are in fact reviews of concerts, they do not offer much material concerned with performance and interpretation but rather focus on the work being performed. A different study is Morgan's (2010) investigation of texts published in *The Gramophone* between 1923 and 1931 and written by critics and readers who were members of the National Gramophonic Society. Through the analysis of these texts Morgan discusses how patterns of listening and thinking about music changed in response to the advent of recording technology and what function the first *The Gramophone* critics held in this process. In this study we find a first case of investigation of reviews of recorded performances, even though these were just a minority of the texts analysed. However, recording technology was at its initial stage and the change in focus from the work to the performance in criticism still had to occur. As Morgan states critiques of performances in the 1920s showed a lack of specificity and detail, thus appearing to today's reader as vague and unprecise. According to Morgan this may reflect partly the non-musical background of the founders and first critics of the magazine; on the other hand, this is also due to the still marginal concern for performative issues by listeners, who were mostly inclined to discuss the work performed and the quality of the recording.

Beside musicology, philosophy of art was long concerned with criticism and related topics. In recent decades analytic philosophers offered important contributions to the critical discourse by extensively discussing issues like the nature and localization of the value of works of art (Beardsley, 1965; Budd, 1995; Dickie, 2000; Levinson, 2004, 2009), the process of criticism and the importance of reasons for value judgements (Beardsley, 1982; Carroll, 2009; Hopkins, 2006), the existence and nature of principles of aesthetic value (Beardsley, 1962; 1968; Dickie, 1987; Levinson, 2002), the intersubjective validity of aesthetic value (Budd, 2007), the nature of aesthetic concepts (Aschenbrenner, 1981; Sibley, 1959), as well as specific issues related to the use of language by critics like the distinction between thin and

thick concepts (the firsts being purely evaluative, the seconds being descriptive concepts with an evaluative component, see Bonzon, 2009; Elstein & Hurka, 2009) and the use of metaphors (Grant, 2010). These papers discuss topics relevant to art criticism in general, and thus can be applied to inform any investigation of this practice. Some of them have been discussed in the previous section on the process of evaluation of musical performances. However, as appropriate to their philosophical nature, they do not offer nor look (systematically) into real world examples of criticism.

Sociology and cultural studies on music criticism

Musicology and philosophy of art are not the only disciplines that dealt with the phenomenon of music criticism. Recently sociology and cultural studies have turned to the critical practice with increasing interest, in particular recognizing criticism the role of gatekeeper of taste (Schmutz, Van Venrooij, Janssen, & Verboord, 2010, p. 501), able to offer legitimation to a cultural institution giving it the status of Art. Baumann (2001) argued that American critics offered a legitimating ideology for Hollywood movies to be acknowledged as art form, and in music the same is claimed to have happened with jazz (Lopes, 2002, cited in Schmutz et al., 2010) and rock (Regev, 1994). The rising interest in criticism from the side of sociology and cultural studies brought to some first systematic explorations of large set of critical writings. Schmutz et al. (2010) investigated changes in newspapers coverage of popular music in the second half of the twentieth century in four geographical areas: the United States, Germany, France and the Netherlands. They analysed 1,867 music related articles published in eight widely circulated newspapers (two per country) in 1955, 1975, 1999, and 2005. Articles were coded according to style (classical or popular music), type (review, interview, news item, preview, announcement, background, opinion, or regular column) and genre (jazz, rock, R&B, country, world music, etc.) and measured in square centimetres. Descriptive statistics calculated for the dataset showed a rising prominence of popular music across decades in all four countries. This seems to point to an increasing legitimacy of popular music (p. 505) and it is accompanied by a shift toward an evaluative and properly critical approach to the emerging art form: while in 1955 news items and announcements were the commonest type of article on popular music, by 1995 reviews covered the highest amount of space in newspapers in all countries except France. Differences were

found between countries with Germany showing a less open attitude toward popular music and the USA and Netherlands being the most open to it.

Also in 2010 the same researchers published results of another study that aimed to compare aesthetic criteria used in popular and classical music criticism (Van Venrooij & Schmutz, 2010). Here, the source material was notably restricted and entailed 122 reviews of albums of popular music published between October 2004 and March 2005 in six newspapers in the three countries: the United States, Germany and the Netherlands. This narrower focus allowed researchers to run analyses on the text content level. Researchers read the reviews and assessed the presence or absence of *popular* and *high art* aesthetic criteria. The choice of what constitute high art aesthetic criteria and popular ones was done top down drawing from diverse literature. Indicators of criteria typical for high art were said to be: discussion of context, talking of performer as creative source, associations and comparisons with high art (recognized masterworks), and proper high art criteria such as originality, complexity, seriousness and timelessness. Indicators of popular aesthetics were: negative stance to high criteria, use of terms that point to participatory experience (e.g., rousing, irresistible, catchy), user orientation (i.e., suggesting for which audience the album could be good), and use of language related to "primary" tastes, like oral and food-related metaphors. These indicators were used as dichotomous variables and ordinary least squares (OLS) regression was applied on the coded reviews to test for differences between countries in the use of the two types of indicators. Review length (in words) was also accounted for during the analysis. Results showed distinctive patterns for the three countries. High art criteria were significantly more present in German reviews than in the American and Dutch reviews, while popular criteria were particularly high in the American and Dutch texts and almost absent in the German reviews.

While the first study offers a glimpse of the extent, form and subject of music newspaper coverage, the second analysis enters the evaluative domain. Its aim though was not to explore expert critics' construction of value judgements but rather to compare the extent to which predetermined topics – understood as a signal of the legitimization of a cultural institution as art form – are discussed in popular music reviewing. Moreover, the focus on popular repertoire implies engaging with a musical style in which the notion of work and that of performance are not as separate

as they are in classical music, and where the construct of interpretation plays therefore a different and arguably marginal role. For these two reasons the relevance of this study for an exploration of the performance evaluation process is limited.

Economics of information on music criticism

If we accept that a main purpose of music reviews is to lead readers' choices regarding what to buy and what to listen to (Cone, 1981; see also Frith, 2009), then it is reasonable to expect that analysis of reviews is of interest in the field of economics of information as well. Mudambi and Schuff (2010) ran a study on customer reviews taken from Amazon.com in an attempt to identify what features make a review helpful for readers. Drawing on Nelson's (1970) model, the authors distinguished between search goods and experience goods, where search goods are items whose "key attributes are objective and easy to compare, and there is no strong need to use one's senses to evaluate quality" and experience goods are those whose "key attributes are subjective and difficult to compare, and there is a need to use one's senses to evaluate quality" (Mudambi & Schuff, 2010, p. 191). They chose three experience and three search goods on sale in Amazon.com and analysed 1,608 reviews concerning those six products. Music is a paradigmatic example of experience goods and therefore they included a music recording (Compact Disc "Loose" by Nelly Furtado) within the first group.

They analysed two explanatory variables: extremity (positive/negative rating given in number of stars) and depth (measured as number of review words). The dependent variable was the helpfulness of the review, measured as percentage of people who answered 'yes' at the question 'Was this review helpful?' provided by Amazon on its website. Their findings suggest that for experience goods like music recordings extreme reviews – that is, reviews which entail a clearly negative or positive evaluation – are perceived as less helpful than moderate reviews. This result could be explained following the paradigm of experience and search goods: extreme ratings for experience goods may easily be seen as reflecting a sheer matter of taste, while a moderate but well articulated review may be perceived as more credible. Also, review depth affected helpfulness ratings, so that longer reviews were perceived as more helpful than shorter ones. This effect however was stronger for search than for experience goods, in line with the authors' hypothesis that in reviews of search goods, which are more fact-based, additional text "is more likely to contain

important information about how the product is used and how it compares to alternatives" (Mudambi & Schuff, 2010, p. 190).

This study presents several limitations for its application to the present discourse. First, reviews under scrutiny criticise a recording of popular music repertoire, which, as already mentioned, reflects a model of musical performance in which the musical work being performed is more intimately entangled with the performance itself than in classical repertoire. Second, the music recording was only one of three experience goods under scrutiny, the other two being an MP3 player and a video game. Third, reviews were customer reviews published online and not the output of the professional activity of seasoned critics. Last, the constructs of reviews extremity and depth were investigated using the quantitative surrogates of star rating and number of words. This approach had the advantage of assuring a higher level of objectivity, but as the authors suggest, further study could employ qualitative text analysis to explore these constructs more comprehensively. Beside all these limitations, however, the results support the thesis that, for music criticism to offer useful judgements of value, it needs to be grounded on valid reasons – that is, reasons that refer to properties intrinsic in the object criticised.

AIM OF THIS THESIS

Given the state of research discussed so far, it is now possible to delineate the main objectives of this thesis.

As has been shown, a new approach to the investigation of the phenomena of music performance evaluation and appreciation may need to involve the analysis of textual descriptions of musical experiences by expert listeners who also have sound linguistic skills and experience in verbalizing their perception and appreciation of music.

Criticism of musical performances published during the course of the past century – that is, since the legitimation of the performer as interpreter and the establishment of a canon repertoire led critics to abandon the discussion of the work being performed to focus on how the performance was realized – represents an extended heritage of such descriptions.

This corpus of texts on music could offer a fertile and still unexplored terrain of enquiry. In fact, despite the attention it has attracted in recent years in different

disciplines, there is currently no investigation of music performance criticism that systematically explores the way seasoned critics make sense of their experience of performances and how they structure and verbalize their evaluations. The aim here is to contribute to filling this gap by offering an investigation of value judgements in criticism of performance.

Insights from this investigation are expected to offer first empirical evidence on the content of this very common form of written response to music, thus adding to our understanding of expert performance evaluation.

As previously discussed, most literature on performance evaluation so far has focused on music assessment in the education (often higher education) environment. This form of assessment differs from evaluation in critical review in its object of assessment (professional versus student performance) audience and purpose (guidance to consumers versus a pedagogically valuable feedback for students). Nonetheless, the two forms of response to music share what is arguably the core aspect of any evaluation of works of art: the focus on the artistic value and the aesthetic experience.[2] As Gabrielsson (2003) suggests, this focus is more prominent outside the educational setting.

In this difference thus resides the potential value of an investigation of expert critical review. Even if not directly translatable to the educational context, insights from an investigation of critics' judgements will offer the performance evaluation discourse a fresh perspective and new input for reflection, informing the development of assessment protocols with evidence on what expert audiences find pleasurable and what concepts and vocabulary expert listeners employ in the conceptualization and assessment of performance.

The main open issue regarding the validity of evaluations of musical performances concerns the criteria to be used; in particular, the possibility of having

[2] Of course, the standard by which the artistic value is judged depends on the evaluation context. Differences between assessment in school exams or competitions and in critical practice are in this perspective however not neat, but rather reflect different points on a continuum that moves from beginner level up to higher education students, new and seasoned professionals. Critical practice deals with aesthetic evaluation at the far right side of this continuum, hence making an understanding of the criteria involved in this practice most useful for institutions and pedagogues engaged in supporting young musicians to progress towards this end.

criteria that are shared between people. Therefore, the aim of this thesis is to examine the nature of judgements given by critics in their reviews of performances, focusing on what properties of the performance are discussed, how these properties are used to build value judgements, and to what extent underlying evaluation criteria are common to different critics. The main research question is:

> What reasons do critics adduce to support their evaluative judgements of recorded performance?

This question is operationalised in the following sub-questions:

1. What do critics write about when reviewing a recorded performance?
2. How are the diverse elements discussed used to build value judgements?
3. To what extent is the emergent model shared between different critics?

In order to explore these questions, a series of data reduction and inductive thematic analyses has been undertaken. Prior to this, in Chapter 2, methodological perspectives are discussed concerning the selection of material and the procedure of investigation. In particular, a review of the literature dealing with texts through quantitative and qualitative approaches is offered; the resulting considerations have informed the design of the subsequently reported analyses.

2 METHODOLOGICAL CONSIDERATIONS

This chapter is structured in two parts. In the first, different approaches to the analysis of texts are examined, and an overview is offered of what tools can be used to extrapolate information from unstructured texts. Drawing from this review, as well as from considerations raised in Chapter 1, the second part presents the material of investigation chosen for the present research and the methods applied to its examination.

DEALING WITH UNSTRUCTURED TEXTS

The analysis of open (or unstructured) text poses substantial methodological challenges for the researcher. A broad distinction in the approach to text analysis is between positivist- and interpretive-oriented methodologies, the former striving for robustness and reliability of results and the latter focused on achieving a deep level of analysis and understanding of the material for which the interpretive contribution of the researcher is essential. Both positivist and interpretive approaches can be data-driven (inductive) or theory-driven (deductive), thus supporting both explorative and confirmative analysis purposes (Guest, MacQueen, & Namey, 2012 chapter 1).

The following sections outline the main methods currently in use in both positivist and interpretive approaches and broadly discuss their major limitations.

Positivist approach

The positivist approach aims to maximize efficiency in terms of manageability of analysis tasks, as well as reliability (i.e., the probability that different researchers repeating the analysis on the same data will obtain the same results) and robustness (i.e., the probability that repeating the analysis with a different sample of material from the same population will lead to the same results) of the investigation. The past few decades have seen the flourishing of a multitude of computerized or computer-assisted methods for the analysis of unstructured texts that tried to answer these needs. The use of terminology in distinguishing between methods is not always clear,

and the same terms can be used with different meanings. In this thesis, the use of relevant terminology is mainly based on the recent work by Feldman and Sanger (2007) and Guest and colleagues (Guest et al., 2012; Namey, Guest, Thairu, & Johnson, 2008).

Drawing from this literature, a main distinction can be made between methods that derive information from the 'raw' data, i.e., words (*content analysis*), and those that use an intermediate document, i.e., a pre-processed version of the texts, to run the analyses (*text mining*).

Content analysis

Content analysis entails the exploration of a large set of text documents by means of interpreting the frequency and salience of words or expressions (Namey et al., 2008). The simplest form of content analysis is the extraction of selected words or word combinations: based on the frequency rate with which those terms and expressions appear, the salience of key ideas or the presence of recurring concepts can be evaluated.

Content analysis can be used to explore not just what people write or talk about, but also *how* they use language to talk about it. Different use of linguistic structures and everyday words – like pronouns, articles or punctuation – or the emotional valence of the vocabulary used can offer insight that go beyond the object described and enter the domain of the author's personality and psychological state. In psychology, the observation of word use has been employed in the last decades to assess personality dimensions and distinguish mental disorders (Tausczik & Pennebaker, 2010). To facilitate this approach, recently Pennebaker, Booth and Francis developed simple software for content analysis that counts words in psychology relevant categories across different text documents (Linguistic Inquiry and Word Count (LIWC), see Tausczik & Pennebaker, 2010).

Content analysis offers two main advantages to researchers dealing with qualitative data: it assures reliability by working quantitatively with the raw data, and it is efficient, in that it makes the exploration of large sets of textual data feasible. Its main disadvantage, however, resides in its powerlessness to account for the context-dependency of terms. In content analysis, words and combinations thereof are examined in isolation; as a consequence, it is not possible to disambiguate between different meanings or syntactic roles one and the same word can have, or to account

for the use of irony, metaphors or idioms. As Tausczik and Pennebaker explain in discussing the potential of LIWC software, despite the fact that "studies are providing evidence that function words can detect emotional and biological states, status, honesty, and a host of individual differences… the imprecise measurement of word meaning and psychological states themselves should give pause to anyone who relies too heavily on accurately detecting people's true selves through their use of words" (Tausczik & Pennebaker, 2010, p. 30).

A way to limit this problem is offered by recently developed software packages (like DQA with WordStat) that permit the extraction of words or expressions together with a certain number of words that come prior to or after the searched terms (*keyword in context* report, Namey et al., 2008, p. 143). This form of reporting, however, calls for the researcher to read through the extracted passages and take decisions to disambiguate the meaning of keywords, thus partially marring the efficiency and reliability advantages of these methods.

Text mining

Text mining is a recently developed area of research in computer science and machine learning that is informed by a variety of fields of study such as information retrieval, natural language processing, and data mining. Text mining refers to the extraction of insightful patterns of information from text documents aimed at the discovery of knowledge relative to those documents (Tan, 1999).

This extraction of patterns of information cannot occur on the raw data directly. Hence in text mining – differently from content analysis – the original unstructured documents need to be pre-processed into structured patterns of data on which the software can run the analyses. Pre-processing may include operations like tokenization and zoning (i.e., partition of the text), deletion of stop words (e.g., function words), stemming, term extraction and labelling, and part-of-speech tag. Through this process the text is reduced to a set of canonical elements familiar to the software. On these elements data mining procedures are applied. The general process of text mining can be summarised in the following scheme (adapted from Feldman & Sanger, 2007, p. 15):

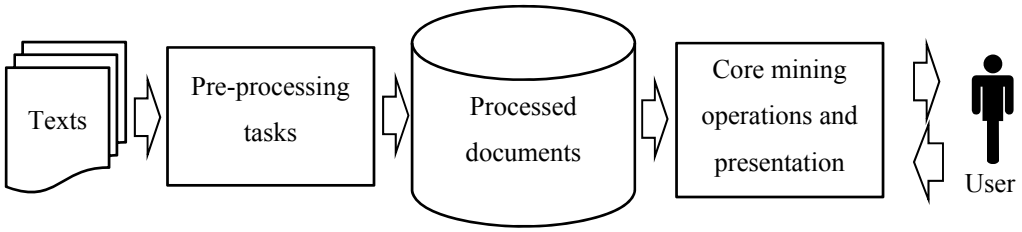


Figure 2.1. Text mining process, adapted from Feldman and Sanger (2007).

The double arrow between the core mining operations and the user indicates the iterative nature of the process. Visualised patterns of data following the mining operations should in fact elicit a series of queries and reflections that inform a next set of (refined) analyses in what should be seen as an interactive human-machine loop (Feldman & Sanger, 2007, p. 13). Core mining operations can come in various forms, but they can be summarized in three major types: *categorization*, *clustering* and *information extraction*.

Categorization

Given a set of categories and a number of documents, categorization is the task of attributing each document to its correct category. Categories (or topics, or themes) can be either given by the researcher (*knowledge engineering approach*) or derived through inductive process from a set of pre-classified documents (*machine learning approach*) (Feldman & Sanger, 2007, p. 64). An example of automated method for text categorization applying machine learning approach can be found in Hopkins and King (2010). In this method, in a first stage, texts are pre-processed by converting to lowercase, removing punctuation, and stemming. After this, a 'bag of words' procedure is applied: for each document, each stem is translated into a dichotomous variable, with 0 indicating that the stem does not appear in the document, and 1 that it appears in the document. Stems occurring in less than 1% or more than 99% of documents are eliminated.

Two samples of text documents are provided to the software: a smaller set previously hand-coded by researchers and a larger set for which an estimate of the categories is sought. Comparing the patterns of dichotomous variables between the two sets, the software automatically assigns each un-coded document to one

category. Next, misclassification probability is computed by splitting the hand-coded set in two, utilizing the first half to estimate the categories of the second half, and then comparing estimates with real data. The results of this check for misclassification are applied to correct the raw estimate of the whole dataset. According to Hopkins and King (2010) between 100 and 500 documents hand-coded by the researchers allow the software to produce estimates of the categorization of residual documents with a root mean square error between 3 and 1.5 percentage points.

Categorizing texts through text mining procedure does not offer new insights of the text content, in that it is up to the researcher to decide beforehand, by reading and studying a sample of texts, what are the categories and themes to be sorted out. Computerized categorization procedure nonetheless represents a useful tool when the set of documents is too large for researchers to code manually, and researchers need an estimate of how many documents in the set fit in each category or theme.

Clustering

Clustering is the process of partitioning a set of documents into groups. The main difference between categorization and clustering is that clustering is an unsupervised process – that is, the labelling of objects into clusters occurs without any prior information given by the researcher (i.e., pre-determined classifier or labelled documents). The assumption underpinning cluster procedures (*cluster hypothesis*) is that "relevant documents will be more similar to each other than nonrelevant ones" (Feldman & Sanger, 2007, p. 82). Hence, clusters production is based on the comparison of patterns similarity between documents.

To be useful, clusters need to maximize similarity within and difference between groups. To this purpose, documents are first reduced to numerical strings, usually employing 'bag of words' procedures as the one seen in Hopkins and King (2010). In a second step, likeness between strings is computed by means of similarity matrices and used to organize documents. The unsupervised nature of clustering processes has two main advantages over categorization methods: first, it does not require from researchers to create a classifier, that is, to decide prior to the analysis which categories to apply nor to engage in a time consuming hand-coding process;

second, it allows for unexpected and potentially insightful categories (and thus concepts) to emerge (Grimmer & King, 2011; Janasik, Honkela, & Bruun, 2008).

One issue with clustering methods is that there are a large number of different algorithms that can be applied to cluster a set of documents, and it is difficult, or impossible, to predict which of these unsupervised approaches will lead to an insightful classification of the texts (Grimmer & King, 2011, p. 1). In answer to this problem, Grimmer and King (2011) developed a computer-assisted method for clustering and conceptualization of documents (CAC). The method consists of running on pre-processed documents a long series of different clustering algorithms and using a visualization tool to allow the researcher to explore and choose between different resulting clusters. The method is still in a validation stage and software to implement it is currently being developed at the Institute for Quantitative Social Science at Harvard University. Tests to assess the quality of clusters developed through this method run on diverse sets of documents (e.g., press releases by Senator Frank Lautenberg's Senate office) show promising results, with clusters produced through the CAC method displaying higher quality than clusters produced by human coders. In these tests however, cluster quality was defined as the average similarity of pairs of documents from the same cluster minus the average similarity of pairs of documents from different clusters (Grimmer & King, 2011, p. 5). It remains to be seen how this notion of quality, based on similarity of numerical strings corresponding to word-stems, reflects a conceptually satisfying differentiation between constructs. More interestingly is the discovery reported by Grimmer and King of an unexpected category through the employment of CAC method, with consequent gain of novel insights.

Information extraction

Besides categorizing and clustering, text mining is used to extract different kinds of information from texts. In information extraction (IE), the application analyses texts *semantically* and offers users the specific information in which they are interested. IE can thus be seen as "a limited form of 'complete text comprehension'" based on natural language processing (Feldman & Sanger, 2007, p. 95).

In order to 'understand' the text semantically, IE applications perform a series of tagging aimed at identifying *entities*, *relationships* and *events*. The process of IE

can be summarised into five steps, each of which brings the analysis to a deeper level of text understanding (Cunningham, 2005):

Table 2.1. Five-task model of IE, adapted from Cunningham (2005).

1) Named Entity Recognition (NE)

Finds and classifies names, places and other entities

2) Coreference Resolution (CO)

Identifies relationships between entities

3) Template Element Construction (TE)

Adds descriptive information to the NE results (using CO)

4) Template Relation Construction (TR)

Finds relations between TE entities

5) Scenario Template Production (ST)

Fits TE and TR results into specified events scenarios

An example: taken the sentence

> "Its whole expanse was covered with tall, juicy grass, and when the wind blew, great waves passed over it with a sound like troubled water" (Michael Ende, The Neverending Story, The Grassy Ocean)

NE then identifies the entities in the sentence, namely *expanse*, *grass, wind, waves*, *sound, water*. CO finds out that *it* in the second statement refers to the expanse. TE determines that the grass is tall and juicy, the waves great and the water troubled. TR discovers that the grass constitutes the whole expanse, that the waves went over the expanse and that the sound of them was like troubled water. ST finally establishes that there was a wind blowing event in which the diverse entities were involved.

Not every IE application goes through all five steps. The simplest form of IE is terms extraction, which involves only the identification of entities in the text (Feldman & Sanger, 2007, p. 95).

One example of IE process applied to a study of online reviews of music recordings was developed by Hu, Downie, West, and Ehmann (2005); this method,

however, was limited to the extraction of metadata relative to the genre of the music piece reviewed and the valence (positive or negative) of the review. A more in-depth analysis can be found in a different IE application by Hu and Liu (2004) relative to the extraction of consumers' opinion related to specific features of products in online reviews. Hu and Liu developed a method that first extracts substantives relative to features of the object reviewed. In a second step, the algorithm identifies attributes of the features (namely adjectives found in proximity of the feature name) and based on a thesaurus, classifies each adjective as either positive or negative. Finally, the application provides the user with a summary of the features and relative opinion orientation. That means, for instance, that analysing reviews of a digital camera the outcome of the analysis will be of the kind: picture quality [253 positive; 6 negative]; size [134 positive; 10 negative]; etc. (Hu & Liu, 2004, p. 755).

As the authors explain, this method is limited to the recognition of properties of the object explicitly named in the text. The identification of features discussed in a non-explicit way – like size of the camera in the sentence *"while light, it will not easily fit in pockets"* (p. 757) – would require a far more sophisticated semantic understanding than the one provided by the software. Also, the application identifies features as entities, hence expressed through nouns (like size or picture quality); properties described by means of verbs are thus not included in the analysis either.

Limitations of the positivist approach

In summary, content analysis and text mining applications offer useful tools to explore large sets of textual documents quickly and reliably. Additionally, allowing for the investigation of very large samples, these methods increase the robustness of the analyses. Content analysis offers a first exploration of texts based on the search for keywords and computation of frequency rates. This approach, however, cannot account for the context-dependency of words, thus can only be seen as a first, complementary step to a further, closer analysis.

Categorization algorithms offer potent tools to facilitate, for instance, indexing and classification tasks such those necessary in libraries and archives or in other contexts requiring the storage of large amount of textual documents. In research they can also be used for confirmatory purpose, to test for instance the percentage of documents belonging to a given category within a set. Categorization methods, however, do not suit inductive analysis, since they work on a predefined classifier.

Besides, these methods assume the existence of categories that are comprehensive and mutually exclusive, so that each document can belong to one and only one category. They do not apply to scenarios that involve more than a few categories or a level of complexity that allows for category overlapping or unclarity.

Clustering and information extraction can go beyond that, offering novel information or summarization of the content of the texts. The fact that clustering works with texts transformed in numerical strings poses the question of how large is the portion of information lost in this process. The semantic approach of information extraction seems to be more ecologically valid. As Feldman and Sanger (2007, p. 94) point out, however, these applications are only useful when three conditions are satisfied:

- The information to be extracted is specified explicitly and no further inference is needed;
- A small number of templates are sufficient to summarize the relevant part of the document;
- The needed information is expressed relatively locally in the text.

Even as advanced applications as the one developed by Hu and Liu (2004) can only extract from text information relative to entities that are explicitly stated. It is telling to see that in testing the strength of their method, Hu and Liu only analysed online reviews of what Nelson (1970, discussed in Chapter 1) would call search goods – that is, items that can be described in terms of objective features for which a common understanding can be assumed (e.g., digital camera). The extent to which similar methods can be applied to investigate descriptions of experience goods like music, which often rely on the use of similes and metaphors, remains to be seen.

Interpretive approach

An answer to the need for a more comprehensive, detailed, and insightful analysis of textual data is found in the interpretive approach, often identified as *qualitative research*. The adjective 'qualitative' fundamentally refers to the type of data and analyses used. Qualitative research employs data that are not represented by quantified values, like texts, images and sounds. These data are not analysed by

means of mathematical and statistic tools rather through systematization, categorization and interpretation (Janasik et al., 2008, p. 440).

Since the development of Grounded Theory in the 1960s (Charmaz, 1995), several different analytical methods have been developed, especially within the domain of psychology and other social sciences. Among them are discourse analysis, conversation analysis, phenomenology, focus group, narrative psychology, co-operative inquiry and interpretive phenomenological analysis (for a discussion of all these methods see Smith, 2008). Common to all these methods is the reading (and re-reading) of the texts under scrutiny with the aim of identifying themes which are then signposted through codes. Themes are later organized and grouped into superordinate concepts.

When it comes to decide which themes to prioritize, the strength of themes is assessed by the researcher not just in terms of frequency of occurrence but also accounting for the salience of the concept as it emerges from the context of the discourse or the language chosen by the participant to express that idea (Smith, 2008, pp. 74-75). Another characteristic of interpretive approaches is their iterative nature. While proceeding in the analysis, the researcher constantly turns back to the already analysed texts to use the newly emerged insights to refine and correct the existing codes. This is necessary to assure that the development of themes remains congruent with the raw data – that is, the texts at hand.

Limitations of the interpretive approach

Qualitative analysis of texts renders it possible to move the investigation beyond the explicit information present in the text to capture context-related meanings of words and sentences and to account for the use of idioms, metaphors and other rhetoric figures as well as irony or sarcasm. Moreover, since no reduction of data into any canonical form is required, in qualitative analysis there is no loss of information involved.

On the other hand, qualitative research is dependent on the interpretation of the researcher; consequently, potential lack of reliability presents a major concern for this approach. It also requires from the researcher a notable investment of resources in terms of time as well as personal knowledge and interpretive acuteness and tact.

Because of its time-consuming nature, qualitative text analysis tends to have an idiographic focus, thus being unsuitable to investigate large datasets: methodologies

like interpretive phenomenological analysis for instance usually work with samples between three and fifteen interviews, and recently the founder of this methodological approach also made the case for the single case study (Smith, 2008 chapter 4).

The reliability concern is also crucial to the choice of pledging oneself to the interpretive approach. Quoting Geertz (in Guest et al., 2012, p. 14): "to commit oneself to a semiotic concept of culture and an interpretive approach to the study of it is to commit oneself to a view of ethnographic assertion as … 'essentially contestable'". As Janasik et al. (2008) emphasise, the analytic outcome of an interpretative analysis can be subject to two kinds of biases: the first related to the researcher's own beliefs, conceptions, way of thinking, knowledge, etc.; the second due to the human tendency to "think in terms of pure categories, or ideal types … and assume that such categories reflect the organization of reality" (Janasik et al., 2008, p. 440). Regarding this last reflection, however, it could be argued that mathematical and statistics methods, which work by reducing complex sets of data into simple, quantifiable models, share the same weakness.

Reliability and validity remain nonetheless important problems in interpretive research. An attempt to offer a solution to this problem comes from grounded theory, which suggests a method for a transparent, systematic analysis constituted by a series of inductive and iterative techniques aimed at the development of theoretical models (Charmaz, 1995). The basic process of grounded theory is very much similar to that mentioned above: (1) reading texts; (2) identifying possible themes; (3) comparing and contrasting themes in order to structure them; and (4) building theoretical models (Guest et al., 2012, p. 12).

The attempt to give validity to the process is mainly achieved through three procedures:

- The analysis should be purely data-driven; for this reason, a literature review should be postponed (Charmaz, 1995);
- The coding of text should be done through a line-by-line reading and coding and a systematic comparison and contrast of each statement with the previous ones (Guest et al., 2012, p. 28);
- The amount of texts to be coded should not be decided *a priori*: the analysis (and data collection) should continue until the saturation point is reached, that

is, until the analysis of further documents no longer add any new themes or insights to the already emerged model.

Adherence to these conditions should provide an analysis which is soundly grounded in the data at hand, so as to assure reliability, and which involves a set of documents as large as it is needed to guarantee robustness and allow for generalizability of results. But as Guest et al. (2012) point out, observing these conditions is in practice extremely demanding and rarely these requirements are fully satisfied:

> Many people claim to be using a grounded theory approach in their analysis, but fully developed applications of the approach are relatively rare. Grounded theory requires a painstaking, line-by-line reading of qualitative data where each statement is systematically compared and contrasted. Most people who claim to use a grounded theory approach do not analyze the data at this level of detail because they lack the time and resources or they lack data of sufficient richness to warrant such a detailed level of analysis—or both (Guest et al., 2012, p. 28).

Hybrid approach: Applied Thematic Analysis

The interpretive approach seems to offer the right tool to investigate the complexity and richness of textual data allowing for a comprehensive, detailed, context-aware analysis of texts. On the other hand, it does not satisfy the requirements of generalizability and repeatability typical of scientific research. Attempts to distance the researcher from the object of analysis through the use of quantitative methods does not offer a valid solution either, in that the results so obtained are reliable and robust but risk being more 'correct' than interesting. The duality between positivist and interpretive approaches recently generated broad discussions on fundamental epistemological issues and in turn gave rise to a series of mixed-approaches, aimed at reconciling qualitative and quantitative methodologies finding linkages between the two and offering a perspective in which these approaches are seen as complementary more than opposed (Janasik et al., 2008).

A recent contribution to the discourse is offered by the anthropologist and social-behavioural scientist Greg Guest and colleagues (2012) in the form of a practice-based, positivist/interpretive approach to qualitative research: Applied

Thematic Analysis (ATA). At the core of this methodological perspective is the conviction that analysis processes should be driven case-by-case from the data at hand, without excluding in advance any theoretical or epistemological approach. Guest *et al.* (2012, p. 4) reject "a compartmentalized view of qualitative research and data analysis" and suggest "a type of inductive analysis of qualitative data that can involve multiple analytic techniques".

Central to this approach is the qualitative text analysis carried out by the researcher. As about what analytical stance should be taken in this process, Guest *et al.* (2012, p. 12) state that ATA is linked to grounded theory in its emphasis on supporting claims by means of evidence grounded in data. Given the four-step analysis process of grounded theory described above, ATA involves steps one to three (reading, finding and coding themes, structuring themes), with a portion of step four (building a theoretical model). Also with grounded theory ATA shares the need for an interpretation of data always coherent with the actual texts at hand. Therefore, the proposed approach is systematic (e.g., in the codebook development, code application and data reduction) and iterative.

Besides these linkages with grounded theory, ATA's primary purpose is to "describe and understand how people feel, think, and behave within a particular context relative to a specific research question" (Guest et al., 2012, p. 13). This makes ATA closer to phenomenology. However, while phenomenology (as well as interpretive phenomenological analysis) focuses specifically on individual human experiences, ATA embraces a broader topic range, that can include also social and cultural phenomena (Guest *et al.*, 2012, p. 18). The main feature that distances ATA from grounded theory and phenomenology is the fact that ATA allows the use of quantification and data reduction techniques, as long as they are complementary to the qualitative analysis and appropriate to the analytical purpose. Given its hybrid nature, ATA offers a solution that maximizes the level of in-depth, qualitative analysis while allowing for the examination of large sets of textual data.

As Guest *et al.* (2012, p. 16-18) emphasise, ATA is not new. It is largely based on "commonly employed inductive thematic analysis and shares features with grounded theory and phenomenology". It should also be noted that the different techniques and procedures described by the author ought not to be taken as static and final. ATA is primarily a pragmatic way of approaching qualitative data that

recognizes the positivist requirement of evidence-based statements and takes the data as paramount in deciding what analytical procedures to follow.

METHODS EMPLOYED IN PRESENT THESIS

The purpose of this thesis, as expressed in Chapter 1, is to offer an exploration of critics' judgements of performances through the investigation of the following questions:

1. What do critics write about when reviewing recorded music performance?
2. How are the diverse elements discussed used to build value judgements?
3. To what extent is the emergent model shared between different critics?

To address these questions, the analysis of a representative sample of music criticism was needed. This part of the chapter is divided into two sections: in the first, the object of investigation of this study is defined and reasons are given for the applied delimitations. In the second section, the process of analysis applied to this collection of texts is briefly presented.

Object of analysis: A sample of criticism

As discussed in Chapter 1, the label 'music criticism' embraces a wide range of meanings, and the activity it refers to can manifest in different forms; to render this study meaningful and manageable it was therefore necessary to delimit the enquiry by clarifying the object of investigation.

Criticism of recorded performances

The broader scope of this research was to contribute to the understanding of the process of evaluation and appreciation of musical performances; hence the investigation focuses on *performance* criticism, as it developed at the beginning of the twentieth century in relation to the performance of the canon repertoire in classical music. It was also important to have an example of criticism strictly linked to the everyday musical practice, directly relevant to any professional musician, and suitable for systematic investigation. Therefore, the research focuses on performance criticism as it is exemplified in *critical reviews of classical music recordings.* There

are multiple reasons for choosing critical reviews of recordings instead of reviews of live performances:

- First, recordings offer an acousmatic experience of the musical performance. Given the ample evidence of influence of visual stimuli in listeners' evaluations, focusing the examination on recording reviews rules out these possible influences. This in turn allows a stronger focus on interpretative issues – whose investigation is a main objective of this research – following Gabrielsson's (2003) request to focus evaluation research on aesthetic aspects of performance.;
- Second, critical review of recordings offer examples of descriptions and evaluations of real world professional performances, thus permitting the exploration of performance evaluation in an ecological context;
- Third, recordings do not simply represent a surrogate of a live performance; they are artistic products in their own right and arguably call for a set of evaluation criteria that do not apply to other forms of performance. No research so far has investigated what these recording-specific criteria may be, thus this study may offer a contribution in this direction;
- Fourth, reviews of recordings offer the possibility to set the critical text against a fixed and reproducible sound object, thus allowing for comparisons between the verbal expression and the artwork of which the text is a review;
- Fifth and last, most of the music listened to today by most people is recorded music, hence this kind of musical object seems the most suitable for an investigation of the critical practice as it relates to the current reality of the everyday musical world.

Critical review by professional critics

In Chapter 1, the necessity of looking at descriptions of performances produced by listeners who possess not only musical understanding but also solid linguistic skills and plenty of experience in utilising those skills to describe and evaluate music was also discussed. Thus the object of this study was further limited to *music magazine* reviews of classical music recordings, *produced by professional critics.*

The distinction between the professional and the non-professional critic is a controversial one. The paths that may lead to the career of music critic are various, and the activity of music criticism itself ranges from occasionally reviewing a

concert for a local newspaper to regularly contributing to a specialized magazine (and earning a living from it). As former *New York Times* music critic Bernard Holland claimed: "you have taken no bar exam, fulfilled no residency, acquired no license to practice. The day I put 'music critic' after my name people started asking me about music; before that, no one asked my opinion about anything" (Holland, 1996). For the purpose of this study, professional critics are defined by the regular activity of publishing criticism of musical works and/or performances in magazines and/or newspapers and being paid to do so.

Gramophone's reviews of Beethoven's sonatas

Given this delimitation of the study object, it was then necessary to produce a corpus of critical review of recorded performance that offers a representative sample of best practice of performance criticism in classical music repertoire. This corpus should be large and varied enough to allow significant overview of the critical practice and focused enough to permit an in-depth investigation. The specialized music magazine *Gramophone* was chosen to provide source material. The choice of one single magazine sets limits to the study in terms of use of language and cultural context. At the same time, the magazine is one of the oldest publications for classical music reviews, and there is no doubt regarding the authority that the musical world bestows on its reviews and reviewers. The magazine was founded in 1923, and since, it has published issues without interruption, offering more than 90 years of music witnessed through the ears and eyes of its critics. This heritage of material was made public in 2009, when the complete archive of the magazine was digitized and published as open access on the *Gramophone* webpage. The authority of the magazine as a leading institution for reviews of classical music recordings, combined with the unique coverage of recordings it offers (over nine decades) and the availability of this material in digital format made the choice of the *Gramophone* as source material a unique opportunity for an in-depth investigation of a sample of best practice of performance criticism.

Within the material available in the *Gramophone*, the object of study was restricted to those reviews concerning one or more of Ludwig van Beethoven's 32 piano sonatas. These sonatas form an essential part of each pianist's standard repertoire – since Hans von Bülow, they have come to be known as the *New Testament* of piano repertoire (Walker, 2010, p. 175) – and are probably among the

classical compositions most often performed and recorded in the last century. They offer a corpus of musical material that is varied enough to provide stimulus for differentiated critical praxis, at the same time representing a comprehensive and close unity. As such, the choice of this repertoire allows, on the one hand, access to a potentially vast and rich amount of critical material and promises, on the other, to give insight into the critical practice that has a direct relevance to the majority of pianists and piano pedagogues.

Given these delimitations, the object of investigation of this study can now be so defined:

> All reviews of recordings of one or more of Beethoven's 32 piano sonatas published in *Gramophone* from its foundation in April 1923 to September 2010 (date of beginning of the present inquiry).

Reviews analysis

Given this definition of the object of study, the next step was to decide what analytical stance to adopt in its investigation. As discussed in the first part of this chapter, text analysis is open to an ample spectrum of analytical methodologies. Even among unstructured texts, reviews of music recordings can be expected to constitute a particularly demanding challenge. Three considerations can help illustrate the point.

First, musical parlance is a complex and still poorly explored field in which everyday words are both used to describe or suggest aesthetically significant features and employed as technical terminology, as in the expressions 'slow *movement* of the sonata' versus 'the forward *movement* and dramatic tension produced by the pianist'. The use of words borrowed from other semantic fields and in general common to everyday parlance as technical terminology represents an obstacle for automated analysis processes, in that it creates a large amount of noise in the data with which the applications must deal.

Second, following Mudambi and Schuff (2010), music recordings are experience goods *par excellence* – that is, products whose understanding and evaluation requires direct experience with the object. That makes it difficult for the reviewer to rely on precise, objective features of the performance to, as *Gramophone*

editor James Jolly states, "characterise a performance with such vividness that the reader takes over from the critic as the final arbiter" (Pollard, 1998, p. 202). If the value of a performance is linked to the kind of experience it can prompt in the listener, reviewers can be expected to rely on figurative language and use of metaphors to elicit in the reader a feeling of what the recording can offer.

Third, an exploration of how expert critics construe their judgements of performances needs to account for different levels of features critics discuss. Praising a performance for 'the well graded crescendo that perfectly conveys the dramatic tension of the passage' could be seen as related to the use of specific interpretive choices (crescendo), the expressiveness of the performance (dramatic tension) or the stylistic appropriateness of the character expressed. This level of analysis moves further beyond the semantic understanding of the explicit features discussed that can be offered by information extraction applications, and it also requires more flexibility in the categorization of text units that the one allowed by clustering methods (that as discussed imply categories that are comprehensive and mutually exclusive).

For these reasons an interpretive approach represents the best solution to tackle critics' writings and account for their density and complexity given by the nature of the object of discussion (performance value) and probably enhanced by the high level of expertise critics possess. At the same time, given the popularity of Beethoven's piano sonatas it can be expected that the collection of reviews – even if restricted to the *Gramophone* magazine – will produce a set of documents too large for an in-depth qualitative investigation. Preliminary exploration of the collected material through content analysis and text mining procedures can then offer a complementary tool of investigation and enable a more focused thematic analysis. Based on these reflections, this research employs the hybrid approach of Applied Thematic Analysis, as it is described by Guest et al. (2012).

One reflection is important at this point concerning scope and methodology of the present research. A qualitative analysis focused on real world critical review of recorded performance is open to a series of different methodological perspectives. In particular, a thorough investigation of the critical review practice would require the examination of critics' writings in the historio-cultural context in which they were produced. To this aim, employing a discourse analysis approach, the content of the critical review text could be set in a wider scenario, discussed against evidence from

other source material, such as readers' letters, interviews with critics, information on the historical and cultural context, music market data by different labels and retailers, and the analytical reality of the recording sound. A similar approach would offer a thorough and historically focused examination of the critical practice in the Gramophone magazine and would undoubtedly deliver rich and insightful findings, relevant to our understanding of the nature and development of music performance criticism in the British music market.

In the present thesis, it was decided, however, not to take such broad approach but rather focus the investigation on the analysis and categorization of the actual content of critics' writings. This decision was led by three main considerations:

- First, a broader examination of recorded performance critical review employing a discourse analysis approach would shift the focus of the research away from what has been stated as the purpose of this work: that of exploring professional critics' judgements of recorded performance to gain insights relevant to our understanding of expert performance evaluation.
- Second, a discourse analysis of music performance criticism in the British classical music market would require a comprehensive examination of the content of reviews (the aim of this thesis) as a prerequisite, prior to discussing this content against other pieces of information. In this perspective, the work reported in this thesis can be understood as a preliminary, necessary step to a further, broader investigation. Given the complexity of this first task alone, and the lack of extant literature on the matter, embarking on a larger exploration that accounts for different sources of information at once would lead to a project requiring resources far beyond the scope of this research.
- Third, the content of critical review as it is expressed in the published *Gramophone* magazine is the only information source consumers and musicians can access – the actual feedback that critics offer to readers. It is then of utmost importance for consumers as well as producers to obtain insights on what is written in these texts, how critics' judgements are structured and justified and what kind of vocabulary is used in the description and evaluation of performance.

Based on these considerations, the focus of the present research was narrowed to the analysis of the content of critics' writings. Although contextual elements have

been discussed and considered where appropriate along the course of the analyses, no systematic examination of source material other than the published reviews was produced. This allowed for the investigation to remain feasible while delivering clear answers to the research questions and insights directly relevant to musicians, music consumers, and music pedagogues. Hence, following the procedure suggested by Guest et al. (2012), the study is divided into two stages: an overview of the collected material with application of data reduction techniques, and a series of inductive qualitative thematic analyses. Taken together, the different analyses offer a thorough examination of the collected material, and bear evidence of the kind of insights that can be gained through an investigation of music critical review at metadata, word-stem, clause, and sentence/paragraph level.

Overview of the collected material

Following the ATA approach (Guest et al., 2012; see also Namey et al., 2008), in a first step a database with metadata of the collected reviews was prepared and descriptive and inferential statistics were run on the whole dataset. Following, content analysis and text mining procedures like word frequency and state-of-the-art categorization algorithms were applied and used as data reduction techniques. Finally, focused thematic analysis employing *keyword-in-context* report was used to further prepare the ground for the inductive thematic analyses by clarifying critics' usage of a common and important, yet ambiguous term in musical discourse: 'expression'. Details of these analyses are provided in Chapters 3, 4, and 5. Insights gained through these preliminary analyses informed and shaped the subsequent investigations and led to the selection of a representative and yet manageable sub-corpus of material suitable for thematic analyses.

Text analysis

In a second step, the selected corpus of reviews was analysed. The analytical purpose of this research was explorative and the process applied inductive. No pre-defined categories were applied at the coding stage. Three analyses were run separately to address the following questions:

1. What performance features do critics discuss in their reviews? To what extent is the emergent model shared between different critics? (Chapter 6)

2. How are considerations on these performance features used to build value judgements? To what extent is the emergent model shared between different critics? (Chapter 7)
3. What elements beyond the performance do critics discuss in their reviews, and how are considerations on these elements used to build value judgements? To what extent is the emergent model shared between different critics? (Chapter 8)

In the text analyses, three techniques suggested by Guest et al. (2012, chapters 3 and 4) were applied to enhance validity and reliability of findings:

- Development and use of a systematic codebook entailing clear descriptions of the theme or theme element(s) represented by the code;
- Avoidance of multilayer-coding (coding based on memos or other codes): all codes are attached to the source data;
- Use of inter- and intra-coder re-checks to control for researcher's biases.

These procedures cannot, by themselves, assure validity and reliability of findings, but they can help avoid some biases and mistakes caused by a lack of systematicity and transparency in the application of codes. Details of the analysis process are given in Chapters 6, 7 and 8.

The final outcome of these analyses is a visual descriptive model of critical review of recorded performance. This model is summarised in Chapter 9 together with the discussion of its limitations and further implications.

3 GRAMOPHONE REVIEWS I: AN OVERVIEW[3]

After the context given in Chapter 1 and the methodological considerations presented in Chapter 2, this chapter introduces the collected sample of critical review of recorded performance. It offers a systematic, explorative analysis of reviews metadata and discusses the relevance of this material's heritage for understanding the processes behind experts' evaluations and their implications for musical practice.

In what follows, the collection procedure is presented and an overview is given of review structure and length, of the repertoire reviewed and of the people (pianists and critics) involved. In the conclusion of the chapter, suggestions are given about the way these first findings inform the analyses that follow.

METHOD

From the online archive of the monthly magazine *Gramophone*[4] all reviews were extracted – with permission of the magazine[5] – published between April 1923 and September 2010 that concerned commercial recordings of one or more of Beethoven's 32 piano sonatas. To assure as complete a collection of material as possible reviews were collected in two successive phases: first using the *search* tool of the internet site and then browsing every issue page-by-page as they appear in the scanned online version (1,050 issues).

Review texts were collected in Microsoft Word documents, ordered chronologically and divided per decade, with 'decade' understood as periods of 10

[3] Content within this chapter has been published within the following: Alessandri, Eiholzer and Williamon, 2014; and Alessandri, Eiholzer, Cervino, Senn, and Williamon, 2011. For full references, see List of Publications.

[4] Accessed at http://www.gramophone.net. The archive was opened in 2009 but is no longer available publicly. Access to the digital collection of *Gramophone* reviews, including all texts used in this study, can now be purchased as an application for iPad, desktop or tablet.

[5] *Gramophone* granted us written permission to store a private e-repository of reviews from the archive. The repository would be: (i) used for research purposes only (entirely non-commercial); (ii) stored electronically and securely (password protected) to be accessed only by the research team (which consisted of the author and her supervisors); and (iii) used in compliance with all other Terms & Conditions listed at the time of the collection on the *Gramophone* website, with exception of the point 2.v.

years beginning with a year ending in 1 (e.g., 1951-1960). The first decade however encompassed only 7 years and 9 months, starting in April 1923 (first issue of the magazine) and the last decade included 9 years and 9 months, missing the last three months of the year 2010 (not yet published online at the time of the collection).

For each review, a database was compiled with the following information: issue (date, page); sonata(s) reviewed; pianist(s) reviewed; label; critic; release status (i.e., new release, re-issue, first release of an old recording[6]); repertoire reviewed (i.e., only one or more Beethoven's sonatas or one or more of Beethoven's sonatas plus other works); presence of comparison(s) with different pianist(s); and length of the review (in words). Descriptive and exploratory data analyses were carried out on the whole dataset. In the present chapter, the results are grouped into four sections that focus on the structure and length of the text, the repertoire reviewed, and the pianists and critics involved, respectively.

RESULTS

In total 845 reviews (334,210 words of critical text) were collected. For six reviews the body of text in the online *Gramophone* archive was damaged, hence for these reviews information regarding text length and, in some cases, the name of the critic, release status and pianist reviewed could not be integrated in the analyses.

The distribution of reviews by decade is shown in Figure 3.1. The spread of reviews presents two distinct periods, 1950 being the watershed between the two. In the first three decades (up to 1950), the publication rate was 2.59 reviews per year, with a trough in 1941–1950 (16 reviews). Subsequently (1950 – 2010), reviews were distributed relatively evenly, with an average rate of 12.94 reviews per year and a peak in 1961-1970 (150 reviews).

[6] These are cases of recordings produced several decades (between c. 20 and 70 years) prior to their public release. In contrast to other recordings – usually released a few months after their production – these recordings did not seem to be meant (or chosen) to be released publicly in the first instance (e.g., radio broadcasts, live concerts). The peculiarity of these recordings is underlined by critics, who emphasise in the reviews' titles – and sometimes again in the body of the reviews – the time and context of production (not mentioned for other recordings).

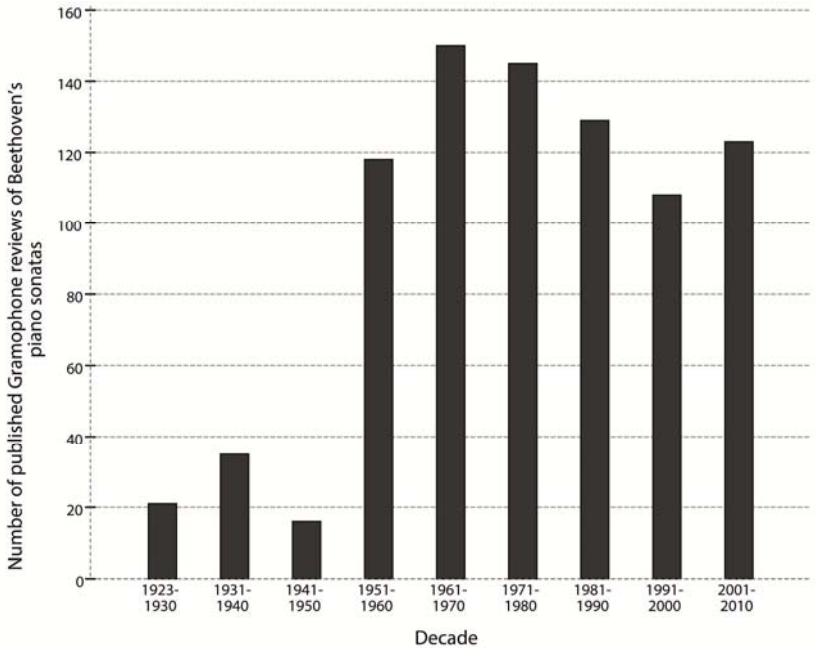


Figure 3.1. Distribution of collected reviews through decades.

The trough in the 1940s can be ascribed to the historical events that took place in the first half of that decade. Despite the effort in continuing publishing, the severe conditions during World War II affected the magazine deeply, both directly – through paper rationing that called for a size restriction up to three quarters of the pre-War size – and indirectly – by constraining the record industry production through also, but not only, shortages of raw materials such wax and shellac (for a detailed description of the repercussions of World War II on the *Gramophone* see Pollard, 1998, pp. 58-77).

Structure and length

Within the *Gramophone* magazine, reviews were found in two distinct sections: "Analytical Notes and First Reviews" and – starting in 1936 – "Second Reviews": a space devoted to recordings that were re-reviewed some time after their appearance on the market, to offer reviewers the opportunity to discuss and describe performances more at length. The "Second Reviews" section disappeared at the end of the 1970s; after this point all reviews of Beethoven's sonatas come under the common label "instrumental". Soon after the launch of the magazine (early 1930s), reviews developed a clear two-part structure: titles containing information regarding the object being reviewed (piece(s); player(s); label; format, price, and when

appropriate original recording) and critical text. At the end of the text the review is often signed with either the name or initials of the author. Starting in 2000, reviews also begin with a one sentence title-like statement.

Critical text parts of the reviews (henceforth, simply reviews) are on average 411.74 words long (SD = 278.81); however their length ranges between 10 and 2446 words. The 10-word review concerns Op. 13, "Pathétique" first movement performed by Frederic Lamond. It appeared in a miscellaneous section on September 1943 with the text "An impressive performance of one of Beethoven's masterpieces; brilliantly played" (unsigned). The 2446-word text is a review by Richard Osborne published on January 1992 concerning EMI's re-issue of Arrau's recordings of Beethoven's five concertos, Variations on an original theme in C minor, and piano sonatas Opp. 27/2, 31/3, 53, 54, 57, 81a, 101, 110, 111 (5 discs).

Review length was found to be associated with different factors such as decade (Kruskal-Wallis test: $H_8 = 60.53$, $p < .001$, see Figure 3.2), and author ($H_{10} = 41.36$, $p < .001$, computed for the 10 most prolific critics, see Table 3.4 in the 'Critics' section).

Also repertoire and pianist reviewed were found to be correlated with review length. Reviews of recordings entailing mixed repertoire (Beethoven's sonatas plus something else) were longer (Mann-Whitney test: $U = 9{,}898.50$, $p < .01$) and more varied in length than recordings of only Beethoven's sonatas (Table 3.1), a fact that could be ascribed to the higher heterogeneity of quantity and nature of the repertoire discussed in those reviews. A moderate positive correlation was found between review length and pianist reviewed, with more often reviewed pianists (see Table 3.2 in the 'Pianists' section) receiving longer reviews (mean = 452.93 words) compared with less often reviewed ones (mean = 369.78 words); $U = 73{,}017.50$, $p < .001$.

Table 3.1. Length of reviews concerning only Beethoven's sonatas and of those discussing mixed repertoire.

	Reviews of only Beethoven's sonatas	*Reviews of mixed repertoire*
Mean length (words)	358.71	454.76
SD	235.55	333.49
Range	10-1830	45-2446

However, it is reasonable to assume that the main factor that determines the length of the text is the editor's choice, usually driven by logistical necessities. Review length is most likely influenced by the ratio between material to be inserted in the magazine and the available space. Lionel Salter in his contribution to Pollard's book (Pollard, 1998) asserts that in the 1990s reviewers were strongly inhibited by such a lack of space. Thinking of the past, he remembers the times of LPs as a golden age in terms of space allotted to each review.

> Looking back, we view with envy the amount of space then permissible for reviews. In early 78rpm days these had been very brief, but with the advent of LP a length of one-and-a-half pages was not unheard of. However, with the ever-increasing number of issues and the consequent pressure on space, more succinct writing became imperative. Decisions always need to be taken on what is really significant and how far to go into detail – bearing in mind the readership extends from 'ordinary music-lovers' to highly informed listeners with an astonishing range of musical, technical and discographical knowledge (whose letters and corrections are much appreciated) (Lionel Salter in Pollard, 1998, p. 197).

These claims seem to find only partial support in the present study findings when examining the average length across decades (Figure 3.2). On the whole, reviews length differs significantly in different decades ($H_8 = 60.53$, $p < .001$). At the beginning of the *Gramophone* life reviews varied extremely in length – probably due to the fact that reviews, as journalistic product, still needed to find their own format and structure. Already in the 1930s, variability decreases and the mean length growths noticeably (even if not significantly). These seem to have been prolific years for the magazine, when there were still few recordings to review but plenty to discuss about them. The 1940s brought a strong cut in review length due to war restrictions, as mentioned above (Kruskal-Wallis pairwise comparison between 1930s and 1940s $H_1 = 3.38$, $p < .05$). From the 1950s, as Salter recalls, reviewers did again enlarge their texts, and the LP era sees a steady increase in length. From the 1980s, however, the situation changes again, settling at a length of circa 350 words at the beginning of the 21st century.

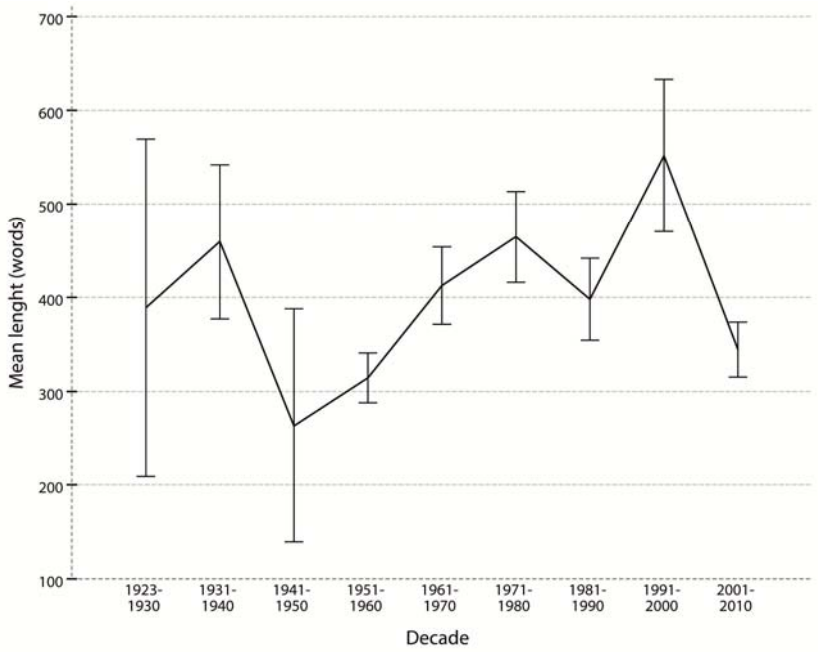


Figure 3.2. Mean value of review length (in words) across decades, with 95% CI error bars.

In this pattern, it is striking to see a 552 mean value scored in the 1990s. It is difficult to explain the reason for this high number, especially in relation to Salter's claims. This decade saw a high concentration of very long reviews (most of them reviews of recordings with mixed repertoire): eleven of the thirty longest reviews of the whole set including three of the four longest (2446, 2287, and 2041 words) are to be found here.

Repertoire

Recordings reviewed may entail a single sonata, a group of sonatas or the whole cycle of 32 pieces. Out of the 845 collected reviews, 322 concern recordings that include one or more of Beethoven's sonatas alongside another composition. These works might be Beethoven Bagatelles or piano concertos or works by other composers, and the section of review concerning these other works ranged from a few words to a more than 90% of the whole text.

Throughout the whole corpus of reviews, the four most often reviewed sonatas are (ranked in descending order): Op. 27/2 *Moonlight*, Op. 57 *Appassionata*, Op. 13 *Pathétique* and Op. 111. With the exception of a small group of sonatas (Op. 31/2 *Tempest*, Op. 53 *Waldstein*, Op. 81a *Les Adieux*, Op. 106, Op. 109 and Op. 110)

all other sonatas are homogeneously spread around 75 instances each (Figure 3.3).[7] Among the least reviewed sonatas we find the so-called 'easy sonatas' (Opp. 14, 49 and 79), together with Opp. 7, 22, 54 and 31/1.

Beethoven's 32 piano sonatas represent, together with the Well-tempered Clavier of Bach, rare cases within classical music: over time, they have developed a strong identity as a group, or *cycle*, almost as if they were one entity.[8] Using Joachim Kaiser's words: "this intimate and adventurous path from c to c – Op. 2 No. 1 begins with the note c and the C minor sonata Op. 111 closes with c – these 32 works, performed always again and again, build a cosmos that is multitudinously rich, and yet as totality completely coherent" (Kaiser, 1975, p. 24, translation by the author).

Within this organic "path from c to c" a few sonatas represent milestones *en route*: that is the case for instance for the three sonatas Op. 2 dedicated to Beethoven's teacher, Joseph Haydn, which exemplify the classic style at the beginning of this path, then the named sonatas, Op. 13, *Pathétique*, Op. 27/2, *Moonlight*, Op. 31/2, *Tempest*, Op. 53 and 57, *Waldstein* and *Appassionata* and Op. 81a *Les Adieux*, which sign different developmental stages along the road, and finally the *late sonatas* group, beginning with Op. 90 and culminating in the colossal Op. 111.

The metaphor of the path and the strong feeling of completeness and variety linked to this cycle is justified by the fact that these 32 sonatas seem to reflect different periods in Beethoven's professional and personal life: from the early Vienna period at the end of the eighteenth century through the *heroic style* period to the last years, signed by the highest technical and musical maturity but also by the tragedy of Beethoven's deafness and increasing isolation. The connection between the 32 sonatas and the composer's life is very strong as are the components of heroism and tragedy linked to Beethoven's image (Burnham, 2000), and it is not

---

[7] This finding is not particularly surprising in that the most often reviewed sonatas are indeed the most popular ones. It suffices to check Amazon.com to have evidence of this. A search for Beethoven's *Pathétique* or *Moonlight* sonatas gives as result more than 1200 each. When searching for any other Beethoven's piano sonatas – excluding the four most often reviewed ones and excluding complete cycles – there are "merely" 451 products all together (search done in the "Music" section of Amazon.com on January 3rd, 2012). Of course a perverse question remains open; if these sonatas are the most famous because of the audience preferences or because of the pianists and labels choice to record those most.

[8] These two sets of masterworks have indeed come to be referred to as the Old and New Testament of piano repertoire.

unusual to hear, for instance, that a young pianist can or should not perform Op. 111, no matter how musically gifted s/he is, since to perform this sonata properly (or even fairly) a certain maturity and experience with life, not just with music, is needed (see for instance Fischer, 1956, p. 14).[9] With this background in mind, the distribution of sonatas for the three periods of L. v. Beethoven's activity was explored separately (Opp. 2 to 28 first period; Opp. 31 to 78 second; Opp. 90 to 111 third).

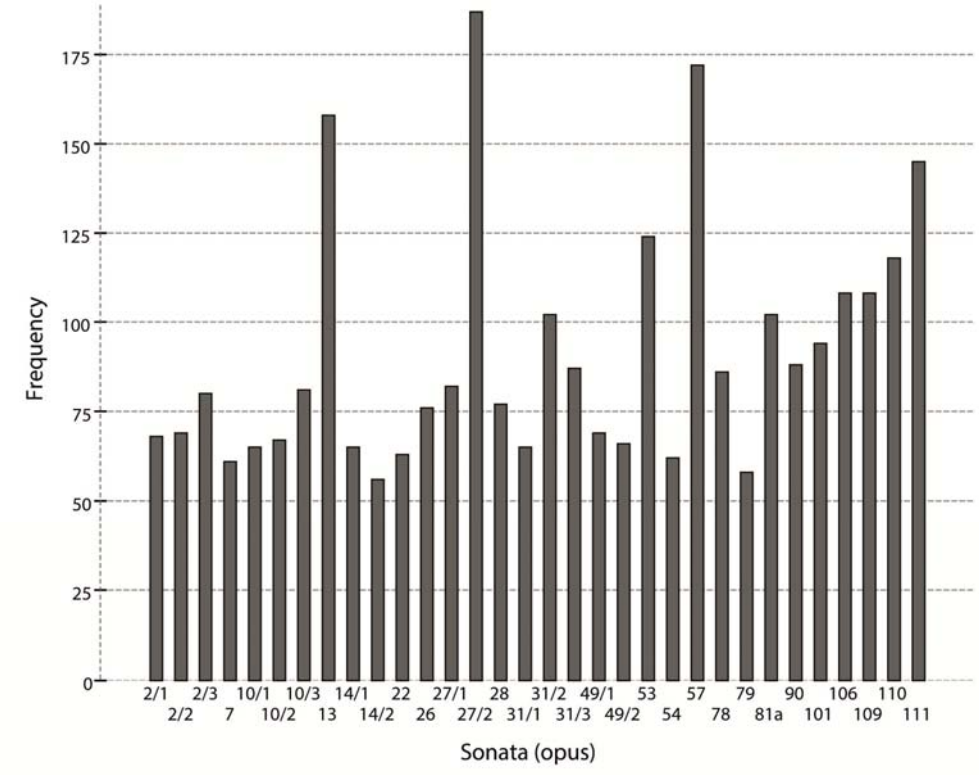


Figure 3.3. Frequency of sonatas reviewed within the whole dataset (1923-2010).

For each review three variables were computed that indicate – for each of the three periods – the total amount of sonatas present in the reviewed recording. Given the different quantity of sonatas that occur in each period (15 for the first, 11 for the second, and 6 for the third) the resulting values were standardized to allow for comparison between the three groups of sonatas. Mean standardized quantity of

[9] See also *Gramophone* review, March 1988, p. 50. Here this view seems implied in Stephen Plaistow comments on Taub's recording of Op.111. After praising the "young American pianist" for his "authentic Beethovenian energy, …fuelled by the mind rather than the fingers alone" he continues: "Who said that pianists have to be old and grey before we can expect them to have insights into Beethoven's last sonatas?"

sonatas in each decade for each of the three periods is shown in Figure 3.4; the first three decades are merged together due to the low number of reviews in those years.

A strong increase is observed over the course of the century in the mean number of sonatas entailed within one review (adding all sonatas together, Kruskal-Wallis test: $H_8 = 118.70$, $p < .001$) as a consequence of the technological developments that allowed much longer recording time at lower production costs.

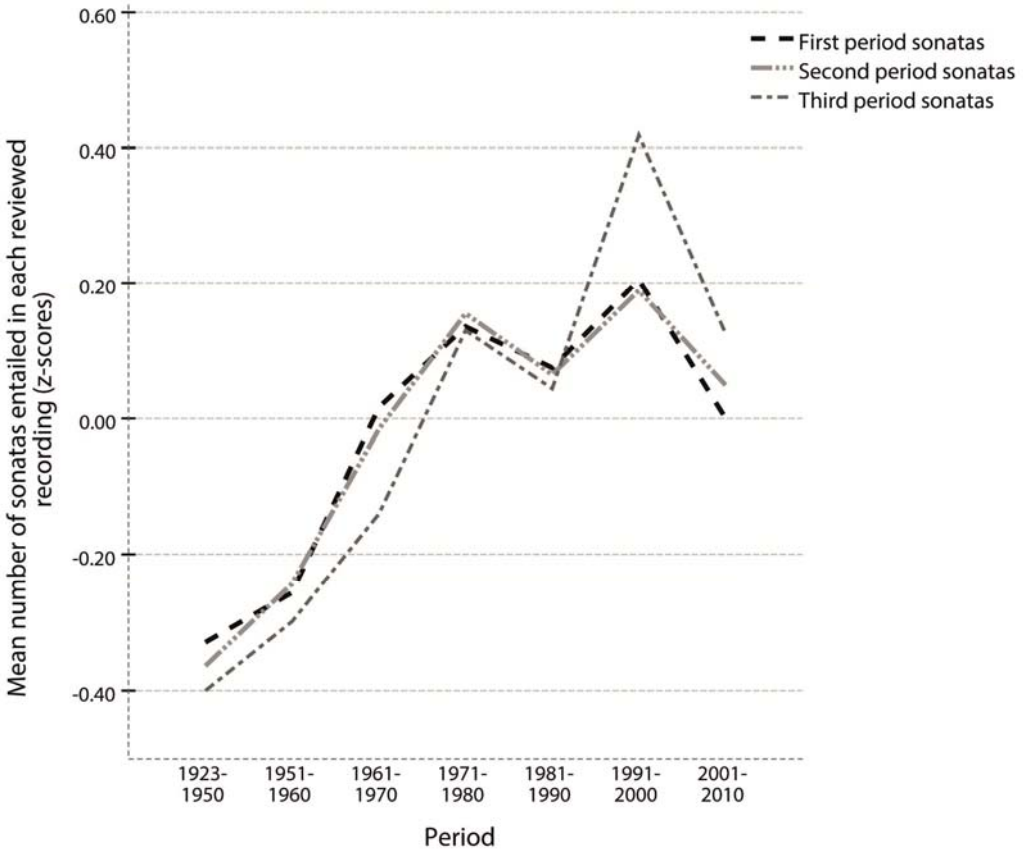


Figure 3.4. Mean number of sonatas (z-scores) in each reviewed recording, for the three compositional periods of Beethoven's activity, across decades.

The three groups of sonatas do not develop equally across decades. Late sonatas were least common at the beginning of the 20th century, slowly increased their presence along the years and reached the other groups of sonatas in the 1970s and 1980s. In the last two decades these late sonatas became the most prevalent group, high above the first and second period sonatas. First period sonatas were the most common in 1923-1950, but the least often reviewed at the end of the century (this despite the presence of the Moonlight sonata, which belongs to the first period

and is the most reviewed sonata overall). Friedman's test showed a significant difference in the distribution of the three groups of sonatas, $\chi^2(2, N = 845) = 73.67$, $p < .001$. Post-hoc pairwise comparisons with Bonferroni correction applied revealed significant differences between sonatas of the first and third period ($Z = 7.44$, $SE = 0.05$, $p < .001$) and of the second and third period ($Z = 5.15$, $SE = 0.05$, $p < .001$).

The same pattern can be observed more in details in Figure 3.5 to Figure 3.8, which show the distribution of sonatas for the four periods 1923-1950; 1951-1990; 1991-2000; and 2001-2010. The first three decades show just a few occurrences of reviewed sonatas; and here a peak can be noticed by Op. 13 followed by Op. 27/2. This peak shifts in the following decades (1951-1990) toward the 'right side' of the Opus number, with Op. 27/2 overcoming all other sonatas, followed by Opp. 57, and then 13. In the 1990s it is no longer possible to isolate few outliers, and the whole late-sonatas block becomes more present (Opp. 90 to 111). Opp. 13 and 27/2 are still present, but other sonatas like Op. 31/2 and 81a come more to the fore. In the last decade (2001-2010) the peak is reached with Op. 111.

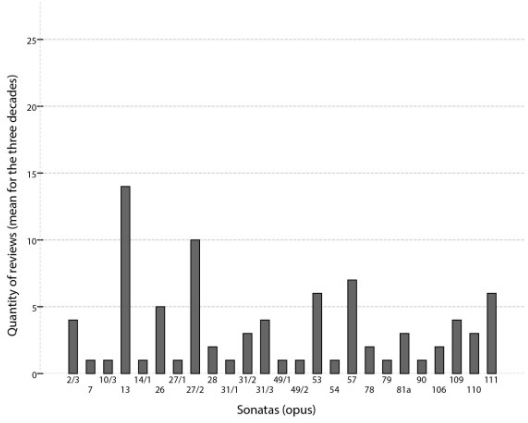


Figure 3.5. Frequency of sonatas reviewed, 1923-1950.

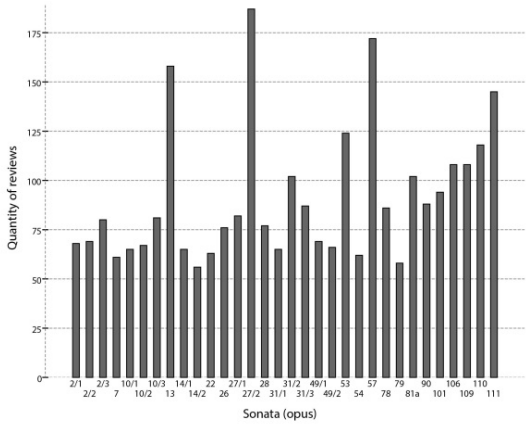


Figure 3.6. Frequency of sonatas reviewed, 1951-1990.

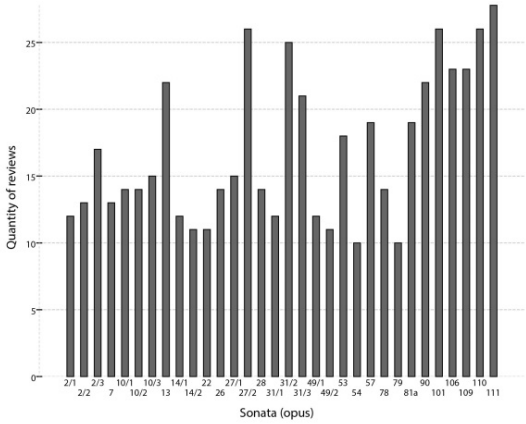


Figure 3.7. Frequency of sonatas reviewed, 1991-2000.

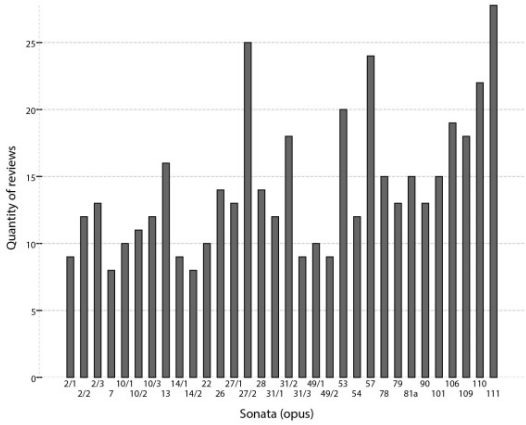


Figure 3.8. Frequency of sonatas reviewed, 2001-2010.

Re-issues

Out of the 845 reviews collected from *Gramophone* 205 (24.28%) were reviews of re-issued recordings. The term *re-issue* may relate to different kinds of products. In the present thesis re-issue is used to indicate *any commercial release of a recording other than its first release*. According to this definition re-issues may be releases of a recording in a new format (e.g., 78rpm released as Long Playing and then as Compact Disc), as well as recordings released in the same format more times by the same or by different label(s) (for instance, once as single disc and once as box set). In these terms re-issues may or may not include different degrees of engineering work.

An example of different kinds of re-issues is offered by Wilhelm Kempff's second recording of the complete cycle of the 32 Beethoven's sonatas. The recordings were produced in 1964-65 in the Beethovensaal in Hanover (Germany),

by Deutsche Grammophon Gesellschaft (DGG) under the Tulip label. First released in 1966 as eight separate LPs, they were re-issued in 1972 as box-set and then again in 1975 as special edition "Hommage à Wilhelm Kempff" (on the occasion of the pianist's eightieth birthday). In 1990 a further re-issue, in Compact Disc format, was released. This was deleted from the catalogue seven years later and replaced by a newly restored version, after having gone through a re-mastering process that assured a cleaner sound quality. Given the continuous success of this recording, in 2008 DGG produced one more re-issue. Along with these main releases, the sonatas were also issued several times as singles or groups of pieces (for instance the late sonatas group alone). They also appeared in 1970 and then again in 1977 in the DGG "Beethoven Edition", a colossal set encompassing all of Beethoven's music recorded so far by the label. Few of these releases found a correspondence in the *Gramophone* pages. The two main reviews concerning this Beethoven's sonatas cycle by Kempff appeared in 1966, for the first release of the set, and then again in 1990 with the release of the Compact Disc version. It also got a shorter mention within the large reviews of the "Beethoven Editions" in 1971 and in 1977.

In the collected sample, distribution of re-issues was strongly associated with the decade (Pearson's $\chi^2(16, N = 844) = 256.92$, $p < .001$, Phi = 0.55, $p < .001$; in the analysis re-issued and partly re-issued recordings, as distinguished below, were merged in one category). Reviews of re-issues first appeared in 1951 (February, p. 24), with Alec Robertson reviewing a Decca Long Playing re-issuing Backhaus's recording of Op. 109, in E major, and Chopin's "Funeral March" sonata. The presence of re-issues increased toward the 1980s, when the ratio between new recordings and re-issues being reviewed was almost 1:1. After 1990 this tendency receded but was partly compensated for by a new phenomenon: old, unpublished recordings – such as broadcast recordings that were never commercially released or old recordings which were not selected for a published release in the first instance – were suddenly made commercially available. The first example of this kind is Josef Lhévinne's Op. 27/2, Moonlight sonata, originally recorded on piano rolls together with several other pieces on a Norman Evans Estonia-Ampico piano at the beginning of the century and released in September 1985 by L'Oiseau-Lyre. The interest in old recordings has increased since then, so that in the 2000s almost one third (26.83%) of reviews concerned this kind of product. Within this picture, starting in the 1960s, a

small number of reviews concerns partly re-issued recordings. That happens when there is a release of a group of sonatas, some of which are newly recorded while some others are taken from previously published material, for instance, in order to complete a cycle (Figure 3.9).

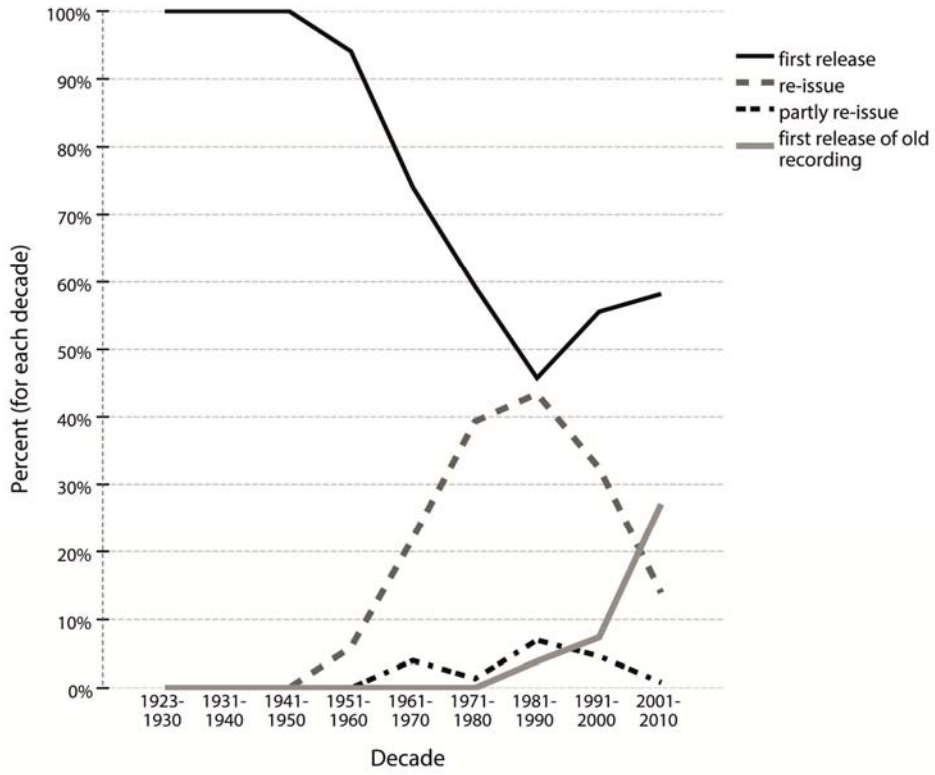


Figure 3.9. Distribution of reviews across decades according to the release status of the reviewed recording.

Pianists

Pianists reviewed in the collected *Gramophone* sample number 216, but merely 17 of them cover 51.95% of all reviews. So, while Arrau was reviewed 53 times and Brendel 52 times, there are 117 pianists who are reviewed just once throughout the century (a complete list of pianists reviewed is reported in Appendix 1).

Out of the 216 pianists, 81 were used by the reviewers for comparisons. Of the 16 performers most often used for comparison, 14 correspond with those included among the 17 most reviewed pianists (Table 3.2). It was reported that 205 of the 845 collected reviews were reviews of re-issued recordings. Of these, 153 (74.63%) are about these 17 most often reviewed performers, so that the ratio between new recordings and re-issues for those pianists is 1.54:1 while for the

residual 199 performers it rises to 6.44:1. This difference is significant according to Pearson's Chi-Square, $\chi^2(1, N = 775) = 67.30$, $p < .001$, Phi = -0.30, $p < .001$ (Figure 3.10).

Table 3.2. The 17 most often reviewed pianists within the collected critical review corpus.

Name	*Frequency*	*Name*	*Frequency*
Arrau, Claudio	53	Barenboim, Daniel	18
Brendel, Alfred	52	Gieseking, Walter	18
Kempff, Wilhelm	49	Gulda, Friedrich	16
Backhaus, Wilhelm	38	Lill, John	16
Ashkenazy, Vladimir	27	Michelangeli, A. B.	14
Richter, Sviatoslav	26	Kovacevich, Stephen	14
Schnabel, Artur	26	Pollini, Maurizio	14
Solomon	25	Serkin, Rudolf	13
Gilels, Emil	20		

Note. Highlighted names in grey refer to those pianists who also belong to the 16 performers most often referred to for comparisons. The two pianists most often used for comparison who do not appear in the table are Orazio Frugoni and Richard Goode.

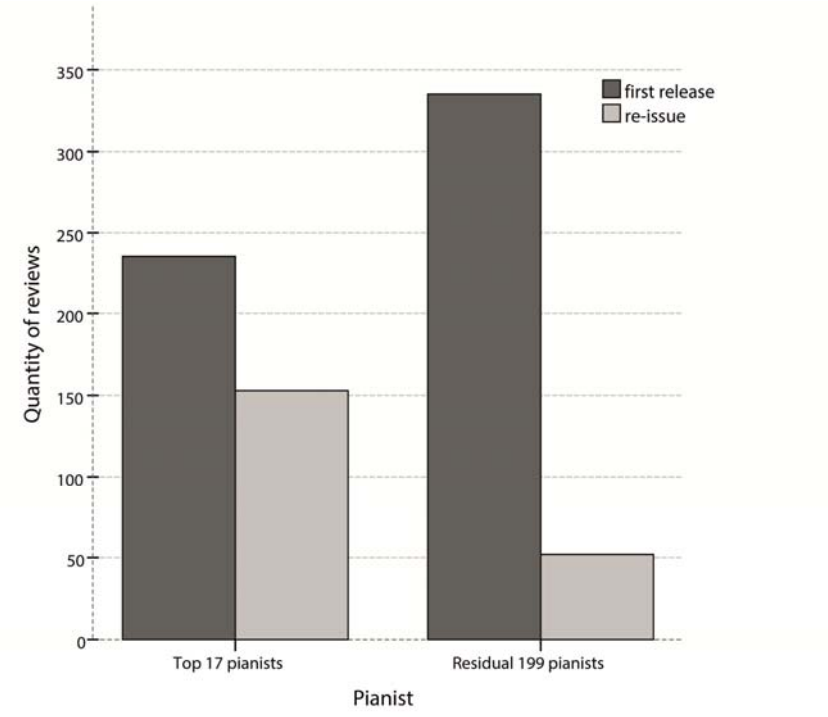


Figure 3.10. Distribution of recordings among pianists according to their release status.

Comparisons between pianists by reviewers, used to explain, justify or clarify a critical statement, were common, found in 41.28% of all reviews and 54.02% of the reviews of recordings entailing only Beethoven sonatas. Beginning in October 1953, comparisons were also stated officially in the titles of the reviews (Table 3.3).

Table 3.3. Examples of different kinds of comparisons found in *Gramophone* (pianists' names in bold).

January 1981, p.48

BEETHOVEN PIANO SONATAS, VOLUME 2 **Bernard Roberts**. Nimbus Direct to Disc D/C902 (four records, nas, £19.40). Notes included.
Roberts does not have all the tonal poise or intellectual quickness of **Schnabel** whose set of these same sonatas (HMV mono RLS754, three records to Roberts's four) is reviewed on page 998 of this issue. Among more recent cycles **Brendel's** (Philips 6768 004, 11/78) is the more enquiring, the more intellectually various, avoiding Roberts's tendency to slow the music unduly in moments of introspection…

June 1974, p. 74

PIANO SONATAS. **Friedrich Gulda**. Decca Eclipse ECS722-3 (two records, 99p each). ECS722: No. 21 in C major, Op. 53, "Waldstein"; No. 28 in A major, Op. 101. ECS723: No. 30 in E major, Op. 109; No. 31 in A flat major, Op. 110; No. 32 in C minor, Op. 111.

Selected bargain comparisons:
No. 21:
Brendel (6/64) (5/70) (R) TV34115DS
Nos. 28 and 32:
Rosen (5/70) 61127
Nos. 31 and 32:
Brendel (8/63) (3/70) (R) TV34113DS

Here are two further discs from Gulda's earlier cycle of Beethoven sonata recordings…
Of the two discs, though, this is of lesser interest, primarily because Gulda's account of the Waldstein Sonata, fleeting, deft and aerial (the semi-quaver flights in the first movement at times so deft they barely sound) is no challenge, ultimately, to the **Brendel** on Turnabout. **Brendel** plays with great economy of gesture, is as poised and fluent as Gulda is; but with **Brendel** I find the music is more strikingly articulated, the virtuoso demands more frankly met…

Note. The table shows a comparison made in the text body (top panel, used throughout the century) and a comparison stated in the titles of the review (bottom panel, used starting in 1953).

Critics

Among the 845 collected reviews, seven reviews (0.83%) were damaged so that it was not possible to read the name of the critic at the end of the text, and 73 reviews were unsigned (8.64%). The residual 765 reviews were signed with either initials or full names. Signed reviews were unusual at the beginning of the century: 62.50% of reviews published until 1950 were unsigned. At this time only few reviewers (at times just three or four) were active at the *Gramophone*, and dealt with the whole set of recordings to be reviewed (see Lionel Salter in Pollard, 1998, see also "Portray of a reviewer" sections in the *Gramophone* magazine). Concerning Beethoven's piano sonatas, beside the magazine founder Compton Mackenzie (who used to sign his reviews with the pseudonym *Z.*) the first main *Gramophone* reviewers were Alec Robertson, W. R. Anderson, and C. Henry Warren, later followed by Trevor Harvey. Their contribution together count for 30.60% of the reviews published in these 30 years according to the signatures present in the magazine.

After 1950 the number of reviewers increased steadily (by 1964 there were 16 reviewers at the *Gramophone*) and so does the number of signed reviews: between 1951 and 2010 only 4.66% of reviews are unsigned. Signatures are restricted to the initials until the end of the 1990s. After this point initials start to be substituted by full names. Among initials, pseudonyms and names, it was possible to identify 52 different critics (see Appendix 2). And among them, just 10 critics wrote 530 reviews – that is, 62.72% of the whole corpus (Table 3.4).

Most of these reviewers' activity is spread across several decades, with an average of 21.32 years between the first and the last published reviews; the highest peak was Stephen Plaistow at 41 years and 3 months. Two exceptions are Andrew Porter, who, concerning Beethoven's piano sonatas, was only active in the 1950s and Jed Distler, who started reviewing Beethoven's sonatas in 2005. Seen chronologically, some of these reviewers significantly shaped the *Gramophone* critical output, contributing substantially to the overall set of reviews for a given period. For instance, Andrew Porter wrote 34.75% of Beethoven's piano sonatas reviews published in the 1950s, while Bryce Morrison and Jed Distler together produced the 58.56% of the reviews that appeared in the 2000s.

Table 3.4. The 10 most prolific reviewers identified in the collected critical review corpus.

Reviewer	*Quantity of reviews written*	*Percentage (all reviews)*	*Period of activity*	*Percentage (for the period of activity)*
Richard Osborne	108	12.8	Apr ’74 – Nov ‘04	27.34
Stephen Plaistow	88	10.4	Sep ’61 – Dec ‘02	16.15
Joan Olive Chissell	65	7.7	Oct ’68 – May ‘93	18.90
Bryce Morrison	60	7.1	Apr ’92 – Jul ‘10	27.91
Roger Fiske	52	6.2	Jul ’55 – Mar ‘86	11.93
Andrew Porter	41	4.9	Apr ’54 – May ‘60	49.40
David J. Fanning	33	3.9	May ’85 – Sep ‘02	17.01
Malcolm MacDonald	31	3.7	Sep ’52 – Jul ‘84	6.95
Jed Distler	28	3.3	Oct ’05 – Oct ‘09	52.83
Alec Robertson	24	2.8	Aug ’34 – Jun ‘54	31.58

But who are *Gramophone* critics? As opposed to other professions, there is no standard path to follow to become a professional critic. *Gramophone* critics therefore stem from very different backgrounds. However, most of the ten critics listed above had some kind of musical education, often complemented by other theoretical studies. Alec Robertson for instance studied organ, harmony and composition at the Royal Academy of Music and was active as an organist and choirmaster as well as broadcaster and lecturer. He was also Head of the Education Centre of The Gramophone Company when Mackenzie launched the magazine in 1923 and he became one of the very first reviewers of *The Gramophone*. Roger Fiske read English at Wadham College, Oxford, where he also obtained later a DMus, and composition and criticism at the Royal College of Music, before joining the BBC as producer and broadcaster. Andrew Porter studied at University College, Oxford, thanks to an organ scholarship while simultaneously reading English.

Some critics did not just have formal music education but also were active professionally as performers and composers. Malcolm MacDonald for instance was an appreciated composer and his compositions received numerous acknowledgements (e.g., the Arts Council Scotland Award in 1946 and the Royal Philharmonic Prize in 1952). He studied composition and conducting at the Royal

College of Music, and philosophy and music at Cambridge. Among the ten reviewers listed above there are also five proficient pianists. That is the case of Jed Distler, whose education was mainly focused on music and who is still active as performer and composer, or Stephen Plaistow, who studied piano and violoncello at Bedales and then harpsichord at Cambridge (where he became president of the Cambridge University Music Club) and who is also still active as pianist, as well as writer and broadcaster. Also Joan Olive Chissell and David J. Fanning studied primarily piano (at the Royal College of Music and at the Royal Northern College of Music, respectively) and they were active as pianists, even if they complemented their studies with some scholarly education (Chissell studied theory and history of music and criticism, Fanning obtained a PhD in music). Bryce Morrison studied English literature and was a music scholar at the King's School of Canterbury, but at the same time he developed as a pianist under the guidance of Ronald Smith and Iso Ellinson in the UK and Alexander Uninsky in Texas, and he is currently active as pianist, piano pedagogue and jury member in international piano competitions.

Not all *Gramophone* critics studied music at a higher education level. Richard Osborne for instance, the most prolific reviewer concerning Beethoven's piano sonatas, read English at Bristol University. Already as a child he produced his first short stories (aged 12 he had a first short story broadcast in the BBC programme *The Children's Hour*) and in Bristol he won the two principal university prizes in literature (information taken from the 'Gramophone reviewers' sections of the *Gramophone* magazine).

DISCUSSION

The results of this analysis of critical review metadata raise several issues regarding the practice of recorded performance criticism and its relationship with the music recording market and music performance studies.

Agony of choice

There is a noticeable change in the repertoire reviewed over the last century. The distribution of sonatas seems to resemble a 'path to maturity' from early sonatas to Op.111. The increasing number of reviews of late sonatas in the later decades of the

century should be taken cautiously, but is intriguing. If this change cannot be explained away as a random phenomenon, various questions arise. Is this tendency reflected in the development of effective record production? If so, is it just a coincidence or does it mirror an effective shift in listeners', performers' and/or labels' preferences, taste and expectations? What role did criticism play in this shift? Does it make sense to claim that, as the performer needs to mature before approaching Beethoven's late sonatas, so does the listener? These and similar questions could be addressed from an historical and cultural perspective as well as from a psychological one, for example following Eliashberg and Shugan's (1997) dichotomy and investigating critics' role as influencers and predictors of listeners' preferences.

Such a study could move beyond issues of repertoire preference and examine preferences for particular interpretations. As the present investigation suggests, there are a large number of commercially available recordings from which listeners can choose. Alone, the *Gramophone* reviews cover 845 recordings, including 205 re-issues, produced by 216 different pianists. Since reviews published in this magazine presumably represent only a small selection of the recordings available on the market (consider, for example, the 23 pianists mentioned in the last section of this discussion who completed the recording of the Beethoven cycle and are not mentioned at all in the magazine), the amount of material seems impressively large. Already in 1951 Alec Robertson, reviewing Arrau's recording of the Moonlight sonata, complained that "we hardly needed another recording [of this piece]" (*Gramophone*, February 1951, p. 24). Since then the same sonata has been reviewed 176 times in the same magazine. And the *Gramophone* reviewer of 50 years Lionel Salter claims that this abundance of recordings puts an "intolerable strain" on the reviewer, when it comes to find "something fresh to say" about the n^{th} performance of the same piece (Pollard, 1998, p. 201).

Given this abundance of recorded material, it is legitimate to ask to what extent critics (and more so consumers) are able or have the necessary time, energy, and financial resources to distinguish between the many different interpretations and to appreciate their differences. Findings in decision-making research suggest that an increase in options (quantity of different versions of an item from which to choose) may paradoxically lead to paralysis of choice and dissatisfaction, even in the arts

(Schwartz, 2008). In this scenario the critic's guidance – working as filter of choice – becomes particularly significant. This is much more so since many critics tend to have long-lasting careers, writing for several years or even decades. And in the second half of the 20th century, they also became increasingly specialized in a specific repertoire and some of them have come to be acknowledged worldwide as authorities in their field (Pollard, 1998, p. 200).

On the other hand, this high level of familiarity with the repertoire and its diverse interpretations may influence critics' attitudes and preferences towards certain performances in ways different from lesser degrees of familiarity, likely to be found among the general public (Levinson, 1987, 2002, 2010). This in turn suggests that what may be considered a good performance by a listener – a good value-for-money recording – may not be considered thus by a critic who has a different level of musical expertise and listening history. Despite a conspicuous corpus of research addressing the influence of musical expertise on reliability and consistency of performance assessments (for a recent overview see Kinney, 2009), no study to date has investigated differences in the preference for one or the other interpretation between listeners with different levels of expertise and no study has taken into account the level and kind of expertise typically exhibited by music critics.

Comparative listening

A further observation that can be drawn from the findings is the weight given to the comparative element in reviewing practice. Comparisons between different interpretations/recordings emerged as a constitutive trait of *Gramophone* reviews, and editor Jolly supports this observation claiming that the comparative element is the "characteristic that has set *Gramophone's* reviews aside from its rivals" (Pollard, 1998, p. 202).

The importance given to comparative judgements in reviews is consistent with the large number of recordings of the same repertoire and the fact that reviewers tend to work over many years, searching for better understanding of how various interpretations differ from each other. However, comparisons in the present study tended to focus on only a small number of pianists. This, as Schick (1996) suggests, could be explained by the sheer number of recordings available, which forces critics to "compare a new release only with their past favourites, which makes the task more

practical but eternally rejects a slighted disk" (Schick, 1996, p. 157). In any case these results raise questions regarding the role of comparative judgement in music appreciation. In music research as well as in the academic context, with few exceptions, performance evaluation is explored through a criterion based assessment procedure – in which a performance is judged in isolation, set against a set of commonly agreed criteria – rather than through norm referenced assessment – in which a performance is assessed through comparison, as being better or worse than another performance (McPherson & Schubert, 2004). The importance that critical review seems to attribute to the comparative element however suggests that it could be useful to reconsider the extent to which listening to various interpretations is actually done, or can be done, in a criterion based way.

Subjective judgement

Beside the comparative element, a further aspect of reviews that invites a reflection on the nature of critical judgements is the increasing importance given to the identity of reviewers. Reviews were mostly unsigned at the beginning of the century. As the number of reviewers increased (by 1964 there were 16 reviewers at the *Gramophone*) signing reviews becomes usual habit, reflecting an increasing importance given to the possibility of tracing back a given review to its author. A new section is introduced in the magazine in January 1964, under the title "Portrait of a reviewer", in which each month one *Gramophone* reviewer is profiled and presented to the readership. At the beginning of the portrait a short introductory text explains the motivation and scope of the section:

> In the years just after the war there were four reviewers on THE GRAMOPHONE panel dealing with the entire output of the industry. Today, there are 16 names contributing to Analytical Notes and First Reviews alone. Many of these have been well known in music journalism for a long time; some, rightly, represent the younger school of critics. Whether old friends or new, we feel that readers may like to know something about the men behind all these initials.

The use of initials seems to suggest a familiar atmosphere, as if *Gramophone* critics were a sort of closed circle or, using Lionel Salter's words, a fraternity:

> [After 1950] More and more reviewers were coming aboard, their names appearing on the masthead after a preliminary run where their reviews were signed with their full names before being promoted to the fraternity, when they were reduced to initials. (Pollard, 1998, p. 200)

By the end of the 1990s there are almost 50 reviewers active at the *Gramophone* and at this point initials in reviews start to be replaced by full names. To understand the weight given to critics' identity it is necessary to reflect on the nature and scope of a *Gramophone* review. Discussing the relationship between reviewers and readers, and the way in which reviews 'function', editor Jolly stated:

> Gramophone's reviews operate on many different levels. The most basic is, of course, "Is this a performance I can live with?" Yet it is not as simple as that; the reader has to build a relationship with the reviewer, he (or sometimes she) has to know that what appeals to, say, RL will have the same appeal for himself. And over the years we all develop a special understanding. Indeed, such is the complexity of this relationship that it can operate on an even more sophisticated level – "I know that if RL doesn't like Maestro X's Sibelius then there's a chance that I may well like it myself". (Pollard, 1998, p. 202)

This emphasises how critical judgements are not objective statements but rather verbal expression of the critic's thoughts, triggered by his perception and filtered through and informed by his personality, knowledge and experience. As consequence, knowing the identity of the reviewer becomes a step central to a useful and meaningful interpretation of the review text. In academic or competitive contexts experts' assessments are usually treated as valid expressions of the value of the performance. These first findings however suggest that critical judgements are seen in a different perspective, as expressions of one (expert) person's impression of the music.

Re-issues

Finally, a reflection concerning the substantial presence of re-issues among the recordings reviewed. Almost one quarter of all the reviews found in *Gramophone* concern re-issued recordings. That fact raises questions regarding the criteria behind the process of selection as to what to review, as well as the nature of a re-issue itself and the objective behind the published review.

In the second half of the century, the growing recording market imposed the need for more stringent selection of the material to be reviewed. The choice of so many re-issues over new recordings could then be striking at first: why should *Gramophone* invest space in discussing performances already described and evaluated in previous years thus ignoring new, possibly great recordings? What is the purpose of re-reviewing the same performance? An answer to this question is inevitably multifaceted. Editor Jolly, describing how the process of selection of recordings changed overtime, claims:

> Today [1998] with some 400 discs arriving each month … decisions as to what to select for review are taken with the knowledge that every so often something superb is going to slip through the net. (James Jolly in Pollard, 1998, p. 203)

That suggests that quality (or assumed quality) is a criterion behind review selection. The choice of re-issues could then be seen as a way to reaffirm the value of an old recording over a new one (and of the magazine's decision to review it in the first place). But that alone cannot be a sufficient reason. Reviews of re-issues were evenly spread among long-lasting critics and other reviewers, suggesting that their presence is not due to seasoned critics' biases in terms of awareness and appreciation of older pianists. The quick growth in number of reviewed re-issues found between 1950s and 1970s can be ascribed firstly to the availability of new technologies, which explained the production of re-issues in the first place. However, the ground gaining movement of historical performance interpreters in those years might have also influenced this tendency, provoking critics to investigate the value of the new performance practice in relation to that of their mainstream counterparts.

The strong presence of re-issues suggests also that different issues of one and the same recording are considered to be two distinct sound objects. This could be understood in two ways. As said, the growth of re-issues reviews starting in 1951 may be explained by the technological innovations of subsequent decades: the introduction of the LP record by Decca in 1950, the following stereo recording in 1958 and later on, in 1983, the Philips/Sony digital recording and the Compact Disc (dates relate to the UK market; see Pollard, 1998). It is a truism to claim that the 78rpm version and the Long Playing or Compact Disc version of Schnabel's recordings of Beethoven's sonatas are not – aurally – the same object. Even within

the same format, different re-mastering processes by different engineers create a significantly different end product. This apparently obvious claim however poses a question regarding the scope of music recording reviews. Should reviewers comment on issues of recording quality?

The average listener views music recordings as portable concerts (Alessandri, 2011) without necessarily being aware of recording issues. If in a concert review we expect critics to discuss the work and its performance, in a recording review we would expect them to take into account a third aspect, namely, the recording *as* a recording. Critics are aware of the complex nature of sound recording and of the different contributions offered by performers, producers, engineers and technical resources, putting them in a unique position to review the recording as a whole. Of course it remains to be seen the extent to which this component enters the overall value judgement of the recording itself.

A second way in which a re-issue can be seen as a product other than its original release is what seems to be suggested by *Gramophone* editor Jolly when discussing the nature and purpose of a music review:

> T. S. Eliot argued that every time a new poem is written the entire canon of poetry is changed irreversibly and, similarly, every time a work is reinterpreted the entire history of that work is subtly altered. When Claudio Abbado records a new Bruckner Ninth, his version has to take its place not just alongside all the other versions with the Vienna Philharmonic, or all the versions that have been recorded by Deutsche Grammophon, but alongside every version that has ever found its way on to disc. (James Jolly in Pollard, 1998, p. 202)

New interpretations can shed light on the nature of older interpretations, and a critic's perspective and appreciation of a given performance can change overtime through exposition to different performances of the same or of other pieces. So for instance, Edward Greenfield reviewing Wilhelm Kempff's 80th birthday edition of Beethoven's sonatas and concertos claims:

> Of these sets the earliest is of the Beethoven piano concertos, first issued in 1962. The fantasy, the sense of joy bringing a smile to the lips, is what above all strikes me afresh on hearing these performances again. That is so even in No. 3, which I remember disappointed me slightly when I first reviewed it for these pages, slower and a little more staid than Kempff's earlier mono version (DG DGM18130, 12/55—now deleted). But in context with the others, the

> slower tempo for the first movement now seems no less convincing, the magic of Kempff wonderfully persuasive in the transition to the second subject for example. (November 1975, p. 151)

In this perspective, reviewing a re-issued recording becomes an occasion to approach an old recording anew and re-evaluate it in the light of other recordings produced so far; when appropriate, to re-affirm its value as interpretation and maybe also its increased value in terms of recording quality; and finally, to make the readership aware of its availability in a new, improved, format. In the light of these reflections, re-issued recordings seem to be objects different from their first releases, standing on their own and with their own right to be reviewed. Their substantial presence in the *Gramophone* material collected seems therefore to be justified.

Re-issues were also associated with the distribution of reviews among pianists. Amongst the 17 most often reviewed pianists are those who are usually acknowledged as great Beethoven interpreters, like Schnabel, Kempff and Brendel. With the exception of Richter, all pianists encompassed in this list completed the recording of all 32 sonatas. Within them are encompassed six of those eight performers who are the only pianists to have recorded the whole of Beethoven's cycle more than once in the course of their lives.[10]

The fact of having recorded more sonatas, even all sonatas more than once, could explain the high number of reviews those pianists received. However, along the century many other pianists accomplished the task of recording all 32 sonatas: by 2009 at least 64 pianists had completed or were in the process of completing the cycle (Alessandri, 2011). Here we have just a selection of 16 of them. Other performers who completed the cycle are mentioned in *Gramophone*, even if only a part (often a small one) of their cycle is reviewed, and 23 of those 64 performers[11] do

---

[10] i.e., Arrau, Backhaus, Brendel, Barenboim, Gulda, Kempff, in addition: Paul Badura-Skoda and Bernard Roberts. Information is taken from a previous discographical project on Beethoven's piano sonatas. See (Alessandri, 2011).

[11] Robert Benz, Muriel Chemin, Dino Ciani, Sequeira Costa, El Bacha Abdel Rahman, Maria Grinberg, Gotthard Kladetzky, Paul Komen, Michael Korstick, Christian Leotta, Michaël Lévinas, Seymour Lipkin, Andrea Lucchesini, Murray MacLachlan, Anne Øland, Georges Pludermacher, Akiyoshi Sako, Russel Sherman, Robert Silverman, David Allen Wehr, Gerard Willems, Yukio Yokoyama, Dieter Zechlin. Of course, this could at least partly be linked to the fact that *Gramophone* has been dealing to a largest extent with British releases. Unfortunately, a clear distinction between records available and records chosen for reviews is not possible due to the lack of comprehensive data on what records were issued in the UK in each given period. As indication however, out of the 23

not appear in the *Gramophone* pages at all. Hence the fact of having produced a high number of recordings of Beethoven's sonatas does not entirely explain the consistent presence of these pianists in the magazine reviews.

These performers did not just record Beethoven's sonatas: they produced recordings that, as it seems, passed the 'test of time'. Re-issues can be produced for marketing reasons, for instance to celebrate a specific circumstance (e.g., Kempff's 80th birthday) or to offer certain pieces in different couplings or groupings (e.g., complete cycle box set or, on the contrary, a choice of a few sonatas such as named or late sonatas). The high number of reviewed re-issues could then be seen as the music world's attention to and celebration of famous Beethoven pianists. However, as said during the past 90 years the main motivation behind the production of re-issues was arguably developments in the recording technology. If this is the case, recordings produced at the early stages of this developmental process were the ones that were candidates for later re-issues. In this perspective recordings produced in the 1980s or later seem to be twice disadvantaged in that the high quality level, durability and stability of the Compact Disc as a medium might have a direct consequence for the recording industry policy: re-issues are no longer needed. Once all great performances of the past will have been proposed in this new format it is difficult to see why a new release would be necessary (with the exceptions, mentioned above, of re-issues produced for marketing reasons).[12] Evidence for this can be seen in Figure 3.9 where we see a decrease in reviews of re-issues within the last decade.

Regarding the pianists reviewed in this corpus of critical texts, we might then consider who would now be at the top of our frequency table had Schnabel, Arrau or Kempff lived two generations later and recorded in the Compact Disc era, and had Ohlsson, O'Conor or Fu'Tsong recorded these pieces in the early stages of sound recording technology. It could also be asked: who would we now celebrate as great Beethoven interpreters?

---

cycles mentioned, 15 are currently available in Amazon.co.uk for purchase, 6 are available but only as import product, and 2 are not available.

[12] Of course, this claim assumes that with digital recording we have reached a kind of 'final stage' of recording quality, assumption that is – at least – highly arguable.

CONCLUSIONS

This chapter has provided an overview of the large sample of recorded performance critical review collected in the *Gramophone* archive, accompanied by reflections on the practice of criticism itself. Some of these reflections open questions that go far beyond the scope of this research and that call for further investigation. In general, the insights gained through the present analysis offer evidence of the potential that critical review has as a source of information and understanding for musical practice. More specifically, as for the purpose of this research, five main reflections emerged in this chapter that informed the analyses to follow:

i. A large corpus of material was collected, with 845 reviews of recordings of Beethoven's piano sonatas found in the British magazine *Gramophone*. The first picture emerging from this material is extremely varied with a high number of critics and pianists involved. However reviews were polarized around small groups of players and authors, thereby supporting the possibility and meaningfulness of an in-depth investigation of a selected sub-corpus of texts.

ii. The history of the *Gramophone* emerged as linked to and shaped by its critics. Different generations of reviewers coming from different musical and scholarly backgrounds succeeded at the *Gramophone* along the century, but most of them once arrived stayed for a long period, writing for the magazine for decades. The collected sample therefore witnesses the activity of a few long-lasting critics who covered the production of Beethoven's piano sonatas over many years. The variety of educational and historical backgrounds that critics bring together with the distribution of reviews and the reflection on the relationship between critic and reader mentioned above, suggest the necessity of taking into account the identity of the review authors when analysing the texts and also when looking chronologically for changes in the use of certain terms or expressions.

iii. Another important distinction found in the data is that between reviews of mixed and non-mixed repertoire. Preliminary observations suggest that reviews

discussing a more varied repertoire, instead of focussing on Beethoven's sonatas only, may represent a different kind of review product, that superficially describes the different recordings, making the reader aware of their availability, but without engaging in thorough critical considerations. As such, these texts may not be the best choice as examples of reviews to be used in a detailed investigation of critics' judgements.

iv. Comparison between different performances emerged as a constitutive activity of *Gramophone* reviews, and this element will need to be given further attention in the following chapters.

v. In general, questions arose regarding the role and purpose of a music review also linked to the nature of recorded music as opposed to live performance. In line with the construct of performance as event discussed in the first chapter, these first results reaffirm the necessity to embrace a wide perspective and – in order to understand what critics say with regard to the performance as artistic product – take into account also non-musical elements that may work as value parameters of a recording.

Based on these findings and reflections, the following chapter moves the investigation beyond the level of metadata and onto the textual domain, examining the content of critical reviews at first through a series of preliminary data reduction analyses.

4 GRAMOPHONE REVIEWS II: TURNING TO THE TEXT[13]

Chapter 3 has offered an overview of the collected critical review sample through an analysis of reviews metadata. This chapter moves the investigation beyond the level of metadata and enters the textual domain. As discussed in Chapter 2, a person-centred, interpretive approach represents the best solution for analysing the present corpus of material. The nature of this qualitative approach however makes it ill-suited to the analysis of large datasets. The analyses reported in this chapter were thus used as data reduction procedures, following the ATA approach (Guest et al., 2012, Namey et al. 2008), to frame the investigation allowing the selection from the initial, vast sample of reviews of a manageable and still representative corpus of material suited for the subsequent inductive thematic analyses.

Specifically, a five-step qualitative/quantitative data reduction procedure was applied to contain the music critical review sample on both inter- and intra-review level. This included:

(i) A thick-grained thematic analysis that produced a first categorization of the topics discussed in reviews;

(ii) An estimation of this categorization for the whole dataset;

(iii) A qualitative analysis of critics' vocabulary, with vocabulary organized in different semantic categories;

(iv) A comparison of the use of these categories between critics and in different periods;

(v) A comparison of word stems between critics and in different periods.

The present chapter reports methods and results of these preliminary analyses. It is divided in two parts that offer insights into (1) the main objects of discussion in

[13] Content within this chapter has been published within the following: Alessandri, Williamson, Eiholzer and Williamon, 2015; and Alessandri, Eiholzer, Cervino, Senn, and Williamon, 2011. For full references, see List of Publications.

reviews (analyses i and ii) and (2) the vocabulary used by critics, with focus on differences in word content between different groups of reviews (analyses iii, iv, and v). Drawing from the findings, in the final part of the chapter a selection of reviews is produced that will serve for the subsequent thematic analyses.

WHAT ARE REVIEWS ABOUT?

Introduction

The first acquaintance with the textual content of reviews was gained through an analysis that produced a thick-grained categorization of what reviews are about. The construct of performance as event discussed in Chapter 1 – an event in which elements like sounds, agent, work, audience and context play essential parts – applies also to recordings. Therefore, as argued in Chapter 3, it can be expected that listening to a music recording implies listening to at least three different objects: the work being performed, the performance of that work, and the way that performance is transferred and conveyed through the recording technology and medium. If that is the case, it would then be expected to find these three objects reflected in the review texts.

In Chapter 1 also it was stressed how the development of music performance criticism in the course of the twentieth century was linked to the establishment of a canon repertoire and of the status of the performer as interpreter. In line with this, and with the increase of reviews of re-issues and first releases of old recordings found in Chapter 3, it could be anticipated that the second of these objects – the way the work is performed and interpreted by the pianist – will be given increasing importance in critical review in later decades of the century.

To test for these assumptions and to gain insights into the weight the different objects have in critics' writings, a two-step hybrid qualitative/quantitative analysis was run, to test two hypotheses:

(i) The larger part of review texts has as its object the following three topics: work performed, performance, and recording.

(ii) The portion of text devoted to the discussion of interpretative issues increases along the century.

In the following sections the analysis procedure is described and the obtained results are presented and discussed.

Analysis (i): Qualitative analysis of reviews content

As mentioned in Chapter 3, for six out of the 845 collected reviews the text reported in the *Gramophone* archive was damaged, so that these could not be used for text analyses. Thus all analyses reported in this chapter refer to only 839 reviews.

Method

A subset of 63 reviews was chosen from the total 839 reviews. This included seven reviews per decade randomly chosen among reviews that

(1) concerned solely Beethoven piano sonata(s) and

(2) had a text length of between 130 and 800 words, to assure having enough text with which to work and to exclude long, article-like reviews that offered a different journalistic product.

Reviews were analysed, and different content sections were hand-coded in the text according to the following three categories: *performance*, *composition*, and *recording*. Categories were understood as mutually exclusive and segmentation was done at sentence level, to allow comparison with the results of the subsequent analysis (analysis ii). A fourth category was left open (labelled 'other') to allow for other, unexpected interesting topics to emerge. The definitions of the codes used in the analysis are reported in Table 4.1. This process was repeated independently by one more researcher (with professional training in piano performance) for a selection of the material (n = 15 reviews), and inter-coder reliability was computed using Cohen's Kappa.

Table 4.1. Code definitions for thick-grained content analysis.

CODE	*Definition*
Performance	Descriptive and evaluative claims about the way in which the musical work is realized, with or without reference to the performance as interpretation of the work.
Composition	General information about the work, work description, context of composition, biographical information about the composer.
Recording	Statements about the recording context, process, quality (e.g., sound quality due to recording procedure/material, use of certain technology, distribution of sonatas through different discs/records).
Other	--

Results

Inter-coder reliability for the analysis of a selection of reviews was found to be Kappa = .84 ($p < .001$), 95% CI (.76 – .91), which represents an almost perfect agreement between coders (Landis & Koch, 1977).

From the text analysis run on the subset of 63 reviews, the categories *performance*, *composition* and *recording* accounted for 79.19% of all text, on average. They were distributed as follows: 54.69% *performance*, 9.99% *composition*, and 11.59% *recording*. The distribution of categories across decades is reported in Figure 4.1. The amount of text given to the category *performance* increased over time. The years 1961-70 do not belong to this picture, showing a low percentage of *performance* related text, accompanied by a peak in the category *recording* and a low percentage of other, non-classified text. The category *composition* plays a major role at the beginning of the century and decreases toward 1960. After this date, almost no text at all is devoted to it.

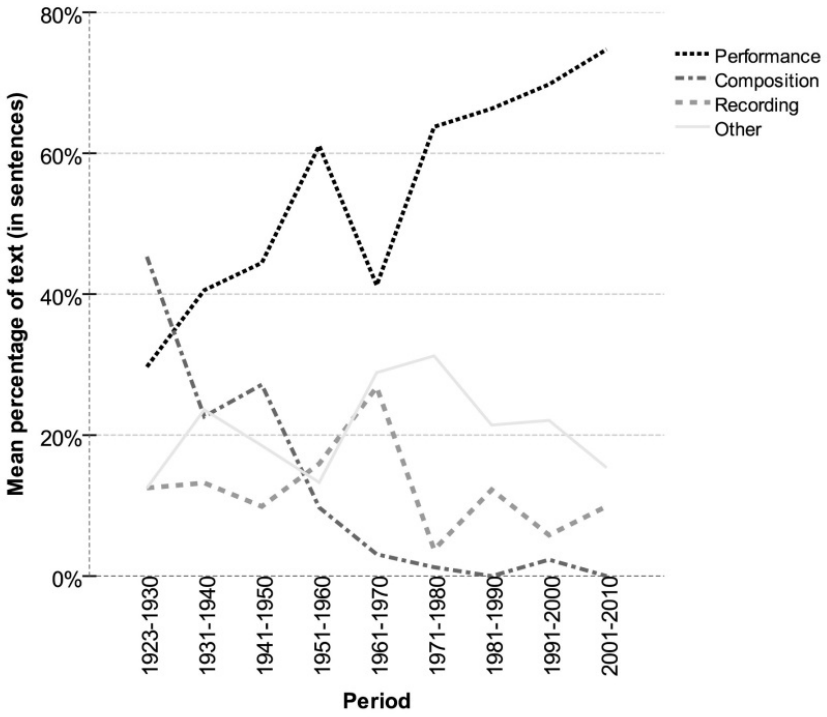


Figure 4.1. Distribution of text according to the four content categories *performance*, *composition*, *recording*, and *other* across decades.

Pearson's Chi-Square was computed using the first and last 251 sentences of the analysed set of reviews (corresponding to the periods: April 1923-December 1950 and April 1986-September 2010). Results showed a significant association between period of publication and amount of text devoted to the different categories (χ^2 (3, N = 502) = 95.90, p < .001, Phi = 0.44, p < .001). Descriptives of the Chi-Square are reported in Table 4.2.

Table 4.2. Contingency table of text categories in the first and last decades of the observed period.

		Category				*Total*
		Performance	Composition	Recording	Other	
1923-1950	Count	98	75	30	48	251
	Expected	139.5	38.5	25.0	48.0	251.0
1986-2010	Count	181	2	20	48	251
	Expected	139.5	38.5	25.0	48.0	251.0
Total	Count	279	77	50	96	502
	Expected	279.0	77.0	50.0	96.0	502.0

The three categories *performance*, *composition* and *recording* accounted for almost 80% of the critical text. The residual 20% was labelled as 'other'. Here information is enclosed about pianist (e.g., biographical information, comments on the pianist's general attitudes, characteristics, habits), content of the booklet accompanying the recording, availability on the market of other recordings of the same piece or by the same pianist, as well as general comments on the personal knowledge or experience of the critic (e.g., what other recordings of that sonata the reviewer has listened to, or reviewer's personal tastes and preferences). In addition, statements were found where the critic explicitly commends or not the recording to the reader, or where s/he reflects on what criticism ought or ought not to be and how it should or should not be done.

Analysis (ii): Estimation of content categories for the whole dataset

Method

In a second step, the hand-coded reviews were used as input data to estimate category proportions across the whole dataset using the R package ReadMe recently developed at the Institute for Quantitative Social Science, Harvard University

(Hopkins & King, 2010). As discussed in Chapter 2, ReadMe is a text mining application that employs learning machine procedures to estimate the distribution of categories among texts based on a series of pre-coded documents. The version used in this analysis was 0.99835, available from http://GKing.Harvard.Edu/readme under the Creative Commons Attribution-Noncommercial-No DerivativeWorks 3.0 License.

For this analysis, sentences – understood as groups of words separated by dots – were taken as semantic unit. Reviews were pre-prepared by taking out all dots that did not signal the end of a sentence (e.g., Op.; No.; E.M.I.; H.M.V.; Mr.; Dr.; etc.). Subsequently, reviews were split in single-sentence documents using an application developed in Microsoft Visual Basic ® 6.0[14]; this resulted in 13,328 documents. These documents were divided into two sets:

- a training set, containing sentences belonging to the 63 pre-coded reviews (n = 817 documents). This set was used by ReadMe to 'learn' how to categorize texts; and
- a test set entailing all other review sentences, for which ReadMe needed to produce an estimate of the content categories.

Hopkins and King (2010) in their validation tasks found that a training set entailing between 100 and 500 pre-coded documents could produce estimates with a root mean square error (RMSe) between 3 and 1.5 percentage points. However, these tests were done analysing valence of political opinions in blogs – that is, comparing the use of positively or negatively loaded words in the texts. It was expected that analysing *Gramophone* reviews according to the labels described in Table 4.1 would pose a more demanding challenge to the ReadMe algorithm than the one reported in the original validation procedures in terms of semantic understanding of the texts required to distinguish between one and the other category. Consider the sentence:

> Beethoven's Op 49 consists of two little sonatas – sonatinas really – of which this is the second; the next sonata he wrote was Op 53, the Waldstein, and the difference in style is remarkable. (Dora Labbette, April 1924, p. 21)

[14] Application developed by Simone Alessandri.

Here the reviewer is clearly discussing the sonata recorded, and not the interpretation of it offered by the pianist. Cues to guess the category correctly may be the fact that no pianist is named, as well as the presence of words like "Beethoven's", "wrote" and "consists". Similarly, take the sentence:

> Imagine Friedrich Gulda's hard-hitting sonority and dry-point articulation welded to Wilhelm Kempff's clipped phrasing and intimate dimensions, and you'll get a general sense of Ciccolini's detail rather than bigpicture-oriented aesthetic. (Jed Distler, April 2007, p. 82)

Here the names of three pianists and words like "sonority", "articulation" and "phrasing" could allow for the attribution of the sentence to the category "performance". However, sentences like the following represent a harder task:

> The expressive adagio that twice interrupts the almost gay trivace certainly does not seem to me like a stab in the heart, though a moving into shadow. (Alec Robertson, February 1937, p. 19)

Here the reviewer is describing the third movement of Op. 109. That he is talking about the work and not its performance becomes clear when reading on in the text; after this paragraph comes to an end, the review moves on to the performance level, starting from the description of the first two movements:

> Of the theme of the six variations one can certainly agree that it is one of Beethoven's loveliest tunes.
>
> Kempff gives an almost completely successful performance. He balances excellently the contrast between the vivace and adagio of the first movement, and the playing of the part-writing in the prestissimo is beautifully clear.

However, the previous sentence taken on its own would be difficult or impossible to categorize even for a human coder, since without contextual support it cannot be excluded that the reviewer is talking of how the Adagio was performed – and not how it is composed. Examples of this kind were found copiously spread across the 63 reviews analysed. Therefore, it was expected that the margin of error in ReadMe estimates for reviews would be higher than the one found by Hopkins and King (2010) in their analyses. To check for this, a validation of the method was run splitting the training set in two and using 500 documents as the validation training set

and 317 as the validation test set. Estimated and actual values for the four categories are reported in Table 4.3.

As expected, RMSe was higher than that reported by Hopkins and King (2010), at 3.56 percentage points. Mann-Whitney test was however non-significant, U = 7.000, p = .44 (exact Sig. 1-tailed). As Hopkins and King (2010) affirm, a RMSe value around 3 percentage points can still be acceptable (p. 241), however, this requires a more cautious interpretation of the estimates produced.

The initial intention of this analysis was to estimate content categories for the whole dataset and for each decade separately. However, the high RMSe found in the validation stage suggested that a single decade analysis would be too finely tuned, requiring a level of accuracy in the software estimates that could not be assumed in this case. Therefore, besides looking at the whole set of reviews, only two more analyses were run on reviews published in the periods 1923 – 1950 and 1991 – 2010 respectively, to compare estimates between the two extremes of the century[15].

Table 4.3. Distribution of content categories as they were estimated by ReadMe and as they resulted from the hand-coding (validation task).

Category	*Estimated*	*Actual*
Performance	53.98 %	51.41 %
Composition	10.91%	16.93 %
Recording	14.61%	11.91 %
Other	20.49 %	19.75 %

Note. Values refer to the set of 317 documents used as test set in the validation procedure.

Results

The aim of the computerized ReadMe analysis was to determine the extent to which the content analysis results of the 63-reviews subset could be taken as representative of the whole 839 collected reviews. In the whole dataset, the three categories *performance*, *composition* and *recording* accounted for 79.32% of reviews text. Of this, 53.50% was attributed to *performance*, 9.09% to *composition* and 16.73% to *recording*. The totality of text covered by these three categories remains constant along the century, but proportions between categories are different for the estimates

[15] The choice of a longer period (1923-1950) at the beginning of the century is due to the small quantity of reviews available in those years.

run on the first and last decades of the century, with a much lower percentage of text devoted to *performance* in the period 1923 – 1950 (36.38%).

Percentages emerged in the ReadMe and in the hand-coding analyses are given in Table 4.4. Correlation between the two sets of values was strong (r_{12} = .91, p < .001). However a few large discrepancies were found that are highlighted in Figure 4.2. Differences can be ascribed to two factors: the margin of error of software estimates and the natural variance within the material (the hand-coding analysis was run on a selection of 63 out of the 839 reviews, while ReadMe estimates were run on the residual 776 reviews). In line with the results of the validation test the largest discrepancies are found in the categories *composition* and *recording*.

Table 4.4. Distribution of content categories as they emerged in the analyses run on the 63-review sample (hand-coded) and on the whole dataset (ReadMe estimates).

Decades	*Category*	*ReadMe estimates*	*63-reviews analysis*
1923-2010	Performance	53.50	54.69
	Composition	9.09	9.99
	Recording	16.73	11.59
1923-1950	Performance	36.38	39.04
	Composition	11.90	29.88
	Recording	28.83	11.95
1991-2010	Performance	60.17	72.47
	Composition	9.02	1.12
	Recording	11.82	7.87

This is true particularly in the period 1923 – 1950. A possible explanation is that technical terms and numbers largely used to indicate the different discs – typical of the beginning of the century – may have led to a higher proportion of sentences being labelled as *recording* in the computerized automatic analysis. The other large difference seems to be in the category performance for the period 1991 – 2010. According to the ReadMe estimates, the category performance covers a 60.17% of the text in this period, opposed to the more than 70% found in the 63-reviews set. However, even with this lower percentage, Pearson's Chi-Square showed a significant difference between distribution of categories in the earlier and later decades (χ^2 (3, N = 200) = 14.316, p < .005, Phi = 0.27, p < .005). Results of this analysis are thus in line with the findings of the qualitative analysis run on the 63

reviews. Together, they support the hypothesis that in the course of the century the discussion of interpretative issues is given increasingly more space in reviews.

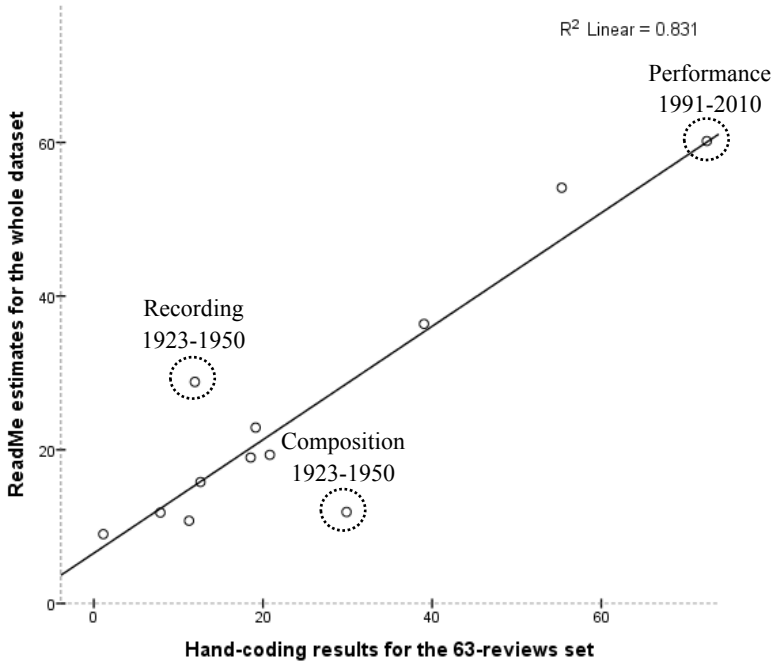


Figure 4.2. Scatter plot displaying ReadMe estimates for the whole dataset against hand-coding results for the 63-reviews set. The dotted circles indicate categories for which the two analyses gave discrepant results.

Conclusions

These first two analyses produced a thick-grained categorization of the topics discussed in *Gramophone* reviews. A small sample of texts was made object of qualitative thematic analysis. Findings of this investigation were substantiated by the estimates computed through automatic software analysis. Both initial hypotheses were confirmed by these findings: the three main objects of discussion in *Gramophone* reviews of Beethoven's piano sonatas are the performance, the composition performed, and the recording process, medium and quality. These three topics cover about two thirds of the critical review texts. Results also reflect an increasing focus on interpretative issues over the course of the century, with later reviews providing more text on performance. This is in line with the growing quantity of reviews of reissues (from the 1950s) and releases of old recordings (from the 1980s) found in the full dataset and discussed in Chapter 3.

For what concerns the subsequent qualitative investigation, the results of these first two data reduction analyses led to two conclusions: firstly, given the large sections of text devoted to the discussion of interpretative issues, a relatively small number of reviews would suffice to obtain an overview of the content of critics' writings. Secondly, in order to facilitate a systematic investigation of the texts, it was decided to structure the thematic analysis in two distinct layers, focusing at first on *performance* features discussed by critics, and moving then onto the residual elements of the end-product recording (*composition*, *recording*, *other*).

CRITICS' VOCABULARY

Introduction

The second set of data reduction techniques addressed the vocabulary used by critics. Limitations of a word-level analysis of texts, linked to the impossibility to account for the context-dependency of terms, have been discussed in Chapter 2. Despite these limitations an exploration of words used can be helpful in framing subsequent analyses.

The aim of this research was to obtain an understanding of the content of critical review of Beethoven's piano sonata recordings that could be as focused as possible, that is, void of spurious influences from, for example, other repertoire, and as comprehensive as possible. Although the spread of reviews across several decades suggests the possibility of observing diachronic changes in critical practice, the overview of collected material in Chapter 3 evidenced a strong polarization of reviews around a few authors. Following this observation, a first goal of data reduction analyses (iii) to (v) was to figure out if a selection of reviews for in-depth study should account more for diversity of historical periods or, instead, critics. To address this question differences in the use of words by different critics in different periods were examined. It was thought that discrepancies between groups would suggest a possible variety of topics discussed, and the consequent necessity of accounting for this variety in the thematic analysis, in terms of selection of material.

A second question raised in Chapter 3 concerned the possibility of including or not in the analysis reviews of mixed repertoire – that is, those discussing works other than Beethoven's sonatas. Here as well, it was thought that an examination of

differences in selected categories of words could help answer the question. In this case, substantial discrepancies in the use of words between mixed and non-mixed reviews would rather suggest the necessity of excluding mixed reviews from subsequent analyses. Lastly, a third goal of this preliminary vocabulary examination was to offer some awareness of what may be topics discussed by critics, thus giving some first guidance in the development of codebooks for thematic analysis, as suggested by Guest et al. (2012, chapter 5).

Analysis (iii): Qualitative analysis of critics' vocabulary

Method

A vocabulary of the critical review sample was compiled using the word cruncher function of the software Atlas.ti version 6.1. This resulted in a list of 17,340 word types. This list was reduced by (i) narrowing the analysis to words that occurred more than 5 times in the whole dataset; and (ii) sorting out function words and proper names. The remaining 2,503 word types were analysed by the author and grouped according to different semantic fields.

Results

The main semantic categories emerged are listed in Table 4.5 together with the mean percentage of text (in words) per review covered by each category. The grouping of words was done bottom-up, letting the different categories emerge from the list. However the Reasoning Model, as it was described in Chapter 1 (Beardsley, 1982), was taken as background to this analysis and it informed the development of categories. Thus, understanding critical reviews as a form of evaluation grounded in reasons, we expected to find three main groups of words that may be informative of the way critics construct their judgements:

- *Purely evaluative terms* (e.g., good, bad, better, great, awful), that is, words that comprise only an evaluative component, but offer no description of the object being evaluated (see Bonzon, 2009);
- Words that express a *reasoning process* on the side of the critic (e.g., despite, consequently, implies, justify, reason, considering, and yet, thus, therefore); and

- All words that might offer some kind of *reason* for the judgement to be made.

Table 4.5. Main semantic categories (in words per review) emerged from the analysis of critics' vocabulary.

Main Category	*Sub-category*	*Percentage (SD)*
Purely evaluative terms		1.64 (1.09)
Achievement		0.82 (0.76)
Reasoning process		1.18 (0.71)
Musical parameters	Tempo & Rhythm	0.66 (0.62)
	Sound Quality	0.48 (0.48)
	Dynamics	0.38 (0.43)
	Phrasing & Articulation	0.16 (0.27)
	Pedal	0.06 (0.16)
Structural parts		1.56 (1.10)
Expression		2.85 (1.17)
Originality & Insights		0.71 (0.55)
Intensity		0.65 (0.61)
Ideal interpretation		0.59 (0.51)
Unity		0.44 (0.43)
Accuracy		0.37 (0.44)
Variety		0.37 (0.40)
Clarity		0.32 (0.40)

The latter represents a wide group of words, and the most challenging in terms of letting categories emerge without imposing pre-conceived ideas on them. A broad distinction that can be made is between words related to specific *musical parameters* (e.g., tempo, rhythm, dynamics), indications of specific *structural details* of the work (e.g., measure, scale, triplet) and qualities of the performance that may provide reasons for evaluation. In this latter group eight main categories emerged: expression, accuracy, clarity, originality and insights, ideal interpretation, variety, unity, and intensity. These are briefly discussed below.

Expression. Expression is probably one of the most important terms in musical parlance when it comes to discussing performance evaluation and appreciation. Hence, it was not surprising to find a large portion of critics' terms

falling under this label. However, during the course of the analysis a difficulty emerged in deciding what terms should belong in the category of 'expression'. The necessity of defining the boundaries of this category raised a question concerning the notion of expression that should be embraced in this study.

Despite the ubiquitous presence of the word 'expression' and related forms in musical discourse, there is no unanimity regarding what 'expression' in music means. In this first analysis, words tagged as 'expression' related to disparate constructs: expression of the character of the piece, of emotions or psychological states; elicitation of emotions or psychological states in the listener; and communication of features of the piece. This was based on the author's own understanding of 'expression' based on her experience as musician. Two problems however emerged: first, the boundaries of this notion of 'expression' are vague and slippery, thus in need of empirical evidence to help delineate them more precisely. Second, it is unclear to what extent this understanding reflects what critics mean when they talk of 'expression'. Given the importance of this notion in music performance, the questions raised from this analysis suggested that a focused examination of the notion of 'expression' in critical review would be necessary before embarking on subsequent studies.

Accuracy. Here are terms that refer to precision and exactitude in performing. It relates to the impression that the performer has control over what s/he is doing but also that s/he takes care of details and nuances, and s/he has the necessary sensitivity to do it.

Clarity. This entails terms related to clarity or lack of it. Clarity in performance may refer at least to two different constructs: technical clarity (linked to the pianist's technical proficiency, but also to sound quality that may relate to the recording process) and interpretive clarity – that is, clarity in conveying the structure of a musical piece. Disambiguation between these constructs will only be possible through a qualitative text analysis.

Ideal interpretation. This is a wide, and difficult to bound, group of words, that relates to the ideas of stylistic appropriateness, and the correlated notion of the existence of a true, or correct interpretation of Beethoven's sonatas against which performances should be set.

Originality and insights. One of the main challenges given to performers is the request of being faithful to the score and the true character of the piece and yet able to offer original and insightful readings of the work. In this category thus enter terms that characterise the performance as being innovative, insightful, or unique as opposed to conventional or predictable. While the word categories mentioned so far represent evaluation criteria that can be found in academic contexts (see discussion on evaluation criteria in chapter 1, see also McPherson & Schubert, 2004), this group of terms would be expected to play particularly important parts in the evaluation of professional performances.

Intensity, variety, unity. Three more categories that emerged from the analysis contain terms that refer to degree of intensity, variety, and unity and coherence. The fact that this triad corresponds to Beardsley's proposed theory of general principles of aesthetic value could point to a bias of the author in this analysis. Despite the attempt to avoid impositions coming from existing literature, it is not possible to exclude an influence. However, the general applicability of these categories is straightforward: praising the wide dynamic range of a performer, the delicacy of a pianissimo, the vigour conveyed by a straight tempo or through an explosive sforzando; admiring the elegant proportion of different dynamic levels or on the contrary criticising a mercurial interpretation for its lack of coherence in its use of phrasing or articulation are just a few examples of how terms related to these three categories can permeate performance criticism.

In addition to the word categories mentioned so far, one more semantic category emerged is that of *achievement*. Terms like 'triumph' and 'success' have a clear evaluative connotation, and in fact they could be seen as being close to the category of purely evaluative terms. But they differ from the latter in that they entail what seems to be a reason for the positive verdict: the accomplishment of a difficult task. It is not clear at this point what role the notion of achievement may play in critics' writings, however, the idea that the evaluation of an artistic product is – at least partly – a measure of its success understood as the artist's achievement is a thesis quite debated in philosophy of music and recently

defended by Carroll (2009). According to Carroll, our perception of an artistic product as the artist's achievement is fundamental to the appreciation of art.

Analysis (iv): Comparison of semantic categories

Method

In a second step, the semantic categories emerged from the qualitative analysis of critics' vocabulary were used to create a personalized vocabulary that was uploaded to the software Language Inquiry and Word Count (LIWC 2007 by Pennebaker, Booth, & Francis, already discussed in Chapter 2) and used to compute frequency between groups of reviews. A few combinations of words were added in the LIWC vocabulary to disambiguate the meaning of specific terms (e.g., 'sound quality', 'recording quality', 'composer's intensions'). The complete vocabulary used with LIWC is reported in Appendix 3. Frequency rates of each semantic category for each review were computed and descriptive and exploratory data analyses were carried out to test for associations between decade, critic, and repertoire reviewed and the occurrence of different semantic categories. Purpose of this exploration was to understand what selection of the 839 collected reviews, if any, could offer a more informative and rich source of material for the subsequent thematic analyses.

In all analyses the first three decades (1923 – 1950) were combined to obtain more similar sample sizes. Measured frequencies were non-normally distributed; therefore median values are reported in the following graphs instead of mean values as measure of central tendency. Moreover, since data did not satisfy parametric assumptions, a multivariate analysis, able to construe a model that would account for the influence of critic, period, and repertoire reviewed at once could not be run. Instead, single Kruskal-Wallis tests were carried out to observe differences in the use of semantic categories between critics, periods and repertoire separately. This approach however does not take into account the interrelation between factors, and as a consequence, it increases the probability of Type I error, creating false positives in the results. Therefore in a second step reviews were split into sub-groups (by periods and critics) and Kruskal-Wallis tests were rerun on those samples.

Results

Critics versus Historical period

Having reviews spread along almost 90 years of recording history, it seems natural that a selection of material for an in-depth analysis should entail samples of different decades to account for developments or transformations of evaluation criteria in the course of the century. Comparison of word categories between decades shows indeed significant differences in all semantic categories observed, with the exception of clarity and ideal interpretation (Table 4.6).

Table 4.6. Kruskal-Wallis tests, independent variable: decade.

Semantic category	H_6	*Sig.*
Purely evaluative terms	61.926	.000
Reasoning process	33.146	.000
Achievement	24.215	.000
Musical parameters	26.161	.000
Structural details	14.393	.026
Expression	35.826	.000
Originality and insights	18.245	.006
Intensity	18.884	.004
Ideal interpretation	8.568	.199
Unity	22.500	.002
Accuracy	18.430	.005
Variety	47.592	.000
Clarity	2.376	.882

These results cannot be interpreted, however, without considering what other factors may account for these fluctuations. In particular, what could have reasonably influenced the choice of words most is the author of the reviews. As shown in Chapter 3, distribution of reviews among critics is strongly polarized, so that as few as 10 critics out of 52 wrote more than 60% of the whole corpus of texts. Critics like Morrison or Plaistow published reviews continuously in the *Gramophone* for over four decades. Systematic differences in writing styles of one or the other critic may, therefore, result in ample variations of frequency rates between decades. And indeed, Kruskal-Wallis tests show significant differences between critics in all semantic

categories except clarity (Table 4.7). Averaged across all 13 categories Kruskal-Wallis was $H_6 = 25.56$, $p = .086$ between decades, and $H_{10} = 52.30$, $p = .037$ between critics, suggesting a higher level of variability between critics.

Table 4.7. Kruskal-Wallis tests, independent variable: critic.

Semantic category	H_{10}	*Sig.*
Purely evaluative terms	80.275	.000
Reasoning process	38.333	.000
Achievement	57.336	.000
Musical parameters	152.105	.000
Structural details	90.121	.000
Expression	47.563	.000
Originality and insights	36.185	.000
Intensity	26.029	.004
Ideal interpretation	19.922	.030
Unity	29.310	.001
Accuracy	24.371	.007
Variety	68.259	.000
Clarity	10.082	.433

This seems to be supported by the visual exploration of distribution of categories along the century (Figure 4.3) and across critics (Figure 4.4). Strong differences can be observed in both graphs, however, it is striking that fluctuations between one and the other critic are in some cases extreme, also between critics writing in the same period (e.g., Morrison and Distler). Explicit references to musical parameters and specific passages or moments within the piece (structural details) are copiously used by Distler and Chissell, but almost absent in Morrison's writings. Morrison is also the critic using achievement related words the most, and, as seen in Chapter 3, almost 28% of reviews published in the 1990s and 2000s were written by him. This could suggest that the increase in achievement terms observed in Figure 4.3 reflects at least partly the high amount of reviews written in the last decades by Morrison, and not a general trend of *Gramophone* reviewers to focus on the construct of achievement.

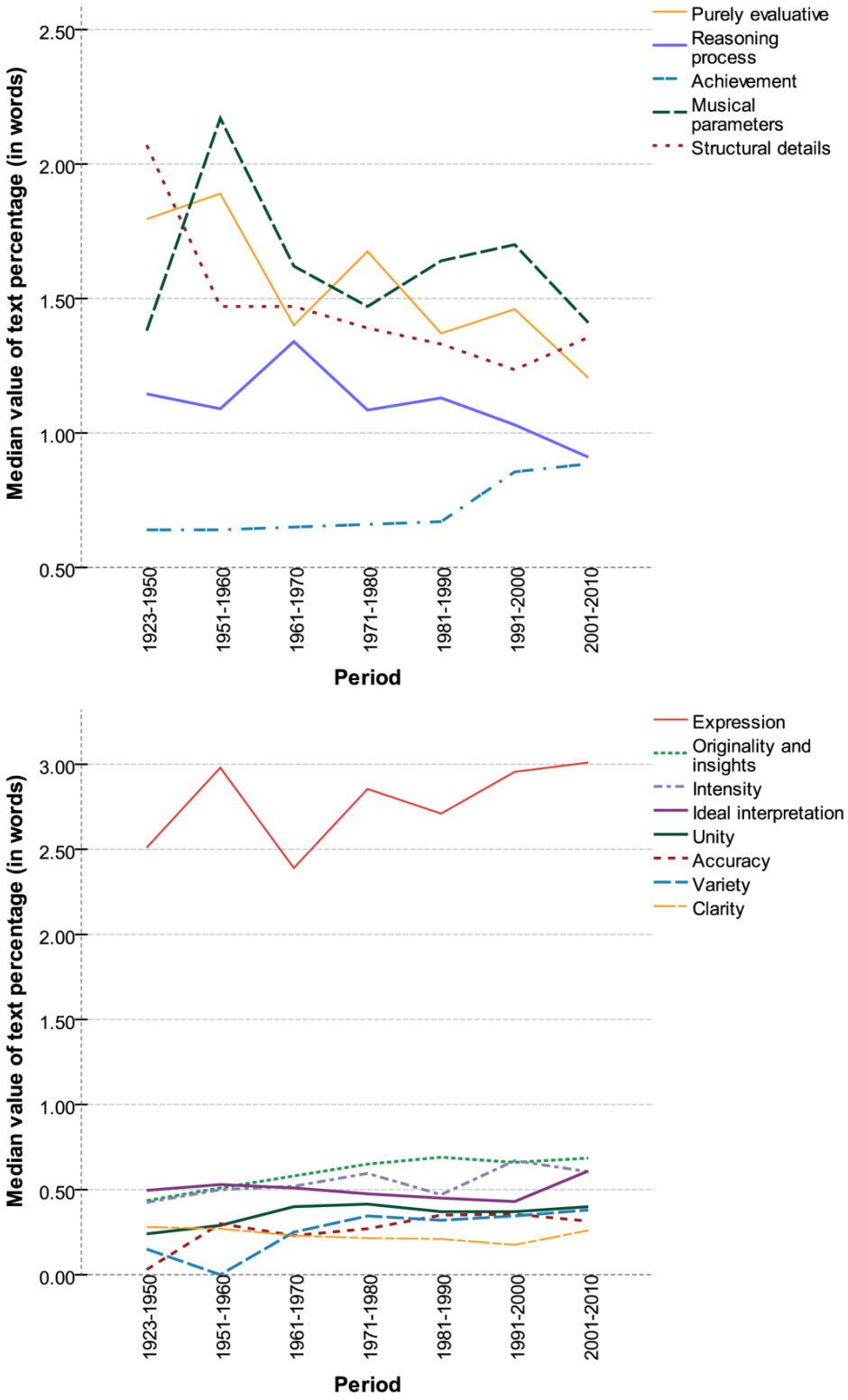


Figure 4.3. Median frequency rates of word semantic categories across decades.

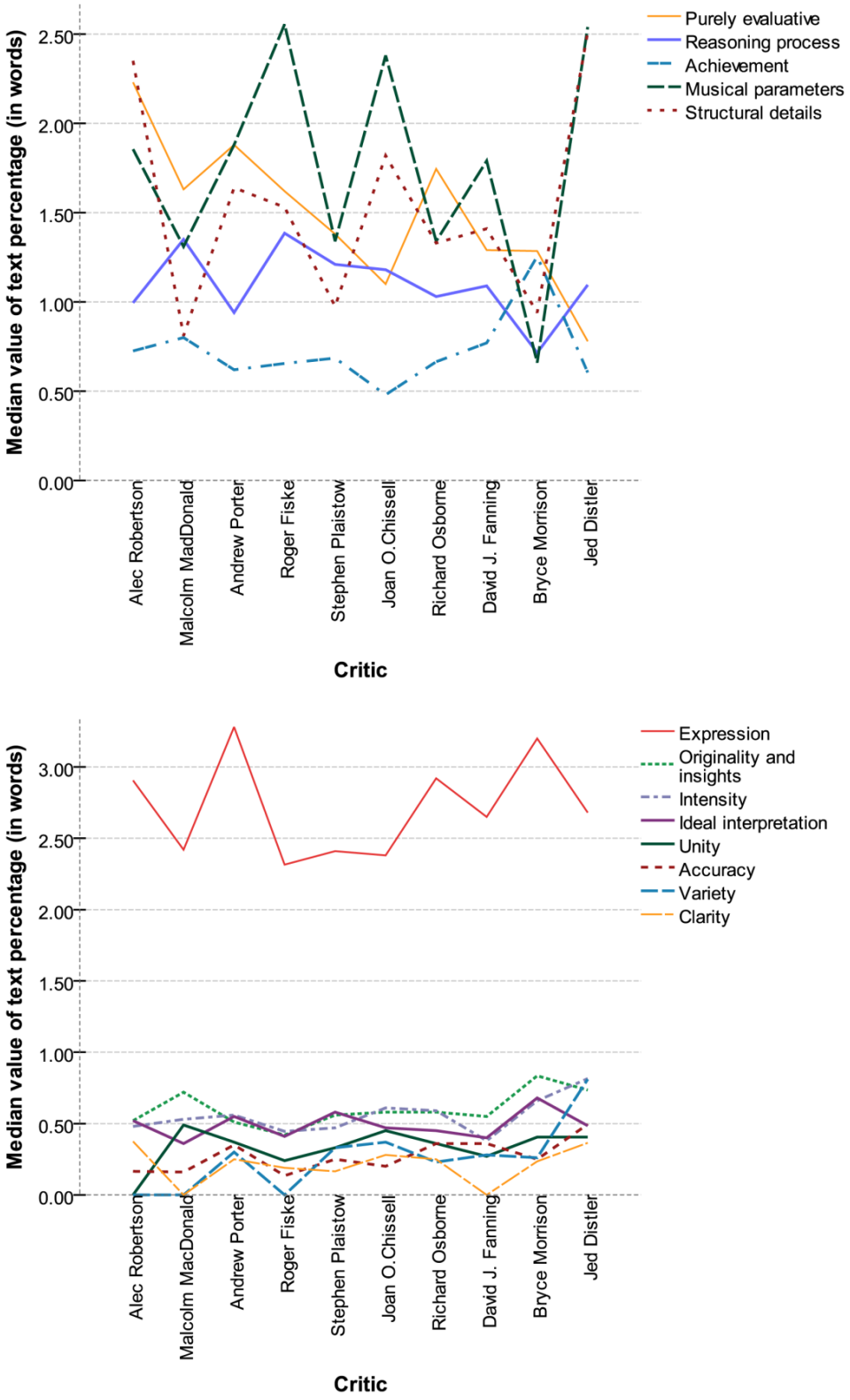


Figure 4.4. Median frequency rates of word semantic categories across critics. Critics are ordered chronologically according to the publication date of their reviews (10 most prolific critics listed in Chapter 3).

Table 4.8. Differences in the distribution of semantic categories between critics (left panel) and between decades (right panel), tested by splitting the reviews corpus accordingly.

Split for decade			**Split for critic**		
Period	*'Critic' significant for*	*Statistic*	*Critic*	*'Period' significant for*	*Statistic*
2001-10	Musical parameters**	H_5 56.100	Osborne	Expression**	H_3 12.366
	Structural detail**	H_5 32.057		Structural details*	H_3 9.420
	Originality and insights*	H_5 15.363			
	Achievement*	H_5 18.344			
	Variety**	H_5 29.538			
	Intensity*	H_5 15.661			
1991-00	Musical parameters*	H_5 *17.196*	Plaistow	--	
	Structural detail*	H_5 17.051			
	Expression**	H_5 23.805			
	Achievement**	H_5 22.345			
	Intensity*	H_5 17.345			
	Unity*	H_5 11.887			
	Variety*	H_5 15.611			
1981-90	Evaluative*	H_6 18.200	Chissell	--	
	Musical parameters**	H_6 27.280			
	Structural detail*	H_6 13.627			
1971-80	Evaluative*	H_5 17.425	Morrison	--	
	Musical parameters**	H_5 24.107			
	Structural detail**	H_5 25.839			
	Originality and insights*	H_5 13.156			
	Clarity*	H_5 18.135			
	Variety*	H_5 12.033			
	Unity*	H_5 12.239			
1961-70	Musical parameters*	H_4 19.937	Fiske	Variety*	H_3 7.982
	Structural details**	H_4 25.749			
	Unity*	H_4 9.778			

Note. Only significant results are reported. ** Significant at p < .05; ** Significant at p < .001*

Splitting the corpus of reviews into sub-groups and testing differences between critics and in different decades separately further supports the initial findings, suggesting larger discrepancies in the use of words between critics than between periods. Results of this split analyses are summarised in Table 4.8. To assure large enough review samples, differences of word use between critics were tested for the periods 1961 – 1970, 1971 – 1980, 1981 – 1990, 1991 – 2000, and 2001 – 2010. For critics, only the first five most prolific reviewers were analysed (who wrote min. 50 reviews each): Osborne's, Plaistow's, Chissell's and Fiske's activity spread over four decades, Morrison's over two.

Use of musical parameters and structural details between critics differed in each decade, together with one or more other categories. Repeating the analysis splitting reviews by critic and testing for period influence, no difference could be found between reviews produced in different decades, with the only exception of structural details and expression words used by Osborne and variety related terms used by Fiske.

Repertoire reviewed: mixed versus non-mixed reviews

One more observation emerged in Chapter 3 concerned the presence, within the corpus of critical review, of texts in which Beethoven's sonatas are discussed together with other pieces, by Beethoven or by other composers. The question was raised if it would be reasonable to exclude these mixed reviews from subsequent analyses since (a) different repertoire may call for different parameters of evaluations in criticism and (b) reviews that have to encompass in the available space the discussion of different pieces, maybe even belonging to different styles and periods, may easily become more vague and less rich in detailed information on interpretive issues.

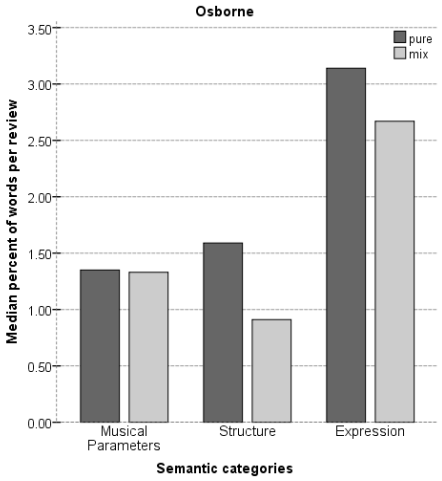
To check for this hypothesis, word categories were compared between mixed and non-mixed reviews. Given the results so far, to account for variability in word use between authors, reviews were further split by critic. The analysis was only possible for five critics (i.e., those who presented a large enough quantity of reviews in both conditions). Out of the 322 mixed reviews 41 entailed just a very few words about pieces other than Beethoven's sonatas, thus these were treated as non-mixed in the analysis. Results are reported in Table 4.9. Mann-Whitney test showed

significant differences in the use of terms related to musical parameters for three out of five critics, expression words for two reviewers and unity and structural details for one each.

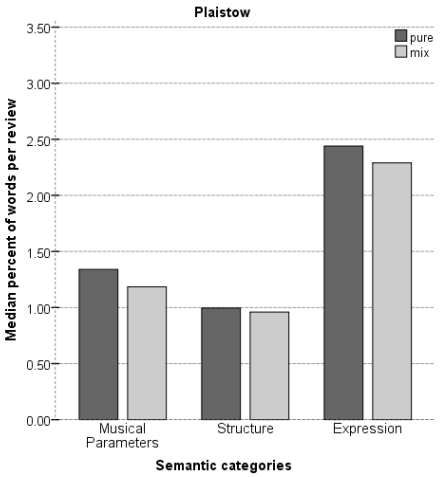
Table 4.9. Significant differences in word use between reviews of mixed repertoire and reviews concerning only Beethoven's sonatas.

Critic	*'Mixed' significant for*	*U*	*Sig.*
R. Osborne	Structural details	817.000	.000
	Expression	1,045.500	.038
S. Plaistow	-		
J. O. Chissell	Musical parameters	330.500	.048
	Expression	625.500	.034
	Unity	615.500	.047
B. Morrison	Musical parameters	201.000	.020
F. Fiske	Musical parameters	204.000	.037

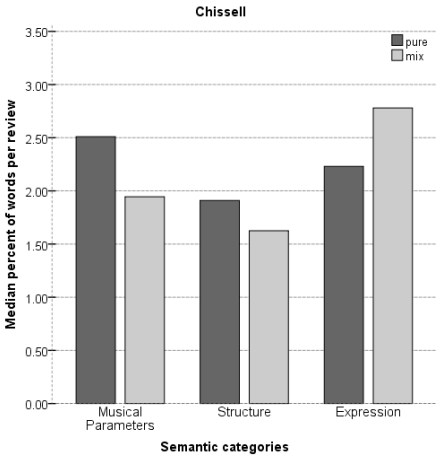
These differences were further explored graphically (Figure 4.5). A tendency can be noticed in employing fewer terms that refer to musical parameters when discussing mixed recordings. This tendency is stronger for Chissell and Fiske and mild in Morrison and Plaistow but absent in Osborne, who, instead, presents lower rates of expression terms and structural details for non-mixed reviews. These results seem to suggest that reviewing performances of diverse repertoire within one review influenced the content of the review itself, limiting the discussion of specific musical features in terms of musical parameters (Chissell, Morrison, Fiske) or expressive features and references to particular moments in the music (Osborne).



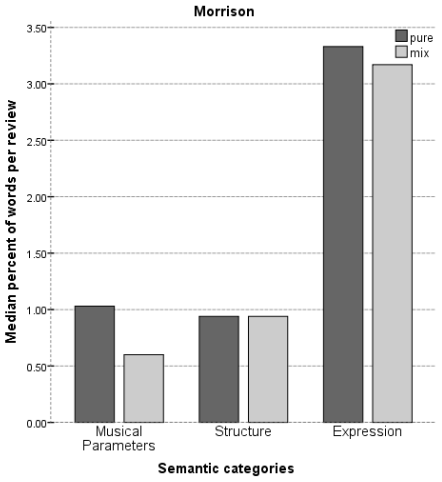
a)



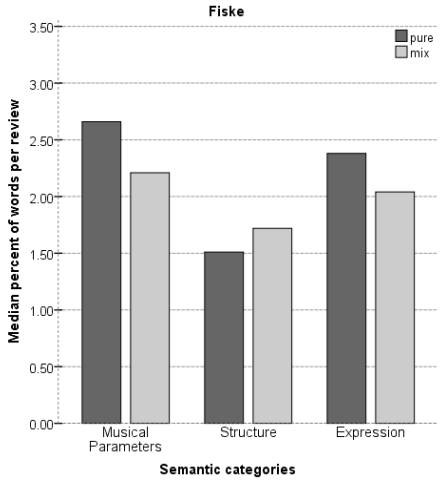
b)



c)



d)



e)

Figure 4.5a to 4.5e. Frequency of semantic categories for reviews of mixed repertoire and reviews of only Beethoven's sonatas. Frequencies are shown for each critic separately; starting from panel a) and down to e): Osborne, Plaistow, Chissell, Morrison, and Fiske.

Analysis (v): Comparison of word stem patterns

Method

Finally, following the comparison of semantic categories used, critics' vocabulary was examined by comparing patterns of word stems using the ReadMe algorithm. This final step added triangulation to the analysis, offering a measure of similarity at word level between different groups of reviews independent from the qualitatively developed semantic categories. It thus offered a further, distinctive perspective on the vocabulary of critics to cross-check the findings of analysis iv.

The ReadMe algorithm groups unstructured texts into pre-defined categories based on similarity of word content between documents. More precise estimations occur for documents that present a higher level of similarity within categories and dissimilarity between categories. Indirectly, this can then be seen as a measure of the degree of differentiation between one and the other groups of reviews.

This assumption however does not hold if documents contain keyword cues that clearly distinguish between categories. As such, this analysis is not suited to investigate differences in vocabulary used between mixed and non-mixed reviews. Indeed, a test run on these two groups of reviews resulted in an RMSe of 0.97 percentage points. This finding is not informative of the actual level of similarity between these groups of reviews, since it can be assumed that the low RMSe was due to the names of works performed entailed in the reviews, which triggered prompt recognition. However, for reviews written by different critics or in different periods no such explicit cues could be assumed. Therefore, this final analysis was only used to cross-check if vocabulary used in reviews is more strongly shaped by the identity of the critic or by the period of publication.

Reviews were randomly split into two sets: a training set (n = 400 reviews) and a test set (n = 439). In two subsequent tests, the software was given as training input information on the reviewer identity and on the decade of publication respectively. Estimates were then compared with real data at hand.

Results

Results are reported in Tables 4.10 and 4.11. Estimates of reviews grouped by critic showed a higher degree of accuracy, with a RMSe of 2.63 percentage points, against the 5.56 percentage points of estimates produced for reviews grouped by decade.

This is reflected also in the correlation coefficients between the two sets of values. Pearson's product-moment correlation coefficient was non-significant ($r_7 = .65$, $p = .06$) for classification of reviews by decade but significant and very strong ($r_9 = .92$, $p < .001$) for classification by critic.

Table 4.10. Categorization of reviews by decade: ReadMe estimates and actual values.

Decade	*Estimated*	*Actual*
1923-1930	5.19	0.00
1931-1940	4.03	4.33
1941-1950	1.70	2.73
1951-1960	20.49	7.06
1961-1970	14.34	18.45
1971-1980	14.42	19.59
1981-1990	13.85	17.08
1991-2000	13.27	14.12
2001-2010	12.71	16.63

Table 4.11. Categorization of reviews by critic: ReadMe estimates and actual values.

Critic	*Estimated*	*Actual*
R. Osborne	10.47	16.40
S. Plaistow	9.72	10.25
J. O. Chissell	7.88	8.43
B. Morrison	6.14	10.02
R. Fiske	7.50	4.78
A. Porter	4.91	5.01
D. J. Fanning	4.99	2.73
M. MacDonald	3.99	4.10
J. Distler	3.24	1.37
A. Robertson	3.27	1.37
Others	37.88	35.53

This further supports the results of analysis (iv) and strongly suggests that reviews vocabulary is intensely characterised by the reviewer, so much so that it was

possible for the ReadMe algorithm to recognise with a higher level of exactness the identity of the text authors through comparison of word stem patterns.

Conclusions

The first of this second set of data reduction analyses (iii) found 13 semantic categories of words, which were then computed for frequency in the sample (iv). Findings were cross-checked through a further comparison of word stem patterns between groups of reviews (v). The analyses revealed that the use of vocabulary in the sample is shaped more strongly by critic identity than by the review time period. That the use of certain semantic categories remains unchanged within one and the same writer even along time spans of several decades suggests that diachronic variations linked to different cultural settings may be observed – if at all – through comparisons between critics and taking into account the historical and cultural background of each of them, and not by contrasting reviews published in different decades. The analyses also showed differences between reviews of mixed and non-mixed repertoire, in terms of focus on musical features and in level of details in which performances are discussed, in line with observations done in Chapter 3. In the light of this, to better capture the richness of the data, it was decided that the selection of a critical review corpus for detailed thematic analyses should be led by critic, rather than by review period and that reviews of mixed repertoire should be excluded from these analyses.

One major difficulty encountered during the exploration of critics' vocabulary concerned the notion of 'expression' that should be used in this research. As said, expression is probably one of the most often discussed features in musical parlance; still, when it comes to define what 'musical expression' is, one encounters diverse notions and definitions. Given the importance and ubiquity of the concept of expression in music, a discussion of critics' judgements of performance cannot but partially be also a discussion of performance expressive features. But how should 'expression' be understood in the analysis of critical review? When does it make sense to say that a critic is discussing the expressivity of the performance? What do critics mean when they talk of expression? These questions naturally emerged during the analysis of critics' vocabulary and called for a further preliminary investigation prior to the inductive thematic analyses, aimed at clarifying the concept of expression

used in critical review. Details of this investigation, the method used and the obtained results constitute the content of Chapter 5.

FRAMING THE ANALYSIS

In the present chapter, a mixed qualitative/quantitative approach has been used to offer a first investigation of the textual content of the critical review sample. In the light of the findings of this investigation, and building on the overview given in Chapter 3, a selected corpus of reviews was produced to be used in the following detailed examination of critics' writings.

Following the analysis of vocabulary used in critical review (analyses iii, iv, and v), reviews of mixed repertoires were excluded from the selection. Further, it was decided that the selection should be led by critic. From the overview of reviews of recordings of Beethoven's piano sonatas published in the *Gramophone* between 1923 and 2010, 10 major critics emerged who produced a significant number of reviews (min. 24) within the collected sample. Obviously, that other critics wrote just one or a few reviews of Beethoven's piano sonatas says nothing about their experience with this repertoire and critical review in general (they could have – and probably had – written reviews of different repertoire and published in other magazines as well). However, it is for these 10 major critics that the material collected offers clear evidence of a high level of experience in reviewing, and specifically in reviewing performances of Beethoven's piano sonatas.

These ten reviewers moulded the criticism of Beethoven's piano sonatas in the magazine with their experience, personality and writing style, and it is thus on their reviews that the inductive thematic analyses of critics' judgements focus. Therefore, 10 reviews were selected for each of these 10 most prolific critics, so as to maximise variability in terms of period, pianist, and sonata reviewed. The 100 selected reviews in the final corpus (35,753 words, excluding titles, critic names and recording details) spanned August 1934 to July 2010, entailed 56 reviewed pianists, and comprised at least 6 reviews for each of Beethoven's 32 sonatas. Details of the corpus of 100 reviews are shown in Table 4.12. Following the thick-grained categorization of critical review content (analyses i and ii), it was decided to split the thematic analysis in layers, focusing first on the passages in reviews that discuss the *performance* (Chapters 6 and 7), and moving then onto the residual parts of the review texts to

investigate what other elements of the recorded performance enter critics' judgements of the final product (Chapter 8).

Table 4.12. Critical review corpus selected for the inductive thematic analyses.

Critic	*Reviews (Gramophone issue, page)*
Richard Osborne	Apr '82, p.66; May '83, p.49; Dec '83, p.84; Aug '86, p.49; Mar '93, p.73; Sept '95, p.83; Nov '95, p.146; Feb '96, p.75; Nov '00, p.86; Nov '04, p.79
Stephen Plaistow	Dec '61, p.57; Jun '62, p.64; Jun '63, p.36; Mar '64, p.63; Mar '65, p.57; Jul '66, p.47; Aug '79, p.69; Mar '88, p.50; Oct '89, p.98; Jan '02, p.81
Joan Olive Chissell	Mar '69, p.66; Jun '69, p.53; Feb '70, p.54; Dec '70, p.86; Jun '71, p.54; Mar '72, p.74; Mar '75, p.81; Oct '80, p.71; Feb '83, p.52; Jun '92, p.66
Bryce Morrison	May '93, p.74; Feb '02, p.63; Dec '02, p.72; Mar '3, p.63; Jan '05, p.76; May '05, p.104; Jun '06, p.71;Jun '08, p.81; Jul '10, p.77(i); Jul '10, p.77(ii)
Roger Fiske	Jul '55, p.44; Nov '57, p.17; Oct '58, p.65; Apr '59, p.64; Nov '59, p.67; Nov '59, p.68; Feb '61, p.48; Aug '63, p.31; Jul '84, p.41; Feb '86, p.52
Andrew Porter	Jun '54, p.42; Feb '59, p.60; Oct '54, p.50; Oct '54, p.51; Feb '55, p.56; May '56, p.49; Nov '56, p.55; Jun '57, p.19; Sept '57, p.17; May '58, p.16
David J. Fanning	Sept '86, p.84; Nov '86, p.78; Sept '88, p.80; Jun '89, p.64; Mar '90, p.69; Sept '90, p.116; Oct '90, p.116; Mar '91, p.85; Apr '92, p.111; Nov '92, p.152
Malcolm MacDonald	Aug '54, p.39; Nov '64, p.52; Jan '65, p.59; Mar '65, p.57; Mar '65, p.58; Jan '68, p.84; Jan '70, p.56; May '81, p.92; Nov '81, p.82; Dec '81, p.84
Jed Distler	Oct '05, p.81; Dec '05, p.97; May '06, p.90; Sept '06, p.80; Nov '06, p.97; Apr '07, p.92; Jun '07, p.84; Sept '07, p.76; Dec '08, p.103; Oct '09, p.88
Alec Robertson	Aug '34, p.29; Oct '35, p.18; Apr '36, p.18; Nov '36, p.17; Feb '37, p.19; Oct '45, p.16; Feb '47, p.8; Feb '48, p.23; Aug '50, p.23; Oct '53, p.22

5 EXPRESSION IN MUSIC CRITICISM[16]

The data reduction analyses reported in Chapter 4 led to the selection of a corpus of 100 reviews that will be used in Chapters 6 to 8 for the inductive thematic analyses of critics' writings. The same analyses however revealed a lack of understanding concerning the notion of 'musical expression' to apply in the present research.

For this reason, prior to the thematic analyses, a further study was run, aimed at clarifying what critics mean – or seem to be meaning – when they use the term 'expression' and derivatives. The end purpose of this was to decide which definition of 'expression' – if any – should be used in the examination of critics' judgements.

This chapter reports method and results of this study. It first introduces the construct of expression as it seems to be shared in different fields of music research. It then reports the method and findings of the analysis. Finally, based on these findings, it offers a few considerations on the problematic nature of 'expression' in musical performance and it explains how the insights gained on the notion of 'expression' in music criticism inform the analyses that follow.

THE CONCEPT OF MUSICAL EXPRESSION

Expression is arguably one of the most discussed subjects in music performance. It is part of the everyday parlance of performers, teachers and listeners and is usually used as a measure of the aesthetic value of the performance. In the context of higher music education, expression typically appears in segmented assessment schemes used to evaluate students' performances (McPherson & Schubert, 2004, p. 64); in music research, it is understood as 'fundamental to performance of every kind' (Clarke, 2002, p. 63) and 'what makes music performance worthwhile' (Juslin, 2003, p. 274).

Despite its ubiquity, the very notion of expression remains ambiguous, and this seems to be reflected in the different ways in which the term 'expression' is used in

[16] Portions of this chapter have been published in Alessandri (2014). For full reference, see List of Publications.

different contexts (Lindström, Juslin, Bresin, & Williamon, 2003, p. 24). In music research, the term 'expression' has been used to refer to 'those continuously variable parameters available to a performer: for the piano, for example, modifications of timing, dynamic and articulation are the only independently variable parameters available' (Clarke, 1991, p. 185), or to the performer's deliberate act of shaping these parameters (Widmer & Goebl, 2004, p. 203), or to a "set of perceptual qualities that reflect psychophysical relationships between 'objective' properties of the music, and 'subjective' (or, rather, objective but partly person-dependent) impressions of the listener" (Juslin, 2003, p. 276).

Expression in music has been discussed and investigated in various research disciplines, particularly in philosophy of music and different branches of empirical music research (Gabrielsson, 1999, 2003; Gracyk & Kania, 2011; Thompson, 2009). In philosophical studies the notion of expression tends to centre on the expression of emotion in music. This is usually framed in terms of the nature of the experience of hearing music as expressive of emotion and the value of that experience (Davies, 2011). In his contribution to the debate, Matravers (2007) proposes that in aesthetics the discussion on the nature of musical expression has reached an impasse. He suggests that the solution should presumably to be sought in the exploration of the experience of expression by qualified listeners. The suggestion is plausible both because musicians are presumably 'experts' in the experience of musical expression and because an adequate notion of expression needs to be as close as possible to actual musical practice. However, as discussed also in Chapter 1, it may be difficult to gain insights into musicians' experiences of musical expression not least because musicians may feel uncomfortable in discussing their own practice and may not be used to verbalizing their implicit knowledge (Lindström et al., 2003, p. 24). Besides musicians, however, it has been argued that there is another kind of listener who can offer not only a high level of musical expertise and exposure but also a trained ability to give specific and detailed descriptions of musical experiences: music critics. This chapter follows Matravers' suggestion by offering an exploration of the use of the term 'expression' in critical review of performance.

METHOD

An analysis of critical discourse that aims to clarify critics' notion of expression poses some important challenges. While theorists, philosophers and scientists are usually required to clarify their own understanding of the concept of expression before examining issues related to it, this does not seem to apply to critics. As discussed in Chapter 1, a piece of music criticism might reasonably be expected to deliver a description, analysis, contextualization and evaluation of the musical work or performance reviewed (Carroll, 2009). All these activities seem to presuppose a common understanding of a set of music related concepts necessary to discuss music, and expression is one of them. This means that critics write under the assumption of a shared understanding of what 'expression' means. Hence, the exploration of what the critic means by 'expression' has to come from the observation of the context in which the word is used and often requires the reader's interpretive judgement.

For this reason, a *keyword in context* (KWIC) approach was chosen for this analysis (Namey, et al., 2008). From the initial corpus of reviews all statements were extrapolated that contained the word 'express' or related terms such as 'expression', 'expressing', 'expressivity', in relation to the performance of the musical work reviewed. Statements that concerned works other than Beethoven's piano sonatas were excluded. Out of a total of 839 collected reviews, 154 contained the word 'express' at least once in relation to the performance of Beethoven's sonatas. Altogether 168 occurrences of express-statements were found across 154 reviews.

Express-statements were analysed independently by the author and one more researcher (with professional musical training), and annotations were added in the text to clarify what critics seemed to mean with the term 'express' in any single statement. Annotations were then discussed between the two researchers, and observations were compared. This led to the development of a codebook (Appendix 4). Subsequently, all statements were re-analysed and coded by both researchers. Reliability between the analysts was found to be Kappa = .84 ($p < .001$), 95% CI (.78 – .91), which represents an almost perfect agreement between coders (Landis & Koch, 1977). Discrepancies in coding (n = 15) were further discussed and agreed upon.

RESULTS

Different uses of 'expression'

Four main different uses of 'express' emerged from the analysis of the critical texts. These refer to (A) the use of certain performance acts; (B) the presentation of the music's content; (C) the manifestation of emotions, thoughts or intentions, either transitively or intransitively construed; and (D) features of the music (Table 5.1).

Out of the 52 critics identified in the corpus of reviews, 30 are represented in this selection of 'express'-statements (Table 5.2). Among them are nine out of the ten most prolific critics discussed in Chapter 3. Critics differ strongly in the ratio between quantity of 'express'-statements and number of reviews written, in particular two tendencies seem to emerge among the nine most prolific critics, with Fiske, Distler, Chissell, Plaistow and particularly Porter showing a large use of 'express'-claims in their writings, while Fanning, Morrison, Osborne, and Robertson almost never used the word 'expression' and its correlates in reviewing. This difference lies mostly in the different frequency of C-claims of expression. For what concerns B- and D-uses, these seem homogeneously spread across critics with the exception of Rast and Plaistow, who made large use of B- and D-statements respectively. In what follows, the four different uses of 'express' are described and briefly discussed. Full-text of the examples used is reported in Appendix 5.

Table 5.1. Distribution of 'express'-statements across the different uses of 'express'.

Use of 'express'	*Number of instances*
A – use of performance acts	41
B – presentation of music's content	15
C – manifestation of emotions, thoughts, intentions	71
D – features of the music	27
Unclear	14
TOTAL	168

Table 5.2. Distribution of 'express'-statements across reviewers.

Critic	*A-use*	*B-use*	*C-use*	*D-use*	*TOTAL*	*Reviews written*	*Proportion of express-claims*
Alec Robertson	1	0	0	0	1	24	4.17%
C. Headington	1	1	1	1	4		
Roger Fiske	5	1	5	0	11	52	21.15%
M. Cooper	0	1	0	0	1		
D. Cooke	1	0	0	0	1		
T. Harvey	0	0	0	0	0		
Joan O. Chissell	7	1	11	2	21	65	32.31%
Lionel Salter	0	0	0	0	0		
J. Budden	1	0	3	1	5		
W. S. Moore	0	0	2	0	2		
Edward Greenfield	0	0	3	0	3		
Andrew Porter	7	0	9	1	17	41	41.46%
R. Layton	0	0	0	0	0		
J. N. Moore	0	0	0	0	0		
Stephen Plaistow	5	0	14	9	28	88	31.82%
Richard Osborne	3	1	4	0	8	108	7.41%
Tim Parry	0	0	0	1	1		
N. Anthony	1	0	1	1	3		
Nicholas Rast	1	5	2	1	9		
David J. Fanning	1	0	0	2	3	33	9.09%
Bryce Morrison	2	1	0	2	5	60	8.33%
Robert Cowan	1	0	0	0	1		
Harrieth Smith	0	0	0	1	1		
J. M-Campbell	1	0	3	1	5		
S. Johnson	1	0	0	0	1		
Jed Distler	1	0	6	0	7	28	25.00%
Max Harrison	0	2	1	0	3		
S. F.*	1	0	0	0	1		
S. I.*	0	0	3	0	3		
H. F.*	0	0	2	0	2		
Unsigned	0	2	1	4	7		
TOTAL	41	15	71	27	154		

Note. Reviewers are listed in order of approximate date of birth. The frequency of express claims as a proportion of all reviews written is given only for the nine most prolific reviewers.

** Unidentified reviewer*

Performance options (A-statements)

Typically, a musical score is under-determined in the sense that its performance indications – for instance, presto and forte – can never be as specific as the performance itself. So it is up to the performer to decide how fast to play presto and how loud to play forte, where to apply a ritenuto and how to realize it, which notes to bring out by means of articulation and accentuation, etc. This set of possibilities that are open to a performer in his/her realization of a music piece are here called 'performance options'.

Performance options can be realized consciously or unconsciously (Clarke, 2002). The realized performance options are performance acts that not only fill the open space left by the under-determined score but also reflect the performer's understanding and interpretation of the work in question. Often, critics seem to use the term 'expression' or its correlates to indicate the way the performer engages in performance options. In particular, critics seek out for discussion those realized options that appear to them to be critically significant, regardless of the emotionally expressive function these performance acts may or may not possess.

These acts typically include slowing down towards the end of phrases, accelerating and getting louder towards the phrase climax, sudden dynamic changes to emphasise significant structural events, lengthening or shortening the duration of notes to enhance a rhythmic pattern, underlining the distinction between two voices by playing one voice louder than the other or by desynchronizing the left and right hand. These and other ways of exploiting performance options are referred to variably by terms such as 'expressive nuance' (August 1989, p.85), 'expressive inflection' (March 1957, p.52), 'expressive gesture' (November 1975, p.101), or, simply 'expression'.

Although any use of performance options seems to be a potential candidate for expressive gesture, at times 'express' seems to be used to refer primarily to performance acts that mirror the notion of Romantic expression (use of agogic and extreme dynamic contrasts being paramount examples, March 1954, p.46; March 1957, p.52). Other times, 'expression' seems to suggest merely giving emphasis, drawing attention to certain notes (January 1970, p.71).

Performance value and A-use of 'express'

There seems to be no implicit evaluative dimension in the notion of expression when critics use 'express' with reference to performance acts. In these cases, critics merely point out the performer's way of dealing with performance options and their choice of devices for realizing them. But saying that a performer engages in a variety of performance options does not, by itself, confer value to the performance; in fact, it is up to the critic to decide whether performance acts have a positive effect or not.

Among the 41 occurrences of A-statements found in the reviews, critics discuss the performer's use of expressive inflections in a positive vein in as few as ten occurrences (24.39%). More often critics either blame performers for relying too much on expressive inflections or for using them in an inappropriate way (26.82%), or they praise them for refraining from playing with too many inflections (24.39%). In merely three cases (7.31%) do critics wish the performer had made wider use of performance options (Table 5.3).

Arguably the most direct consequence of the use of expressive gestures is the emphasis of a musical detail. Delaying the onset of a chord brings attention to it, as would the playing of the chord suddenly louder or softer or playing it in a different timbre. Expression offers performers the possibility to bring to the fore important elements of music, to underline 'special moments', and this is something critics point out in praising performances (July 1973, p.60; November 1976, p.115; October 1990, p.114).

Table 5.3. Distribution of A-statements according to the valence of critics' judgements and the use or not of expressive inflections by the performer (as discussed by the critic).

Critic's judgement	*Use of expressive gesture*	
	Using expressive inflections	Refraining from expressive inflections
Positive judgement	10	10
Negative judgement	11	3

Note. Total occurrences of (A) statements: 41, 7 occurrences are not linked to a clear positive/negative judgement.

But if there is a perception of too many highlighted events, individual moments lose their significance: in a performance in which everything seems to be expressively emphasised, paradoxically, expressiveness is flattened out, and 'special

moments' remain engulfed within it. That happens, according to Fanning, in Kovacevich's recording of Op. 111, in which the over-pedalled and over-dramatized use of performance options 'flattens out the contour of expressive incident' (October 1992, p.138), while in Arrau's recording of Op. 2/1 and Op. 7 Plaistow regrets that the 'anxiety that no point should be missed' prompted the pianist 'to underline and over-emphasise everything', resulting in a 'forced and almost anguished air about his attempts at espressivo' (June 1966, p.47). Also, the emphasis on details brought about by the use of expressive inflections has, if excessive, a detrimental effect on the musical flow. This is the case in particular for expressive timing, as in Arrau's performance of Op. 110, Arioso dolente, in which the critic felt that the excessive use of expressive hesitations causes the 'pulsating left-hand accompaniment' to come 'near to losing its identity' (September 1966, p.63).

On the one hand critics blame performers for 'over-doing' in the exploitation of performance options; on the other they praise the use of expressive inflections that sound natural, spontaneous, 'where expression appears to simply flow from the notes' (August 1994, p.77; also September 1985, p.66; February 1987, p.68; April 2005, p.83). Critics emphasise the value of 'expressive economy' (November 2008, p.85), praising the performer for succeeding in keeping balance between 'steadiness and freedom of expression' (March 1954, p.39), between warmth of expression and emphasis on details and larger-scale relationships (June 1957, p.19; March 1975, p.85; November 1975, p.101; March 1985, p.56). In addition to 'over-doing', performers are also criticised for applying expressive inflections in contradiction to the score indications (April 1957, p.52; May 2001, p.79) or for using them in ways that are ineffective in conveying the music's expressive character (April 1956, p.55) or again, for using performance devices that are considered 'out-dated'. Maybe a telling example of this is the desynchronization of hands: an expedient often used at the beginning of the twentieth-century (Philip 1992, p.47). Delaying the entrance of the melodic note for expressive reasons in pieces other than Romantic repertoire would hardly be perceived as natural and non-affected today. In the corpus of reviews studied here considerations on the (mis)use of this expressive device start to appear in 1979 with Plaistow commenting on Brendel's rendition of Les Adieux that 'not every pianist these days … would dare to play his left hand before his right in the pursuit of true and natural expression' (May 1979, p.76). And in 1998 Cowan

dismisses the use of split chords as an 'old-fashioned' device (October 1998, p.129, on Backhaus).

Together, these findings seem to depict a critical view that easily tends to disapprove the use of expressive gestures for being detrimental to the overall musical value. This is true particularly for expressive timing that is more often discussed by critics in a negative vein (Table 5.4).

Of course these results need to be read in the wider context of the reviews at hand. In particular, Leech-Wilkinson (2009a) argues that the critics' (and musicians') negativity towards an excessive use of agogic and other expressive gestures could reflect a reaction against a performance style typical of the period prior Was World II that was particularly rich in such expressive inflections.

Table 5.4. Valence of critics' judgements on use of agogic.

	Agogic used	Agogic not used
Positive judgement	7	6
Negative judgement	12	1

Given the small and imbalanced sample size it is difficult to provide sufficient evidence for this hypothesis.[17] A comparison of statements made by 11 critics born after 1925 with those made by 5 critics born in the first decades of the century[18] seems to show a nominal increase in the percentage of A-statements valuing positively the use of expressive inflections. This result does not seem to support the hypothesis of an increased sensitivity towards expressive gestures as reaction to the pre-War performance style even though it is counterbalanced by a slight increase in the percentage of statements that praise pianists for refraining from such inflections (Table 5.5).

[17] The problem is aggravated by the paucity of information concerning what music critics were exposed to in which period of their lives.

[18] 1925 was chosen as cut-off date to parallel Leech-Wilkinson (2009b, pp. 252-253). He identifies a change in vocal performance style starting with Elizabeth Schwarzkopf (b. 1915) and Dietrich Fischer-Dieskau (b. 1925) and notes the increased relevance after the Second World War of pianists like Schnabel (b. 1882) and Kempff (b. 1895), 'who have been playing all along in a more restrained fashion'. Among A-statements 16 different critics could be identified. For one review (September 1985, p.86) it was not possible to verify the critic's identity (review was signed 'SF').

Table 5.5. Distribution of A-statements and valence of judgements by critics born before and after 1925.

Judgement	*Use of expressive devices*			
	Critics born before 1925		Critics born after 1925	
	Use of inflections	Refraining from use of inflections	Use of inflections	Refraining from use of inflections
Positive	2 (13.33%)	3 (20.00%)	8 (32.00%)	6 (24.00%)
Negative	4 (26.67%)	2 (13.33%)	7 (28.00%)	1 (4.00%)
Neutral	4		3	
Total	15		25	

Therefore, it may be more instructive to examine the actual distribution of statements by individual critics, especially since the small number of observations makes these results very sensitive to idiosyncratic differences. Table 5.6 shows valence of A-statements by critics listed in order of approximate date of birth. Columns (1) and (2) represent statements that suggest a positive disposition towards expressive nuances, columns (3) and (4) a negative one. Positive and negative statements seem to be quite evenly spread, although a few reviewers born around 1950 present only negative statements. Only five critics present more positive than negative statements, and four of them were born after 1925.

What is additionally noteworthy is the seeming decrease of using A-statements among critics born after the Second World War. This could suggest that in order to examine critics' attitude towards expressive nuances it may be necessary to extend the analysis to those statements that comment on the specific use of agogic or other expressive devices without using the word 'expression' to refer to them. This will be partly done in the subsequent analyses, discussed in Chapters 6 and 7.

Table 5.6. Relationship between 'expression' and valence in A-statements by critic.

Period of birth	*Date of Reviews**	*Name*	*Valence*			
			Expression POSITIVE		*Expression NEGATIVE*	
			Use of expression praised	*More expression wished for*	*Expression praised for not exceeding*	*Use of expression criticised*
<1925	1934-50	A. Robertson[19]				
	1955-86	R. Fiske		x	x	xx
	1960	D. Cooke**		x		
	1968-93	J. O. Chissell	xx		xx	x
	1966-67	J. Budden				x
ca. 1925-50	1954-60	A. Porter	xx		xxx	x
	1961-02	S. Plaistow**	xxx			x
	1974-04	R. Osborne**	xx			x
	1993-10	B. Morrison			x	x
	1993-09	R. Cowan				x
ca. 1950-60	1986-94	J. M.-Campbell				x
	1986-02	D. Fanning				x
	1993-94	N. Rast**	x			
	1994-98	S. Johnson			x	
	2003-06	N. Anthoni**		x		
	2006-09	J. Distler			x	

Note. Critics are listed in order of approximate date of birth.
** Beethoven's piano sonatas reviews in Gramophone*
*** critics with more positive than negative statements*

Presentation of the music content (B-statements)

The content of a piece of music consists of musical patterns: melodic, harmonic, rhythmic patterns and their relationships. Performers not only present these sound patterns, they also present them in certain ways. In his discussion of what it means to perform a composition, Walton (1988) calls this 'portrayal' of sound patterns. A

[19] Alec Robertson only presents one *neutral* occurrence of A-statement. Neutral statements were omitted in this table for clarity of reading.

performer may emphasise certain similarities between patterns and obscure others; may present one pattern as a restatement or as a variation or a development of another, as expository statement or as closure, and so on. Furthermore, performers may present patterns not only in different structural or functional roles; they may also present them as having different emotional features, for instance as being graceful or sad, or as changing from hope to despair.

The presentation of musical content is based on how the performer chooses to engage performance options. In this sense A and B statements of 'express' are closely related, but B types do not typically refer to performance options directly. Rather, B-statements refer to the character and quality of the content's presentation. For instance, we are told that in Alfred Brendel's rendition of the Waldstein sonata, 'the registrally distinctive dialogue of the introduzione is eloquently expressed' (November 1993, p.119), about O'Conor's recording of Beethoven sonata Op.22, we learn that 'internal detail and external form are beautifully expressed' while Jando is said to offer 'a convincing expression of the music's [E flat major op.7] dramatic content' (August 1994, p.72); and in Binn's rendition of op.27/1 'the texture and gracious content' find a 'particularly vivid and happy expression' (March 1982, p.68).

In these and similar cases terms other than 'express' – such as 'present', 'bring out', 'play' or 'perform' – could serve equally well. But note that the critic's reference to those presentations is always accompanied by a qualifying, mostly evaluative, term: whatever content is expressed it is expressed beautifully, eloquently, clearly, convincingly, with perfect control, coherently, vividly and so on. Note also that what can be expressed need not be limited to formal or structural features of the music (e.g., 'dialogue' of voices), it may also include drama, emotional qualities, patterns of tension and relaxation, and more abstract things like the revolutionary spirit.

Manifestation of inner states (C-statements)

These statements reflect ordinary, dictionary-type usage the most closely: expression as the outward manifestation of a person's inner states in her or his actions and

behaviour.[20] Thus 'Backhaus's innate sense of classical style has its full expression in the finale…' (June 1951, p.22), or his 'strongly dramatic intention … is finely expressed by his treatment of the ad libitum at the recapitulation' (January 1953, p.31). The performer's understanding of classical style and his dramatic intention are made manifest in his performance. It may be critically relevant whether the performance merely betrays the performer's thought or whether there is perceived intention behind the thought's manifestation. In any case, the performer's interpretive stance towards a piece, her conception of the piece, of how it should be played, is a subject of critical consideration.

Intransitive use of 'express'

'Express' is often used intransitively as adjective or adverb as when a performance is praised for being expressive or the performer for playing expressively. While these statements represent the most frequent use of 'express' found in the present corpus of critical review (55 occurrences), they offer minimal indications as to how they should be understood. In particular, the statement 'This performance is highly expressive' could plausibly be construed as meaning that the performer has used generously a wide variety of expressive devices (e.g., variations in tempo, dynamics, articulation) or that the performer has used a wide variety of expressions (e.g., musical sadness-expressions, joy-expressions, grief-expressions). However, in reviews, the characterization of a performance as being expressive is often given as if it were an independent value-adding feature of the performance. Claims such as 'The Beethoven sonata is much more satisfying. … The playing is expressive and sympathetic.' (October 1961, p.74) or 'In Les Adieux … Backhaus is inexpressive' (October 1954, p.51) construe the fact of being expressive (or inexpressive) as sufficient reason to explain the value (or lack thereof) of the performance.

Interpreting these statements with reference to the exploitation of performance options or the use of a variety of emotion expressions does not seem to justify the evaluative dimension with which 'expressive' is embedded. In fact, as discussed previously, the generous use of expressive inflections is, alone, not sufficient condition for a performance to be evaluated positively. Understanding these

[20] Oxford dictionary defines expression as (a) the action of making known one's thoughts or feelings or (b) a look on someone's face that conveys a particular emotion. *Oxford Dictionary Online*, http://oxforddictionaries.com/definition/english/expression. Accessed November 30th 2012.

statements to mean that the pianists utilized a wide variety of expressions does not seem to offer a sufficient condition either because such a performance may endow the music with expressive features that the music was not thought to have possessed in the first place, as in Pescia's recording of Op.110 criticised for its 'overstated emoting' that 'cheapens the effect of a tragedy-laden torpor' (October 2009, p.88).

How should then these statements be understood? A possible interpretation is offered by Robinson (2007). Following Robinson, in the arts as well as in daily usage we can distinguish between the expression of an emotion E, and the degree or level of expressiveness that this expression of E possesses. So, for example, an up-side down smiley and Edvard Munch's Melancholy painting can both be said to be expression of melancholy and sadness. However, while the downward smiley is a seemingly inexpressive expression of the emotion of melancholy Munch's painting gives 'a vivid sense of what it is like to be in a melancholy state' and can thus be said to be an expressive instantiation of the expression of this state. In music, jingle bells can be said to be an expression of joy and cheerfulness, but a not too expressive one, as opposed to, for instance, the triumphant culmination of Beethoven's Egmont Op. 84 (Robinson, 2007, p. 32).

The distinction between the instantiation of the expression of E and the degree of expressiveness of that instantiation offers a possible interpretation for the intransitive use of 'express' by critics. When a performance is praised for being expressive, or a performer for playing expressively, critics seem to refer to the level of expressiveness that the expression in the performance possesses, even if no indication is given of what the expression may be an expression of. In this view, the intransitive use of 'express' as in 'the performance is expressive' is a derivative of the standard notion of expression as manifestation of emotions that centres on the *how* rather than on *what* is expressed.

This is plausible if we think that expressing a certain emotion, state, or thought per se does not seem to be of great aesthetic interest. When it comes to expression, what matters in the arts is not so much that emotion E is expressed but how it is expressed; not the expression itself but the fact that the expression is evocative, beautiful, pleasurable, or expressive. So a performance can be praised for being more or less expressive (February 1967, p.60), "very", "intensely" or "soulfully" expressive (August 1963, p.31; July 1998, p.73; September 1990, p.116); a player

can be criticised for being "less expressively telling" than another (August 1994, p.73) or praised for being "expressively powerful" (December 2008, p.103).

Table 5.7. Valence judgement distribution of the 55 intransitive C-statements.

Positive judgement	*Negative judgement*	*Neutral*[21]
Expressive: 31 Balanced: 4	Lack of expression: 16	4

In the reviews at hand 55 instances of intransitive C-statements of 'express' were found (Table 5.7). Most of them (85.45%) were statements in which expression was used as an independent value-adding feature of the performance, either praising the performance for being expressive or criticizing it for being not or insufficiently expressive. In four cases, expression was praised with the added condition of balance: the performance is good for being expressive and yet not pedantic or mannered, or without expression becoming detrimental to the natural tension of the phrase (December 1976, p.94; March 1974, p.56; March 1967, p.54, October 2005, p.81).

Music qualities (D-statements)

In addition to the three uses discussed so far, 'express' is also employed at times to describe features of the musical composition itself, rather than its performance, in particular features of the music that a good performance should bring out. Thus, music may be said to be of a "deeply expressive nature" (September 1937, p.19) that the performance manages to realize; or there may be "expression" in the music (October 2009, p.88) that is brought out best by merely following the score's instructions; or "expressiveness" which is intensified by the performer's "exploitation of the music's intrinsic possibilities" (August 1994, p.78). Again, the way the performance brings out (or not) the features designated by 'expressive' and related words is determined by the performer's handling of performance options.

[21] The four cases labelled as 'neutral' refer to statements in which expression was not used directly as an evaluative feature of the performance at hand, like in 'In the latter sonata's first movement, Biss makes the most of the development section's seeming rhythmic disintegration, although his habitually telegraphed ritards soften the austere surface that equally expressive yet more literal readings convey' (December 2007, p.57).

DISCUSSION

Four uses of 'express' in critical discourse have been identified, and observations have been made on the relationship between expression and performance value. In this section a few further considerations are given on the complexity of the notion of expression that emerged from this analysis.

Physical versus Psychological dimension of expression

The two most common uses of 'express' that have emerged construe expression as the manifestation of someone's inner state (C-statements) and as the use of certain performance acts (expressive inflections) by the performer (A-statements). The former use is closer to the standard usage of 'expression' and to the understanding that permeates the philosophical discourse, while the latter seems to reflect more the use of 'expression' often found in empirical research.

This bi-dimensionality and the way critical talk may slide easily and often not noticeably from one dimension to the other surfaced as one of the significant aspects of the 'express'-vocabulary in the critical discourse. These two dimensions could be labelled as psychological and physical dimensions. The use of 'express' in C-statements essentially involves the thought of someone's inner state to be outwardly manifested in his or her way of performing a musical work. This is the psychological dimension. By contrast A-statements – for instance, the praise of an excellent gradation of a crescendo – do not presuppose this thought nor do they typically seem to suggest it. Where the idea of some inner state being outwardly manifested in the performer's behaviour or actions or where the reference to some inner state, especially to emotions, is not essential to understanding the critical statement that uses 'express' or related terms, it can be assumed that the discourse is limited to the physical dimension.

That in critical discourse physical 'express'-statements may easily slide into psychological ones and that this slide seems natural and often goes unnoticed may be related to different factors. The way the performer realizes performance options may endow the music with a certain emotional character that the critical listener may perceive as the music's or the performer's expression of emotion. Or it may arouse an emotion in the listener, which they may take to be the music's expressive character or the performer's expression of emotion. In addition, performers often

intend to project, by virtue of playing the music in a certain way, a particular mental state (e.g., the delayed onset of a note may be used to suggest that reaching that note occurs under intense physical and emotional pressure) or they may try to control their playing by means of images of inner states, such as nervous energy, exuberance, grief, and so on (Woody & McPherson, 2010, pp. 411-414).

This duality of the meaning and use of 'expression' makes critical discourse complex. The problem is compounded when questions of value are added to the discussion. Variations in timing, dynamics and articulation referred to as 'expressive hesitations' may or may not be constitutive of expression in its psychological understanding, whereas the aesthetically relevant effect that these variations may have locally can both add to and detract from the overall expressiveness of the performance and (or) the work.

Expression in criticism and in music research

The way expressive hesitations are discussed by critics relates to the notion of expression as continuous variations in different musical parameters and thus provides a link with scientific studies of performance expression that are concerned with the central task in music performance (within the Western classical tradition): that of deciding how the notated values have to be played in order to present the musical work's structure and its other aesthetically significant properties. The fact that a musically meaningful performance of a score typically implies that some of the notated values are not realized as written is a standard feature of music performance. Indeed, the literal rendition of a score would likely produce musical non-sense. This may be called realizing performance options.

Critical discourse, on the other hand, seems to apply the term 'expression' to indicate those realized performance options that are relevant for the expressive performance of the work (i.e., B and C-statements, cf. Table 5.1). It seems reasonable to consider this set of realized options as related to, but not coinciding with, the larger set of continuous variations in timing, articulation, and dynamics: the micro-variations denoted by the term 'expression' in research. For instance, routine micro-variations in timing like slowing down at the end of a phrase would not be singled out by critics as examples of expressive gestures. It is only when these micro-variations reach a certain degree, when they develop into an expressive ritenuto or

rallentando, or are used in unexpected ways to emphasise a particular, non-obvious musical pattern that 'expression' statements are used. Similarly, critics would hardly consider as an expressive act the random variations due to unintended body movement characteristic of any human performance that in the research context are understood as being a component of expression (Juslin, 2003).[22]

Finally, in critical practice 'expression' is used at times in a narrower sense to refer to conventions typical of Romantic performance practice, so much so that 'expression' seems to become a synonym for agogic. This, just as the sensitivity to the use of expressive timing discussed above, is linked to the musical and cultural background of listeners and the nature of the repertoire reviewed. These distinctions, however, only emphasise the fluid nature of the notion of expression in critical practice and the consequent difficulty in applying a common vocabulary across the various disciplines engaged in describing and understanding the nature of musical performance and its impact on the listener.

CONCLUSIONS

The present chapter offered a focused, detailed analysis of the way 'expression' and derivatives are used in critical reviews. Findings showed that critics rely on the explicit use of 'expression' relatively infrequently (18.36% of reviews) and that when they do, they use the term in a fluid and multi-layered way. One and the same term is used to indicate very different properties of the performance – linked to technical aspects of the musical delivery as well as to more abstract psychological constructs. 'Expressive' can even be used as a purely evaluative term, as a kind of synonym for 'good' or 'beautiful', and in some cases is used outside the performance dimension, to indicate aspects of the music composition.

The purpose of this analysis was to clarify what definition of 'expression' could be used in the present research. The results suggest that the notion of expression in criticism is too complex and diversified to be discussed under one single label without losing important information. In order to capture the richness of this notion in its diverse meanings and nuances, it was then decided to split

[22] That said, of course critics are themselves subject to perceptual mechanisms and would not be able – in standard circumstances – to distinguish between intentional and unintentional acts in performance.

'expression' in its individual components and let those emerge from the text analysis independently. As such, in the analyses that follow, no general definition of 'expression' was used. On the other hand, in the few cases in which critics used the term 'express' and derivatives explicitly, the four-uses classification developed in this study was employed to interpret the relevant statements.

This investigation of the notion of expression in critical review completed the block of preliminary analyses run on the collected sample. This block constituted the first part of the present research, aimed at winning a first understanding of the material at hand, and exploring the different ways in which the textual content could be tackled. The following three Chapters, 6 to 8, form the second part of the thesis. They focus on the corpus of 100 reviews selected in Chapter 4 and report the series of inductive thematic analyses that led to the development of a visual descriptive model of recorded performance critical review. The first of these analyses, reported in Chapter 6, examines what performance-related features are discussed in critical review.

6 CRITICS' JUDGEMENTS OF PERFORMANCE[23]

With this chapter begins the second part of this thesis, entailing the core thematic analyses that led to the development of a model of critical review of recorded performance (Chapter 9, p. 292). After the overview of review metadata (Chapter 3), the data reduction analyses (Chapter 4), and the clarification of the notion of 'expression' in music criticism (Chapter 5), this and the next two chapters directly address the core question of the present research (p.81):

What reasons do expert critics adduce to support their value judgements?

Findings in Chapter 4 showed that the predominant object of discussion in critical review is the *performance* of a given work. Given the complexity and variety of topics discussed in relation to the performance, two chapters are devoted to this: the present one addresses the question '*What* do expert critics write about when reviewing a performance?', while the second examines *how* the diverse elements critics write about are used to build value judgements, thus focusing on the relationship between valence (positive or negative judgement) and different properties of the performance.

METHOD

Material

The data reduction techniques applied in Chapter 4 led to the selection of a corpus of reviews that could be representative of the whole dataset and suited to an inductive thematic analysis. This corpus of 100 reviews written by 10 different critics (10 reviews each) is the object of analysis in this and the next two chapters (details of the reviews encompassed in the corpus are reported in Table 4.12, at the end of Chapter 4). Reviews were pre-prepared by visually separating (highlighting) parts of the text

[23] Content within this chapter has been published within the following: Alessandri, Williamson, Eiholzer & Williamon, 2015. For full reference, see List of Publications.

concerning the *performance* from the rest (following the codebook developed in Chapter 4). This allowed separate analyses to be carried out on the two parts of the text (performance related and extra-performance), while maintaining the textual context during the analysis process. The analysis reported in the present chapter focused on the performance related part of the review text.

Thematic analysis

As discussed in Chapter 2, a major challenge of a person-centred approach to the analysis of unstructured texts is its dependency on the analyst's interpretation. Here, a new analysis protocol was developed, based on the work of Williamson, Jilka, Fry, Finkel, Müllensiefen, and Stewart (2011) and on the strategies suggested by Guest et al. (2012). To add validity to the analysis, the protocol involved the participation of two researchers in the development of the codebook and an iterative process of text comparison and code revision.

The two researchers were chosen so to reflect the standpoints of two common categories of review readers: the professionally trained listener, who is familiar with the work discussed and comes to the repertoire with both knowledge and strong personal preferences, and the more generally musically trained listener, who has a solid grasp of the musical vocabulary but is not necessarily familiar with the repertoire and the technicalities of the instrument reviewed.

I, as first analyst, had the perspective of the informed listener, familiar with the repertoire and with first-hand experience in performing the sonatas at professional level. The second analyst – a researcher with extensive experience in large scale thematic analysis and who was native English speaker – was also musically trained, but at a non-professional level and in a different instrument (guitar). Thus she added to the analysis the perspective of the musically trained, music amateur reader with no specific technical knowledge of the pieces reviewed. This variety of perspectives was sought to permit the development of a model whose application and understanding could be open to a wider audience, and would not require professional musical training.

The inductive thematic analysis involved three main stages.In the initial stage, a subset of 10 reviews (one for each of the 10 critics) was hand-coded by the two researchers independently, using line-by-line open coding. After the line-by-line

open coding, each researcher organized their codes into themes that summarise the content of the reviews. To enhance validity and assure that the empirical material was reflected in the themes, the use of multi-layered coding (codes based on notes or other codes) was avoided. Instead, all themes were attached to the source data. Emergent themes were then compared between the two researchers. To minimise subjectivity of interpretation, each researcher in turn explained a theme, proposing a definition and justifying it by means of examples from the data.

At this stage, the different experiences and perspectives of the researchers enriched the discussion on the interpretation of the text, leading to negotiations on the development of theme definitions. One example of this was the theme ***Timing***, for which the second coder proposed a definition based on the idea of 'variations in speed'. Even though justifiable in terms of physical properties of the sound, the use of the term 'speed' to define time was difficult to accept for me as trained musician and pedagogue, since (i) it suggests an oversimplification of a perceptively utterly significant component of performance and (ii) it seems associated with the ideas of technical proficiency and virtuosity more than with that of a musically meaningful timing. To maintain a definition that would be accessible to a wider audience and close to the physical characteristic of the musical sound, the term 'speed' was maintained in the final definition. This was however accompanied by the concept of 'beat frequency' to better capture the musically meaningful component of timing. Following this process of discussion and negotiations an agreed codebook of emergent themes was developed (see Appendix 6).

In the next stage, the author applied the codebook to the whole dataset of 100 reviews, revising themes and definitions only where appropriate. Segmentation of text was performed at clause level. One new theme (***Evaluation_****Taste*) emerged at this stage; this theme was discussed with the second coder and added to the theme codebook. After about one third of the documents were coded, the saturation point was reached. After this, no new themes were found and the whole dataset was analysed using the completed codebook (Atlas.ti 6.1 was used for the analysis, see Appendix 7 for sample of coded material).

Finally, upon completion of the full coding stage, lists of quotes were analysed and compared for each theme, to check for coding mistakes, ambiguities or distinctions between themes that needed clarification. In this stage sub-categories

within themes emerged and relationships (linear or hierarchical) between themes were clarified so as to maximize differentiation between and homogeneity within themes. Coding for the whole dataset was revised at the end of this process to adjust it to the newly emerged model. This led to the development of a visual descriptive model of performance judgements in critical review.

RESULTS

Critical review emerged as a very dense form of writing. The 100 reviews resulted in a total of 6,012 codes with an average density of 6.66 codes per clause. Density across critics ranged from 5.01 (MacDonald) to 9.76 codes per clause (Distler). The fact that codes were so closely spaced limited the kinds of analysis that could be run on the results: Qualitative thematic analysis usually enables the exploration of patterns of themes through the observations of co-occurrences between codes. The high number of codes per clause found in reviews made co-occurrence tables unusable. A different approach was then taken to examine how themes relate to one another: this is reported at length in Chapter 7. What the present analysis permitted however was the creation of a comprehensive map of the topics discussed in reviews, and this is what is presented here.

On completion of the analysis there were three superordinate theme families – **Primary Descriptors**, **Supervenient Descriptors**, and **Evaluative Judgements**, with 12 dominant themes, which comprised a further 33 sub-themes. Figure 6.1 visualises the emergent descriptive model, with superordinate theme families located in the left hand side of the figure.

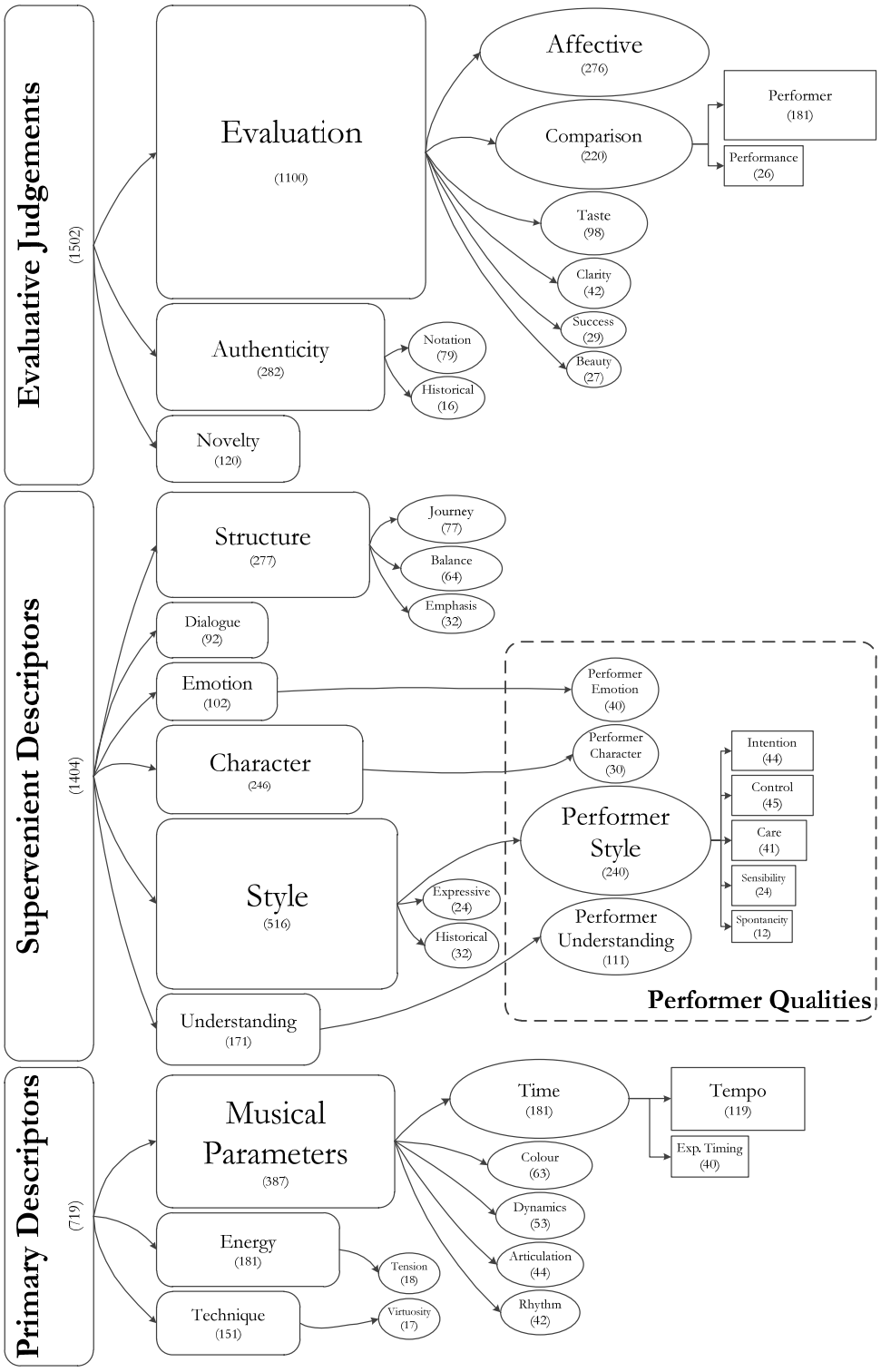


Figure 6.1. Performance-related themes discussed by critics. Superordinate theme families are located in the left-hand side of the model. Themes are visualised hierarchically moving from left to right, and from rounded rectangles, leading to oval, and when necessary down to square shapes. Arrows reinforce the visualization of this hierarchical structure. Shape size roughly suggests the relative weight of themes, in terms of frequency of occurrence. In parentheses under each theme name is the number of times the theme was coded in the texts. Each time a sub-theme was coded, the relevant higher-order themes were coded as well.

Evaluative Judgements comprises comments on the value, importance or merit of the performance. This was the largest superordinate theme family, with 1,502 occurrences. This family also entailed the single largest and most widely spread dominant theme in the whole analysis, ***Evaluation*** (1,100 occurrences, found in 100% of reviews, see Table 6.1, p. 198). **Primary** and **Supervenient Descriptors** entailed characterizations of the performance. **Supervenient Descriptors** was the prominent family between the two and the most varied in the whole analysis, encompassing 1,404 occurrences and 20 sub-themes. Within this family, a group of sub-themes has been highlighted in the model (*Performer Emotion*, *Performer Character*, *Performer Style*, and *Performer Understanding*), that characterises the performance focusing on its agent rather than on the performance itself, thus assigning qualities to the performer (**Performer Qualities**).

In the following section a description for each theme is provided – with superordinate family names in bold, dominant themes in bold italic and sub-themes in italic – together with theme definitions from the codebook and examples from the texts. Issue, page, and critic's name are given for each example. Numbers in parentheses after theme names indicate the frequency with which the theme was coded in the text. Hierarchical relationships between themes are further marked off by indentation.

Superordinate theme family 1: Primary Descriptors

The first superordinate family (719, lower section in Figure 6.1) encompasses three dominant themes that characterise the performance focusing on specific actions or qualities of the musical sound: ***Musical Parameters, Technique***, and ***Energy***.

Musical Parameters (387): This large dominant theme entails comments on the nature of the musical sound that can be either local, concerning single notes or phrases, or global, concerning the whole piece, a movement or a section of movement. Within this dominant theme there are five sub-themes.

> *Dynamics* (53) and *Time* (181) comprise comments on loudness and speed of the musical sound respectively.

> "Brendel has the rare ability to play very quietly and to make the sound rise from almost nothing." (Fiske, August 1963, p. 31)

Comments on speed were further divided into *Tempo* (119) – including comments on speed (beat frequency) on global level – and *Expressive Timing* (40) – grouping comments on temporal variations (from the underpinning beat frequency) on local level, also including comments on pause duration:

> "The slow movement is surely fastish for a soulful Largo" (Chissell, March 1969, p. 66)
>
> "In the earthy G major Sonata Goode hams up Beethoven's wit by extending the rests, sometimes by virtually a whole beat" (Fanning, September 1990, p. 116)

Colour (63) focuses on qualities of sound that relate to timbre and texture:

> "…the sonority is more astringent" (Distler, September 2006, p. 80)

Articulation (44) focuses on the way in which two successive notes are connected. It includes comments on accentuations as well as technical terms used to indicate ways of connecting notes (staccato, legato).

> "…some of his sforzandi, notably in the scherzo, are understated to the point of inaudibility" (Fiske, October 1958, p. 65)

The last sub-theme within ***Musical Parameters*** is *Rhythm* (42). Rhythm can be defined as a pattern of accents (see Cooper & Meyer, 1960, p. 6), whose perception is co-determined by other parameters like pitch-contour, articulation, dynamics, or tempo. As such, no assumptions were made about rhythmic content: passages in this code must have explicitly included the term 'rhythm' and its variants:

> "The first movement of the A flat has little rhythmic impulse" (Porter, September 1957, p. 17)

Energy (181): This dominant theme captures aspects of the performance that convey strength and vitality.

> "But in the final resort it is the voltage that counts in this eruptive fugue" (Chissell, March 1972, p. 74)"
>
> "…the prestissimo is impetuous but not undisciplined" (Porter, June 1957, p. 19)

The only sub-theme of ***Energy*** focuses on *Tension* (18). Here are coded comments that entail the term 'tension' and related words (tense, release, etc.).

> "This is not a reading which is consciously daemonic, fluid or exquisitely 'painted', the tensions eerily depressed" (Osborne, August 1986, p. 49)

Technique (151): In this third dominant theme the critic focuses on the mechanistic qualities of musical delivery. This includes comments on pedalling, hand de-synchronization and repeats and ornaments realization.

> "…in spite of some nimble fingerwork in the quick music the sonata never quite makes its usual impact" (Fiske, July 1984, p.41)
>
> "Paik shows off by fingering the prestissimo octaves rather than playing them as glissandi" (Distler, October 2005, p. 81)

The only sub-theme within ***Technique*** is *Virtuosity* (17), which collects passages in which the term 'virtuosity' and its correlates (virtuoso, virtuosistic, bravura, etc.) are overtly mentioned.

> "What immaculate virtuosity in the finale of Op 10 No 2" (Morrison, June 2006, p. 71)
>
> "…he neither subjects the notes to his virtuosic will, nor demeans his own technique by mimetic attempts at audible disorder" (Osborne, December 1983, p. 84)

Superordinate theme family 2: Supervenient Descriptors

Within this superordinate family (1,404) are dominant and sub-themes that portray the performance focusing on higher-order properties, that is, properties building on combinations of **Primary Descriptors** qualities. The largest dominant theme in this group is artistic ***Style***, accompanied by ***Structure***, ***Character***, ***Understanding, Emotion,*** and ***Dialogue.*** It is the most varied group of themes and the richest in metaphors and similes (middle panel in Figure 6.1).

Within this large family, a group of sub-themes emerged that focuses on the qualities of the performer, rather than the performance. These four sub-themes (*Performer Style*, *Performer Emotion*, *Performer Character,* and *Performer Understanding*) are presented separately under the heading **Performer Qualities**.

Style (516): The second largest dominant theme emerged from the full analysis encompasses characterisations of the performance that describe the manner of execution. It includes a large number of terms and expressions used metaphorically.

> "The fourth variation is turgid" (Fanning, March 1990, p. 69)
>
> "Casadesus strangely suggests at times a little French acrobat hopping through his paces" (Porter, May 1956, p. 49)
>
> "Notice how he eases his way into the first movement's opening measures as if sneaking on stage" (Distler, December 2005, p. 97)

The sub-theme *Historical* (32) focuses on manner of execution linked to different performance practices or historical periods.

> "This is the romantic approach to such music" (Fiske, November 1959, p. 68)
>
> "Serkin is the most classical" (Plaistow, October 1989, p. 98)
>
> "… a sort of Beethoven playing which has surely been outdated since Schnabel" (Porter, June 1954, p. 42)

A further sub-theme characterises the performance in terms of its *Expressive* (24) content. These passages suggest artistic styles that make use of expressive inflections or that are generally described as expressive (see A and C-statements in Chapter 5).

> "He plays the first page with almost no expression at all" (Fiske, November 1959, p. 68)

Structure (277): This dominant theme includes comments on the way in which the performer portrays the design of the music, its elements, patterns and relationships between them (as well as patterns and relationships that ought to be there but are not realized). It includes comments on phrasing and a conspicuous number of visual metaphors.

> "The tempo is spacious, apt to Gilels's mastery of the music's asymmetric lines and huge paragraphs, paragraphs as big as an East Anglian sky" (Osborne, December 1983, p. 84)
>
> "Larsen comes off sounding relatively sober, four-square and uneventful" (Distler, October 2009, p. 88)

A sub-theme of ***Structure*** is *Journey* (77). This includes comments that convey the idea of movement. The portrayal of music is described as a dynamic process:

> "His supple unwinding of the Trio is most attractive" (Porter, May 1958, p. 16)

> "His Minuet and Trio do not have the spring and unhurried bounciness of Fischer's" (Porter, May 1958, p. 16)
>
> "The Scherzo begins its enchanted journey" (Osborne, February 1996, p. 75)

Another sub-theme is *Balance* (64), which focuses on the portrayal of musical design in a coherent, unified or well-proportioned way.

> "Wührer has balanced to perfection the component sections of this moderato cantabile" (Porter, June 1957, p. 19)
>
> "…only a few movements achieve the continuity and natural expressiveness proper to them" (Plaistow, July 1966, p. 47)

A last sub-theme of ***Structure*** is *Emphasis* (32). Here are comments on a portrayal of the musical design that brings to the fore specific elements or details of the music.

> "…a more than usually striking significance is given to the four-note phrase so reminiscent of the opening of the Fifth Symphony" (Robertson, October 1935, p. 18)
>
> "Elsewhere there's a forced and almost anguished air about his attempts at espressivo, as if anxiety that no point should be missed has prompted him to underline and over-emphasise everything" (Plaistow, July 1966, p. 47)

Character (246): This dominant theme entails characterisations of the performance in terms of mental and moral qualities of an individual or of an atmosphere.

> "It is very possible to prefer more lenience at the beginning of this work" (MacDonald, March, 1965, p. 57)
>
> "Balm or solace indeed after the dark and ceremonial Funeral March" (Morrison, May 2005, p. 104)
>
> "His opening to Op. 101 … is suitably devotional" (Morrison, May 2005, p. 104)

Understanding (181): Comments on qualities of the performance and its realisation that reflect reasoning and use of intellect:

> "The brief reminiscence of the opening bars might have been more ruminative" (Fiske, August 1963, p. 31)
>
> "The Scherzo … is also equivocal" (Osborne, December 1983, p. 84)

Emotion (102): Characterisations of the performance in terms of affective states. The decision about what terms delineate an affective state was made based on the list of

stems provided by the Geneva Affect Label Coder (GALC), described in Scherer (2005).

> "…what a sense of elemental rage in the heaven-storming finale of the Appassionata" (Morrison, January 2005, p. 76)

Dialogue (92): Comments on the communicativeness of the performance, as well as speech metaphors.

> "There's a failure of communication somewhere here and I'm just not getting the message" (Plaistow, July 1966, p. 47)
>
> "Above all these are 'speaking' performances" (Morrison, June 2006, p. 71)
>
> "There are solecisms in Richter's playing" (Fanning, April 1992, p. 111)
>
> "…a far greater use of declamatory effects and rhetorical tropes than was the case in either of the two earlier cycles" (Osborne, February 1996, p. 75)

Performer Qualities

Under this heading are grouped **Supervenient Descriptors** that focus on the player, rather than on the performance. They describe his/her traits or dispositions towards the music. Here are comments on the *Performer Style*, *Performer Understanding Performer Character*, and *Performer Emotion*. These comments were found in 93 out of the 100 reviews.

Performer Style (240): While ***Style*** describes manners of execution, *Performer Style* entails comments on manner of execution that reflect the pianist's attitude towards or approach to the work.

> "Solomon played this movement with immense reverence as though he thought it the greatest piano music in existence" (Fiske, November 1959, p. 68)
>
> "Ogdon can play like a listener – that is, with an unselfconscious, unforced continuity" (Fanning, November 1986, p. 78)

Within *Performer Style* five further sub-themes emerged that focus on *control*, *intention*, *care*, *sensibility*, and *spontaneity*.

Control (45) entails comments on the performer's aesthetic and technical command of the performance. It also includes comments on the performer's effort or difficulty in performing.

> "I'm bound to acknowledge Gilels's peerless control over tone, tempo and phrasing" (Osborne, May 1983, p. 49)
>
> "The finale has an effortless continuity" (Chissell, June 1992, p. 66)
>
> "Formidably in command of the music" (Morrison, December 2002, p. 72)

Intention (44) entails comments on inferred performer's intentionality, his/her preferences and decision processes:

> "…here she screws up the tensions of the music (evidently intentionally)" (Plaistow, June 1963, p. 36)
>
> "A lot of things don't seem to come off as he intends" (Plaistow, July 1966, p. 47)

Care (41) focuses on the performer's carefulness in dealing with aspects of the music or performance.

> "…one still senses the meticulous, almost pointillist care over each individual note" (Fiske, November 1957, p. 17)
>
> "Yet Paik also can be cavalier regarding details of accentuation, dynamics and tempo" (Distler, October 2005, p. 81)

Sensibility (24) entails comments on the performer's ability to appreciate and respond to complex aesthetic stimuli, his/her sensitivity to the presence or importance of certain musical features.

> "In Op. 110 he is most exquisitely sensitive to the phrases" (Porter, October 1954, p. 51)
>
> "Heidsieck's tempo for the ensuing Allegro molto e con brio is much more sensible than Frank's" (Chissell, June 1971, p. 54)

Finally, *Spontaneity* (12) comments on the performer's deliberation in realizing the music.

> "The effect overall, each time, is overworked" (Plaistow, January 2002, p. 81)
>
> "…these five sonatas show that Schnabel's performances, however deeply considered, emerged fresh and spontaneous" (Morrison, January 2005, p. 76)

Performer Understanding (111): This sub-theme of ***Understanding*** entails comments that reflect the interpreter's comprehension of the music and his/her discernment or imaginative power in its realization.

> "Maria Donska makes her own contribution by playing all three sonatas perceptively" (MacDonald, November 1964, p. 52)
>
> "I find here a serious meditation devoid of poetic impulse" (Robertson, October 1953, p. 22)
>
> "Its great virtue is that it is obviously a serious reading" (Porter, June 1954, p. 42)

The critic may suggest the performer's vision of the music, question his/her understanding, or discuss his/her agreement with it.

> "I should be most interested to hear an explanation of this interpretative eccentricity" (Robertson, April 1936, p. 18)
>
> "Yet I still doubt his wisdom in choosing a starting tempo of crotchet 72" (Chissell, March 1972, p. 74)
>
> "I did not always agree with his view of it" (Fiske, November 1959, p. 68)

Performer Emotion (40): This sub-theme of ***Emotion*** focuses on affective states that are construed as qualities of the performer.

> "In Op. 110 Serkin seems to have regained poise" (Plaistow, October 1989, p. 98)
>
> "…a rough patch in the Scherzo of Op 110 clearly bothered him" (Osborne, November 2004, p. 79)
>
> "Sheppard revels in the whimsical Menuetto" (Morrison, June 2006, p. 71)

Performer Character (30): Finally, this sub-theme focuses on mental and moral qualities of the performer.

> "Ashkenazy tends to drive home his points too fanatically" (Fanning, March 1990, p. 69)
>
> "Bernard Roberts is a Beethoven interpreter of sterling integrity" (Osborne, November 1995, p. 146)
>
> "Gulda's mix of severity and inwardness is, again, enthralling" (Morrison, December 2002, p. 72)

Superordinate theme family 3: Evaluative Judgements

The first two theme families comprise judgements that portray aspects of the performance. The last and largest theme family (1,502) focuses on judgements on the value, importance, usefulness or merit of the performance.

Evaluative Judgements emerged as a pervasive and substantial constituent of critical review with an average of 13.29 occurrences per review (as opposed to 10.36 occurrences/review of **Supervenient Descriptors**; 6.36 of **Primary Descriptors**)

and a high frequency of co-occurrence with the other families (67.92% of **Primary Descriptors** and 61.77% of **Supervenient Descriptors** co-occurred with **Evaluative Judgements**). This third superordinate family is visualised at the top of the model in Figure 6.1 and entails three dominant themes: ***Evaluation***, ***Authenticity*** and ***Novelty***.

Evaluation (1,100): This is the largest single theme to emerge from the analysis and the only one present in each of the 100 reviews. It includes judgements about the value or merit of the performance as a whole, of performance temporal segments (e.g., second movement) or of performance features, as well as comments on degrees or amounts that clearly delineate a valence of the judgement. These judgements entail little or no descriptive content (e.g., "superb", "bad", "to be reckoned with", "screws up", "too much", "unduly").

Evaluation can be expressed in isolation, as a pure evaluative judgement of the performance or temporal fragments of it:

> "In total, a fine performance" (MacDonald, May 1981, p. 92)
>
> "Serkin's introduction to the final fugue is superb" (Chissell, March 1972, p. 74)

Most of the time, however, ***Evaluation*** terms are presented within a sentence as judgements of certain **Primary** or **Supervenient Descriptors**:

> "Papadopoulos has already scuppered himself with a disastrous drop in tempo for the Fourth Variation" (Fanning, November 1992, p. 152)
>
> "The Adagio, as so often with this player, lacks tenderness and is too heavy" (Robertson, April 1936, p. 18)
>
> "Her tempo fluctuations in the first movement … illuminate rather than detract from the structure" (Distler, June 2007, p. 84)

The largest sub-theme within ***Evaluation*** includes *Affective* (276) judgements that reflect perceptions of the performance or its features, focusing on the listeners' affective reaction:

> "…it is strangely moving" (Robertson, August 1950, p. 23)
>
> "I admire the delicate playing, but am not filled with a sense of wonder and serene joy" (Porter, September 1957, p. 17)
>
> "These are dauntingly patrician readings of three famous sonatas" (Osborne, May 1983, p. 49)

> "…some may find Schiff's arpeggiation of the second theme cloying" (Distler, December 2005, p. 97)

Also, here are comments reflecting perceptions of the performance which aim to add to the listener's understanding of the music:

> "Angela Hewitt … offers intelligent, stylish and often illuminating interpretations" (Distler, November 2006, p. 97)

Another large sub-theme is *Comparison* (220), in which judgements are made in relation to another performance by another (*Comparison_Performer*, 181) or the same pianist (*Comparison_Performance*, 26):

> "Though not quite up to Arrau's, he plays the so-called Les Adieux sonata most beautifully" (Fiske, November 1959, p. 68)
>
> "Brendel's reading of the Pastoral has changed - and its status has stratospherically soared - in two interrelated respects" (Osborne, February 1996, p. 75)

A further sub-theme of ***Evaluation*** is *Taste* (98). Here judgements are presented as the critic's personal perception. These comments, present in 47 out of 100 reviews, tend to be holistic, focusing on qualities of the performance at global level:

> "I personally am much more attracted by Arrau's approach but I realize that others may prefer Richter-Haaser's" (Fiske, November 1959, p. 67)
>
> "I invariably find myself won over by Ashkenazy in this sonata - no matter what formidable counter claims have come before. Yet chacun à son goût" (Chissell, March 1972, p. 74)

Finally, three minor sub-themes within ***Evaluation*** focus on judgements of *Clarity*, *Success*, and *Beauty*.

Clarity (42) relates to either technical qualities of the performance or structural clarity with which the music is portrayed.

> "…the clarity of the toccata-like part writing and off-beat accents make Brautigam's conception work" (Distler, December 2008, p. 103)
>
> "But it is Brendel who gives you the clearer semiquavers in bar 3" (Chissell, December 1970, p. 86)

Success (29) focuses on the performance as product of the pianist's achievement.

> "There is an imaginative failure here" (Osborne, August 1986, p. 49)
>
> "It seems to me that Miss Donska here succeeds without a doubt" (MacDonald, November 1964, p. 52)

Beauty (27) was coded only when the critic used the word 'beauty' or a variant:

> "The fugue is beautifully done – particularly the reprise" (Porter, June 1957, p. 19)

Authenticity (282): This is the second dominant theme within the **Evaluative Judgements** family. It entails comments built on assumptions about the composer's thoughts, the period style, and in general the existence (or not) of a valid or true interpretation of the given work.

> "Only in the Coda does Beethoven himself seem to speak for a moment" (Robertson, April 1936, p. 18)
>
> "He plays the jolly little scherzo and the difficult finale with much virtuosity - the right and only way" Robertson, April 1936, p. 18)
>
> "The discrepancy in timing looks drastic, but … it reflects nothing more than two equally valid views of the music" (Fanning, November 1992, p. 152)

Discussion often focuses on the correspondence between the performance and inherent qualities of the music that ought to be realized and that the performer did or did not achieve:

> "…the spirit of the music has been exactly caught" (Chissell, August 1963, p. 31)
>
> "The sense of mystery is missing" (Porter, October 1954, p. 51)
>
> "…one could make a case for Freire's unfolding animation as being true to this music's intended introductory function" (Distler, September 2007, p. 76)

A sub-theme within ***Authenticity*** is *Notation* (79), discussing the correspondence between performance and score indications.

> "Taub anticipates the meno allegro indication by a couple of bars and I wish he didn't apply the brakes quite so soon" (Plaistow, March 1988, p. 50).

Another sub-theme, *Historical* (16), discusses the performance in relation to the context or composition of the work or the assumed composer's intentions.

> "I am not sure the effect can ever succeed on a modern instrument, but at least Richter-Haaser's attempt is nearer the composer's wishes than Kempff's and Backhaus's total rejection of the sustaining pedal" (Fiske, February 1961, p. 48)

Novelty (120): This final dominant theme encompasses characterizations of the performance or of its features that reflect originality. It also includes comments that refer to the originality of the pianist as interpreter, highlighting interpretative style.

> "Arturo Pizarro now re-emerges on Linn Records with performances of four Beethoven sonatas sufficiently individual and freshly conceived to make them emerge as new-minted rather than over-familiar" (Morrison, March 2003, p. 63)

Critics' agreement

The previous sections have described the map of performance-related themes emerged from the analysis of the 100 selected reviews. These themes indicate aspects of performance critics discuss in their judgements. As argued in Chapter 1, within research investigating the perception and appreciation of music a major concern is the extent to which judgements may be shared between people, or the extent to which different listeners may focus on those same aspects of the performance if let free to do so. An important aspect of the emergent model is thus the level to which it can be taken as representative of a common trend among different critics.

Differences between critics in the relative use of themes may be linked to personality, musical background and reviewing and linguistic style. However, one more important factor should be taken into consideration in the present analysis. The corpus of critical review at hand entails reviews of Beethoven's piano sonatas recordings. However, each review discusses a different disc or set of discs. Looking for differences and commonalities between critics in the use of the emergent themes thus means comparing reviews that discuss *different performances*, most often of *different musical works*. A performance that lacks – say – rhythmic stability, might then trigger comments on the rhythm that may not be necessary in a performance being technically impeccable but lacking in energy.

Keeping this in mind, it is reasonable to expect a certain variety between one and the other critic, due to the fact that different critics are indeed reviewing different objects. These influences (reviewed object, personality, writing style and musical background) are compounded in the material and cannot be taken apart in the present study. What can be examined though is the extent to which, notwithstanding these confounding factors, the relative weight given to each theme is consistent between reviewers.

Table 6.1. Distribution of dominant (bold) and first level sub-themes across the 100 reviews and for each critic separately (10 reviews per critic).

	Theme	*All reviews (N = 100)*	*Robertson (n = 10)*	*Fiske (n = 10)*	*Chissell (n = 10)*	*Porter (n = 10)*	*Plaistow (n = 10)*	*Osborne (n = 10)*	*MacDonald (n = 10)*	*Fanning (n = 10)*	*Morrison (n = 10)*	*Distler (n = 10)*
Evaluative Judgements	***Evaluation***	**100**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**
	Affective	89	7	10	9	9	9	10	7	10	9	9
	Beauty	20	4	4	3	4	*0*	2	1	*0*	*0*	2
	Clarity	31	5	1	6	3	3	1	*0*	3	3	6
	Success	23	2	2	1	3	2	2	4	3	4	*0*
	Comparison	63	4	8	5	8	4	7	1	7	9	10
	Taste	47	7	7	6	4	5	5	3	3	3	4
	Authenticity	**89**	**9**	**9**	**10**	**9**	**9**	**7**	**8**	**10**	**9**	**9**
	Notation	48	4	7	4	4	6	2	2	7	4	8
	Historical	11	*0*	2	3	1	1	1	2	*0*	1	*0*
	Novelty	**59**	**7**	**5**	**8**	**4**	**4**	**5**	**3**	**8**	**8**	**7**
Supervenient Descriptors	***Style***	**95**	**10**	**9**	**10**	**9**	**9**	**10**	**8**	**10**	**10**	**10**
	Performer	80	6	6	10	8	9	7	6	10	9	9
	Historical	23	1	4	4	2	3	2	*0*	3	3	1
	Expressive	17	*0*	3	5	2	2	*0*	*0*	*0*	1	4
	Structure	**85**	**8**	**7**	**9**	**10**	**8**	**10**	**5**	**9**	**9**	**10**
	Balance	43	1	2	5	4	6	5	2	7	4	7
	Emphasis	24	5	2	1	1	4	3	*0*	2	2	4
	Journey	48	3	5	4	4	7	7	3	3	6	6
	Character	**79**	**6**	**8**	**8**	**8**	**6**	**9**	**5**	**10**	**9**	**10**
	Performer	23	1	1	2	1	1	4	0	7	5	1
	Emotion	**59**	**6**	**7**	**6**	**6**	**7**	**5**	**4**	**5**	**6**	**7**
	Performer	26	2	6	4	1	2	2	*0*	3	2	4
	Dialogue	**52**	**2**	**5**	**4**	**6**	**7**	**5**	**3**	**6**	**7**	**7**
	Understanding	**79**	**7**	**9**	**8**	**8**	**8**	**7**	**5**	**9**	**9**	**9**
Primary Descriptors	***Musical Parameters***	**89**	**10**	**10**	**10**	**9**	**9**	**7**	**8**	**9**	**7**	**10**
	Time	73	7	9	10	6	8	5	5	8	5	10
	Dynamics	37	7	3	5	5	5	0	1	4	3	4
	Colour	41	4	2	5	4	5	5	*0*	6	1	9
	Articulation	29	5	3	4	*0*	2	*0*	1	2	2	10
	Rhythm	26	3	*0*	4	4	4	3	2	3	2	1
	Energy	**72**	**4**	**7**	**9**	**7**	**8**	**7**	**5**	**6**	**9**	**10**
	Tension	11	*0*	*0*	2	*0*	2	3	*0*	1	*0*	3
	Technique	**61**	**6**	**6**	**9**	**4**	**4**	**7**	**4**	**7**	**5**	**9**
	Virtuosity	15	1	1	1	0	1	2	1	2	3	3

Note. Themes are treated as dichotomous variable: for each review, a theme was given the value of 1 if it occurred at least once in the text, a value of 0 if it did not occur in the text.

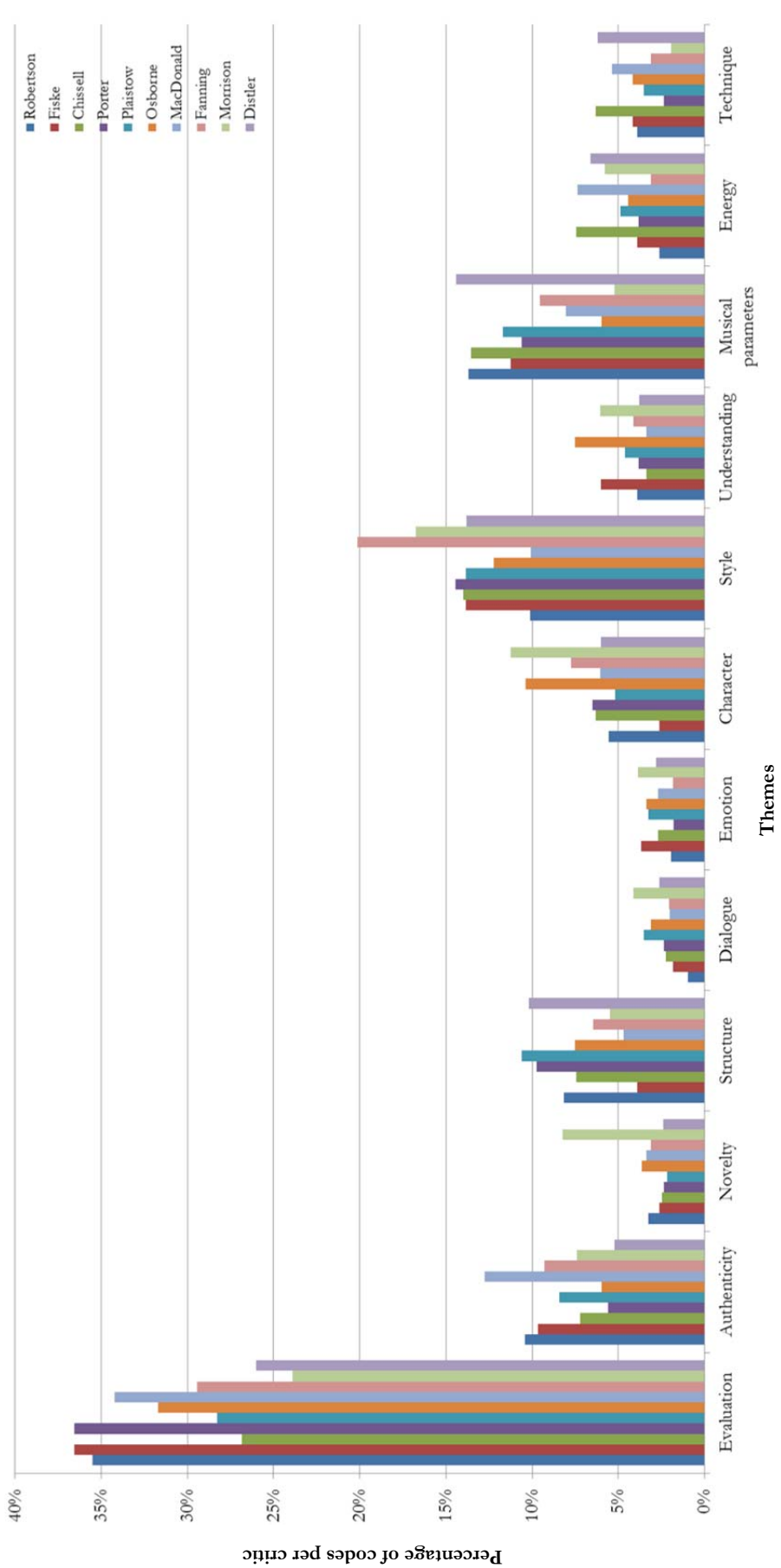


Figure 6.2. Distribution of codes across dominant themes for each critic. For each critic, the relative frequency is shown with which each theme was coded in the text.

Figure 6.1 shows the distribution of dominant and first level sub-themes across the 100 analysed reviews and for each critic separately, with themes treated as dichotomous variables: the frequency shows for each critic the number of reviews in which the theme occurred at least once. All twelve dominant themes and a third (36.36%) of the first level sub-themes were reflected in the writings of each of the ten analysed critics.

Figure 6.2 shows then the relative frequency with which each dominant theme occurred in the reviews of each critic. To partially compensate for the variety due to different performances and musical works reviewed, code occurrences for each critic across the ten reviews were added together. Following Simonton[24] (2004), Cronbach's Alpha was computed as a measure of internal consistency in the relative use of the 12 dominant themes between critics. This showed a high level of agreement, $\alpha = .986$.

DISCUSSION

The purpose of the analysis presented in this chapter was to understand the content of a representative corpus of critical review ($N = 100$) to answer the question: what do expert critics write about when reviewing performance?

The answer can be summarized in terms of **Primary Descriptors** (properties of the musical sound, level of energy, and mechanics of musical delivery), **Supervenient Descriptors** (higher-order impressions of the performance), and the value that any of these properties, or combinations thereof, possess. Although patterns between themes could not be systematically examined due to the density of the text, the emergent visual model offers first evidential support to the view of music critical review as a form of reasoned evaluation (Beardsley, 1968).

An important point is that the present model – resulting as it does from the analysis of the end-product of the critical process – does not allow us to distinguish whether critics' evaluations are inferred from **Primary** and **Supervenient**

[24] Simonton applied the Cronbach's Alpha to test internal consistency in the adjudication of awards and nominations between seven film award organizations (like the Academy of Motion Picture Arts and Sciences (Oscar). The coefficient was taken as a measure of the agreement between the seven organizations on the cinematic contribution of a given film. In the present analysis, the alpha coefficient is used as a measure of agreement between critics in the weight (in terms of frequency) given to a certain theme.

Descriptors (performance X possesses properties A, B, and C, therefore it is good) or simply connected to them (evaluation comes as instinctive response, and then reasons are sought). Further studies, focusing on the temporal component of the critical process, are needed to address this issue.

Performance properties

The different performance properties (**Primary** and **Supervenient**) identified in the model, though differently organized, reflect musical factors commonly used in performance assessment in music education (McPherson & Schubert, 2004, pp. 63-64, see Chapter 1, Table 1.2) and therefore concur generally with previous literature in this area (Bergee 1997; 2003, Thompson & Williamon, 2003; Kinney, 2009; Fiske, 1977; Wapnick, et al., 1993). In addition to these elements, however, critics' judgements also focus on ***Novelty***, artistic ***Style*** and the *Affective* response of the listener.

One aspect of performance commonly present in assessment schemes is musical expression. The notion of expression has been made object of focused investigation in Chapter 5. Findings showed that critics use this term, and its correlates, to indicate at least four different properties of performance: specific actions or qualities of the musical sound; ways of portraying the musical design; communication of higher-order qualities (such as emotions); or as an undefined, positively loaded evaluation. In addition, critics used 'expression' also to indicate qualities of the music work that should be brought out by the performer. Since the nature of musical expression emerged from this preliminary analysis as multi-layered and ambiguous, it was decided to use no *a priori* assumption on what constitutes expression, but rather let the different components of this composite construct emerge from the data. There is indeed evidence for the presence of expression in the present model, although – as expected – not as one cohesive theme. All components of expression identified in Chapter 5 are represented in the visual model, but they are interconnected with other larger themes (***Musical Parameters***, ***Structure***, ***Emotions***, ***Character***, ***Style***, ***Evaluation***, ***Authenticity***).

One trait of critical review that is quite atypical of music written responses in education settings is evaluative judgement that depends on a listener's personal perception and preferences. This trait is represented in the present visual model by

the sub-theme *Taste*. The co-existence in reviews of absolute and relative (taste-dependent) evaluative judgements may relate to the nature of the reviewed products, which – in the *Gramophone* – are all high-level professional recordings. As Levinson (2010) suggests, judgements of value in the arts cease to be meaningful at a certain level, and beyond this point individual preferences become a decisive choice criterion. This finding resonates with the observation emerged from the metadata analysis in Chapter 3 on the importance of the critic-reader relationship: critics' judgements are read as the expression of one (expert) listener's opinion, rather than as an absolute assessment. Knowing the identity of the man or woman behind the review becomes thus relevant to the understanding and interpretation of the judgement.

Overall, the present analysis has revealed a degree of overlap in the content of critical review and aspects that drive written response to music performance in education settings, while still identifying properties that are more unique to professional critical review.

The findings also permit wider reflection on the use of different aesthetic criteria in critical review. Van Venrooij and Schmutz (2010) listed indicators of high art aesthetic criteria versus popular aesthetics derived from the literature as part of their study of popular and classical reviews. High art criteria included discussion of context, the performer as creative source, comparisons with high art (masterworks), originality, complexity, seriousness and timelessness. By contrast, indicators of popular aesthetics included participatory experience (rousing, catchy) and the use of language related to 'primary' tastes, like oral and food-related metaphors (pp. 405-406). This dual categorization can also be found in the present corpus. Comparisons between interpretations and performers (high art) were found in 64 of 100 reviews, in line with findings in Chapter 3. However, popular aesthetics criteria were also common, such as those that indicate listener responses to the music (*Affective*: 89 of 100 reviews). Thus, following Van Verooij and Schmutz, it can be concluded that the present corpus of classical music critical review provides a combination of high art and popular aesthetics for the reader.

Criticism as evaluation

Another wider issue of concern in the present chapter was the debate regarding the importance of evaluation in critical review discussed in Chapter 1. In fact, ***Evaluation*** was the largest theme found to permeate critics' judgements of performances in the present model. ***Novelty*** and ***Authenticity*** were also widely spread and presented further evaluative dimensions. This result reflects the importance of evaluation in music critical review (Calvocoressi, 1923; Newman, 1925; Walker, 1968; Cone, 1981; Carroll, 2009). The finding does not concur however, with the results obtained by Conrad et al. (2005), who found that less than half of 181 music critics saw evaluation as an important element in their writing. An explanation for this apparent discrepancy lies in the variety of musical critique activities. As mentioned in Chapter 1, among the critics surveyed by Conrad et al. (2005), 53% stated that half or more of their writings were "profiles of musicians, composers and musical figures" (p. 16). In line with this, 41% of critics defined themselves not as critic, rather as "arts reporter", "music writer", "program annotator", "general assignment critic", or "entertainment writer" (p. 12).

Seen in this light, the general debate on the nature of evaluation in art criticism is limited by factors such as different media (general newspapers vs. specialist magazines) and art domains (music vs. visual arts, and within music, live vs. recorded performances). For instance, Danto's view of art criticism as a descriptive rather than an evaluative practice (Rubinstein, 2006) may reflect the fact that consumers of visual art and music are not subject to the same immediate burden of possible purchase choices, thus art consumers may not require critics to act as guides for this purpose. In sum on this point, the results of the present analysis suggest that evaluation is a major component of classical music critical reviews of recorded performances.

Performance as intentional act

Finally, a further result from the present analysis of critical review was the focus on presumed qualities of the performer, the agent of the performance. This finding – reflected in the themes grouped under **Performer Qualities** – suggests that the intentionality perceived behind performance actions play an important role in the appreciation and interpretation of a performance. This is despite the fact that these

comments are based on assumptions about the performer which, in the case of the present recordings, the critic could not even see.

This result is in line with theories of the role of intentionality, from the philosophy of art. Levinson (1996), in his discussion on 'performative' versus 'critical interpretation', argues that a person cannot reliably interpret performance actions as reflecting the critical conception of the artist, since no one-to-one correspondence can be established between the two. Nonetheless, such thoughts are common, playing an important role in our understanding and appreciation of the music. In his discussion on the interpretation of artistic works Currie (1993) calls this process 'intentional explanation' (p. 416): ascribing intentions to the artist such that his/her behaviour is viewed as depicting his/her intentions. According to Currie, intentional explanation allows us to create a coherent narrative of the work and is thus essential to our understanding. Carroll (2009) also claims that when we evaluate a performance one of the things we judge is the performer's achievement – 'success value' (p. 53). To assess this aspect we need to know what the artist intended to achieve, how ambitious his/her intentions were, what risks s/he had to take, and so on.

The importance of intentionality has implications for the understanding of the listening experience in general. In recent years there have been notable efforts in the development of computer systems for expressive music performance (for a review see Bresin & Friberg, 2013; Kirke & Miranda, 2013; Timmers & Sadakata, 2014). However, the results of the present study of critical review confirm that the opportunity to entertain thoughts concerning the person behind the performance, his/her will, decisions, emotional state and moral qualities, remains a significant part of the music listening experience.

CONCLUSIONS

This chapter reported method and findings of the first thematic analysis run on critics' judgements. Focusing on the performance-related part of reviews only, it produced a visual descriptive model of performance features that critics seek out for critical attention in their reviews. The model revealed a high level of consistency in the relative use of the different themes between critics – even between critics born generations apart. Emerging observations both confirmed and challenged common

wisdom on music criticism and performance evaluation, thus adding to our understanding of these phenomena.

However, the highly dense nature of critical review writing did not allow the analysis to be moved beyond the level of theme description to explore patterns between different themes. In particular, the model highlighted the presence in reviews of three major components: **Evaluative Judgements**, **Primary Descriptors** and **Supervenient Descriptors**. To answer the main research question, thus examining what reasons critics adduce for supporting their value judgements, it is necessary to move a step further, and explore how **Evaluative Judgements** are connected to the different kinds of **Descriptors**. What do critics wish for when they discuss *Dynamics* or ***Emotion***? Can a beautiful sound *Colour* become a value-detracting feature of a performance in the wrong context? Is *Care* always something positive, or can it slide into fussiness? To what extent do **Performer Qualities** enter the final judgement?

These questions are addressed in the following chapter, through an analysis of the valence in critical review, and the relationship between valence and the different **Primary** and **Supervenient Descriptors** of the performance.

7 VALENCE OF PERFORMANCE JUDGEMENTS

The visual descriptive model of performance judgements developed in Chapter 6 offered a map of what critics write about when reviewing music performances. The present chapter examines how the different elements identified in the model are used as reasons to support evaluative judgements.

This chapter thus focuses on the valence expressed by performance judgements, and how this relates to the different performance features critics discuss. It reports the findings of a three-step analysis aimed at answering the question: What do critics appreciate or wish for, when discussing performance property *X* (e.g., *Dynamics*, ***Emotions***)? In so doing, the analysis reported in this chapter directly addresses the main research question of the present thesis. Digging deeper into the nature of critics' judgements, it explores how the different elements discussed are linked together. The final result of this exploration is a model of performance evaluation criteria in critical review.

METHOD

Material

The text used in Chapter 6 was also the object of the present examination. This entailed performance-related statements extracted from a corpus of 100 reviews published between 1934 and 2010, written by 10 different critics (10 reviews/critic).

Analysis

The high density of the critical review text emerged in Chapter 6 did not permit an exploration of patterns between **Evaluative Judgements** and **Primary** and **Supervenient Descriptors** through an analysis of code co-occurrences. A different approach was thus sought to examine systematically the valence expressed by performance judgements. Based on the analysis in Chapter 6 and on preliminary investigations of the text, two initial observations were made: first, valence in critical review is expressed explicitly through **Evaluative Judgements**, but also implicitly

through the use of valence loaded **Descriptors** (e.g., 'nimble fingers' vs. 'overtaxed fingers'). Both levels should be accounted for in an analysis of valence in critical review. Second, to capture the valence implied in the text, a text segmentation is often required that goes beyond the fine-grained, single-clause level used in Chapter 6 to embrace several clauses or even sentences. Following these observations, a three-step analysis protocol was developed to examine the valence component of the different elements identified in the model in Chapter 6.

Valence in critical review

In a first step, performance-related review text was analysed anew and coded according to its valence content. It was decided *a priori* to use four different comprehensive and mutually exclusive valence categories: ***positive***, ***negative***, ***neutral***, and ***unclear***. After a preliminary analysis a fifth category – ***mixed*** – was added, to capture text units that entailed both positive and negative valence. Definitions for the five valence codes are reported in Table 7.1.

Table 7.1. Codebook used for the analysis of valence content in critical review.

CODE	*Definition*	*Example*
Positive	Statements with clear positive valence.	"It is played magnificently. Schnabel gives a most dramatic reading of the work, leaving us in no doubt as to its essential bigness" (Robertson, August 1934, p. 29)
Negative	Statements with clear negative valence.	"The section of the slow movement has a certain beauty which I feel Schnabel spoils by too dynamic a treatment" (Robertson, April 1936, p. 18)
Mixed	Statements entailing both positively and negatively loaded parts, which cannot be taken apart without losing the meaning of the text unit.	"The last movement, needless to say, is played in the grand manner and is undeniably exciting, but without the fine nuances of phrasing and articulation Gieseking gives us" (Robertson, October 1953, p. 22)
Unclear	This groups statements: (i) for which it is not clear if they entail some valence or not; and (ii) which seem to entail some valence, but for which it is not possible to decide if this is positive or negative.	"The final fugue is something more than a struggle against appalling odds" (Fiske, August 1963, p. 31)
Neutral	Statements that are purely descriptive, they entail no valence.	"Kempff, by the way, does not follow the Schnabel edition, so that there are some textual differences in the two performances" (Robertson, November 1936, p. 17)

The coding was performed by two researchers separately (same who performed the analysis in Chapter 6). That the second coder was also native English speaker assureed as complete a comprehension of idioms and implied valence as possible. Each researcher coded the whole text independently according to the five pre-defined categories. Segmentation was performed at the smallest multiple-clause level necessary to perceive clearly the valence of the text. Upon completion of the coding, agreement was computed. Statements that presented a lack of agreement were discussed between the two coders; researchers took turns to explain their reasoning behind the coding of each statement. This led to a revised version of the coded documents agreed upon by both researchers that offered a valence-based categorisation of critical statements.

Relationship between valence and performance descriptors

Based on these results, in a second step separate lists of valence loaded statements (those entailing ***positive***, ***negative*** or ***mixed valence***, see Appendix 8) were retrieved for each one of the **Primary** and **Supervenient Descriptors** found in the analysis of Chapter 6 (co-occurrence lists, in total 30 quote lists). These quote lists were analysed by the author to identify what aspects of each performance property (e.g., *Dynamics*, ***Emotion***, or ***Energy***) are praised by critics. This led to the development of a set of value adding qualities discussed in reviews.

Performance evaluation criteria in critical review

Finally, the emergent value adding qualities were further analysed by the author to identify higher-order evaluation criteria underpinning them, adapting the procedure proposed by Beardsley for the establishment of basic criteria of aesthetic value (Beardsley, 1968). For each value adding quality the questions were asked: 'Why is this positive? How does it add to the value of the performance?' The process was repeated until a property was reached, whose positive value could no longer be explained by appealing to features of the work itself (where 'features of the work itself' is understood broadly to embrace features the work represents, suggests, or symbolises).

For instance, 'a rich and resonant sound' can be said to be positive in that it renders the performance more intense. Explaining why the resonant sound is good can be done by appealing to a more general principle: that intensity is desirable in the

performance. Explaining on the other hand why intensity is desirable would need an explanation that goes beyond what is entailed in the performance. Therefore, following Beardsley, intensity can be taken as higher-order evaluation criterion.

Applying this procedure to all value adding qualities found in step-two of the analysis led to the development of a model of performance evaluation criteria in critical review.

RESULTS

Valence in critical review

In total, 943 text segments were coded across the 100 reviews. Percentage of agreement between the two coders was 75.95%, Cohen's Kappa = .65 ($p < .001$), 95% CI (.62 – .69), which represents a substantial agreement between coders (Landis & Koch, 1977).

The discussion of discrepancies between coders revealed that disagreements were mainly due to three reasons: ambiguities in the reading of the text, given for instance by comparative judgements or conditional statements (statements in the form 'if you like X you will like this performance'); nuances in the interpretation of the value component of words, partly due to the different perspectives and levels of familiarity with the repertoire the coders had (e.g. the characterization of a performance of Op. 2/1 as 'Haydnish' has a different valence once the reader knows that this sonata was dedicated to Haydn by Beethoven);and misjudgement of word meaning or idiomatic expressions due to the fact that one of the coder was fluent but not native English speaker (e.g., terms like 'fastidious' have a clearly negative connotation in Latin languages but not so in English).

Some of the discrepancies were solved by creating three *ad hoc* rules:

(i) Comments that are partly unclear and partly clearly positive or negative are assigned the code following the valence of the clear fragment. Example:

> "…there is plenty of matter for discussion in Schnabel's interpretations, besides lots for any pianist to learn and profit from" (Robertson, April 1936, p. 18)

Here the "matter for discussion" could be interpreted as something either positive or negative. However, the second part of the sentence is clearly positive. The whole segment was then coded as ***positive***.

(ii) Comments that come in the form 'If you like property X, then you will like/dislike performance P' or its variations are coded as ***mixed***, in that they suggest that the value of the performance is dependent upon the listener perspective or taste. These statements often correspond with comments coded under ***Evaluation***_*Taste* in the model presented in Chapter 6. Example:

> "If you like Beethoven's dynamics undefined and a presentation of him in a thoroughly unbuttoned mood you will warm to Medtner's interpretation of the first and last movements of the Apassionata" (Robertson, February 1947, p. 8)

(iii) Statements that compare two performances and do not offer enough information to understand which one is the object of the review and which is used as comparative element, are coded according to the valence of the terms used. Example:

> "I preferred Ashkenazy's for its stronger voltage and drive" (Chissell, February 1970, p. 54)

This statement could be seen as positive (for Ashkenazy) or negative (for the performer set against Ashkenazy). The terms used ("preferred", "voltage" and "drive") however carry a positive valence, therefore the sentence was coded as ***positive***.

The application of these rules and discussion/clarification of valence loaded words and expressions led to an almost complete agreement between the two coders. For two statements though, no agreement could be found: these two statements were thus excluded from subsequent analyses. Therefore, the results that follow are based on a sample of 941 text fragments.

The great majority of critical review text (87.57%) emerged as clearly valence loaded (***positive***, ***negative*** or ***mixed***), with a strong prevalence of ***positive*** comments. ***Neutral*** statements were rare and a small amount of comments were coded as ***unclear***. Table 7.2 reports the counts for the five categories.

One characteristic of critical review emerged in this analysis is the juxtaposition within the same review of positive and negative judgements. On average, each review entails 50.08% (SD = 0.26) of positive statements, 23.61% (SD = 0.23) of negative statements and 16.71% (SD = 0.19) of mixed statements. Across the 100 reviews, merely seven reviews entailed only positive statements, and just one review encompassed only negative ones. The latter is a very short review of Backhaus's Pathétique and Moonlight sonata, entailing just one performance-related sentence:

> "The performances I found disappointing, and I would suggest there exist a number of couplings of these two sonatas that are superior to this one" (Plaistow, June 1962, p. 64)

Table 7.2. Frequency of code occurrence for the five valence categories.

Valence category	*Number of coded text segments*	*Percentage*
Positive	468	49.73%
Negative	221	23.49%
Mixed	135	14.35%
Neutral	69	7.33%
Unclear	48	5.10%
Total	941	100%

Relationship between valence and performance descriptors

In this section results are reported of the 30 sub-analyses run on the co-occurrences between valence loaded statements (***positive***, ***negative*** or ***mixed***) and each one of the **Primary** and **Supervenient Descriptors** discussed in Chapter 6. For each descriptor, the aim of the analysis was to clarify the relationship between the valence expressed in the statement and the property identified by the descriptor. This led to the development of a list of value adding qualities of performance used in critical review.

In what follows, value adding qualities are presented organized by descriptor. For each performance descriptor, qualities are reported that were mentioned at least three times in the text, together with examples from the reviews. As in Chapter 6, layout and format of the text is used to highlight hierarchical relationships between descriptors; names of value adding qualities are reported in italic, not capitalized.

Results are organized in two main sections that focus on the relationship between valence and **Primary**, and between valence and **Supervenient Descriptors**, respectively. Along the whole result section, number in parentheses after descriptor names indicate how many times the descriptor was linked to a valence loaded statement.

The analysis was limited to the exploration of the relationship between valence and each single descriptor. Linear relationships between descriptors were not

investigated systematically. Recurrent connections emerged during the analysis are discussed along the presentation of the relevant descriptors.

Valence of Primary Descriptors

Out of the 719 occurrences of **Primary Descriptors** found in the critical review (Chapter 6), 470 (65.37%) are connected to a valence loaded statement. The percentage is the highest for ***Energy*** (79.00%) and lowest for ***Technique*** (29.14%). The analysis revealed six recurrent value adding qualities common to several descriptors: *appropriateness*, *clarity*, *variety*, *energy*, *control*, and *accuracy*. In addition, a series of descriptor-specific qualities were found. Table 7.3 summarises the value adding qualities found in the analysis of **Primary Descriptors**.

Musical Parameters (283):

Tempo (81): The musical parameter most often embedded in a valence loaded judgement is *Tempo*. Four themes emerged from the analysis of critics' evaluation of *Tempo*: *fast tempo*, *slow tempo*, *appropriateness*, and *balance* in tempo relationships. Often critics praise a *fast tempo* or wish for a faster one (n = 28, 34.57%). A fast tempo is praised for it facilitates a unified and fluent portrayal of the music structure (*fluency*) and an energetic, exciting performance (*energy*).

> "…his [faster] tempo helps to keep the line buoyant here and the material of the episodes belonging to the rest" (Plaistow, August 1979, p. 69)
>
> "In both the Waldstein's and Appassionata's first movements, Brautigam's fast tempi generate drama and tension" (Distler, December 2008, p. 103)

At times however a fast tempo is criticised for working against other aspects of the performance – or a slow tempo praised for facilitating those – particularly technical *clarity*, *affective power* and the conveyance of a feeling of *control* (*slow tempo*, n = 17, 20.99%).

> "…at this speed he cannot command the poetry of Solomon's wonderful interpretation" (Robertson, July 1955, p. 44)
>
> "But it is not long before this very fast tempo works against clear articulation: runs in triplets become just a blurred flourish, and fortissimo broken-octave triplets … a technical labour" (Chissell, March 1969, p. 66)
>
> "the tempo for the finale … feels a notch or two fast, especially since it is not coupled with much sense of rhythmic enthusiasm" (Fanning, October 1990, p. 116)

A further value adding quality related to *Tempo* is its *appropriateness* (n = 18, 23.46%) to the music character or score indications.

> "Gieseking tears off the first movement of the Pathétique to a tremendous pace, perhaps a little too fast to convey its tragic grandeur" (Fiske, November 1957, p. 17)
>
> "…one fears that his brisk pace for the second movement leaves Beethoven's humbler Allegretto at the starting-gate" (Distler, December 2008, p. 103)

Particular attention is given to shifts in *Tempo* between different sections of a piece or between pieces (*balance* n = 15, 18.52%). Independently from the speed, *Tempo* is praised for being steady, and tempo changes should only occur when the music score asks for them:

> "Nevertheless this last movement never really settles down to a comfortable, steady tempo" (Porter, October 1954, p. 50)
>
> "One or two mildly disturbing things happen in the F minor; for instance unaccountable, almost bizarre changes of tempi in the finale" (Fiske, November 1957, p. 17)

When *Tempo* changes occur, relationship between tempi should add to the coherence of the overall interpretation:

> "Serkin's introduction to the final fugue is superb: the unpredictable tempo changes are finely integrated" (Chissell, March 1972, p. 74)
>
> "The first movement – until the rest has been heard – may perhaps be thought a shade slow, not ebullient or starkly enough; but a bewitching performance of the Scherzo, at a very lithe gait, gives retrospective point to the earlier speed" (Porter, November 1956, p. 55)

Colour (53): This is the second musical parameter most often connected to a valence loaded statement. Four characteristics of timbre and texture of the musical sound are praised or wished for in reviews: *richness*, *variety*, *appropriateness*, and *control*.

Critics praise "richness" (Fanning, June 1989, p. 64), "warmth" (Distler, September 2006, p. 80), and "depth" (Robertson, August 1950, p. 23) of tone (*richness*, n = 22, 41.51%) as opposed to a "thin" (Robertson, October 1958, p. 65), "distant", (Distler, May 2006, p. 90) or "jangling" (Robertson, April 1936, p. 18) tone. An intersection emerges between *Dynamics* and *Colour* with emphasis given to rich, resonant sound at *f* and *ff* dynamic levels (n = 4).

> "His launching of the work gives warning of its stature – the fortissimo opening chords are richer in tone than Brendel's" (Chissell, March 1972, p. 74)
>
> "…there were times in the variations where I felt the need for … a less ungrateful forte tone" (Plaistow, June 1963, p. 36)

Another value adding quality within *Colour* is timbral *variety* (n = 12, 22.64%). Critics express a desire for a wide "palette of tone-colours" (Porter, November 1956, p. 55) and criticise performances for their "uniformity of timbre" (Fanning, September 1988, p. 80) or "sameness of tone-colour" (Porter, June 1957, p. 19).

The third theme related to *Colour* is *appropriateness* (n = 4, 7.55%). As *Tempo*, also timbral qualities have to be in line with the character of the music piece. A beautiful or rich sound is usually positive, but in some occasions it can be unsuitable to convey the music character or composer's (assumed) idea.

> "A beautiful cantabile distinguishes his playing of the F sharp minor melody beginning at bar 27, infinitely seductive but out of place here" (Robertson, November 1936, p. 17)
>
> "The result is a strange jumble of sound, but in a sense it is what Beethoven wanted" (Fiske, February 1961, p. 48)

Finally, a few times *Colour* is praised for being well *controlled* (n = 4, 7.55%), the critic appreciating the ability to use and shape timbre and texture:

> "…I'm bound to acknowledge Gilels's peerless control over tone" (Osborne, May 1983, p. 49)
>
> "There is enjoyment to be had from hearing the textures so adroitly controlled" (Fanning, September 1988, p. 80)

Dynamics (44): Two themes emerged in relation to the evaluation of *Dynamics*: *variety*, and *accuracy*.

Critics look forward to a differentiated, wide-ranged use of *Dynamics* (*variety*, n = 25, 56.82%), and criticise performers for their dynamic evenness:

> "There is much coarse music, played with those alternate loud and soft contrasts of tone that so soon become wearisome" (Robertson, April 1936, p. 18)
>
> "The finale of the A major sonata, Op. 101, suffers similarly from this pervasive, dynamic evenness: the ff at the climactic point of the fugal development (bar 223), for example, goes for next to nothing" (Plaistow, December 1961, p. 57)

Particular importance is given to the use of pianissimo as expressive tool (n = 11, 25.00%):

> "Both players miss the quiet thrill that should come from the drop into pp from forte at bar 166 of the Coda by using too much tone" (Robertson, November 1936, p. 17)

> "Brendel has the rare ability to play very quietly and to make the sound rise from almost nothing" (Fiske, August 1963, p. 31)
>
> "…because of a reluctance to drop to piano or pianissimo in the last, the Waldstein has difficulty here in catching fire" (MacDonald, January 1965, p. 59)

Accuracy (n = 8, 18.18%), reflects the attention given to the differentiation between dynamic levels and the exactness in following the score dynamic indications. Performers are praised for their "careful dynamic grading" (Porter, November 1956, p. 55) and criticised for ignoring or failing to notice Beethoven's dynamics (Plaistow, January 2002, p. 81; Fanning, April 1992, p. 111).

Rhythm (37): With respect to *Rhythm* – understood as patterns of accents (Cooper & Meyer, 1960) – critics appreciate a rhythmic pulse that is firm and even (*steadiness*, n = 13, 35.14%), and at the same time charged with strength and vitality (*energy*, n = 9, 24.32%).

> "Her way with the giant fugal finale, too … includes an end sadly out of rhythmic kilter" (Morrison, July 2010, p. 77)
>
> "…the first movement … is full of superb rhythmic energy" (Robertson, April 1936, p. 18)

These qualities are the more appreciated when balanced, so Brendel's slow pulses in his Hammerklavier performance are praised for being "far-seeing but flexibly ordered" (Osborne, April 1982, p. 66), and Kempff's Pathétique is said to deliver a "special joy" thanks to its "rhythmic drive" that never descends into "graceless flight" (Osborne, September 1995, p. 83).

Two further minor points in regard to *Rhythm* are *clarity* (n = 3, 8.11%) and the feeling of *control* on the side of the pianist (n = 3, 8.11%).

> "…he takes a few bars to define the stinging dotted rhythm with Brendel's clarity" (Chissell, March 1972, p. 74)
>
> "My other serious quibble concerns Foldes's control of the rhythmic flow of the music" (Plaistow, December 1961, p. 57)

Table 7.3. Value adding qualities emerged in the analyses of **Primary Descriptors**. Number in parentheses show how many times a descriptor was associated with a valence loaded statement (table continues on next page).

Descriptor	*Value adding qualities*	*Details*
Musical Parameters (283)	--	--
Tempo (81)	Fast tempo (28)	Fast tempo is praised or wished for as it adds to *energy* and *fluency*.
	Slow tempo (17)	Slow tempo is praised or wished for as it adds to *clarity*, *affective power* and *control*.
	Appropriateness (18)	Tempo in line with the music character and demands.
	Balance (15)	Coherent and balanced relationships between tempi.
Colour (53)	Richness (22)	Resonant, full, deep sound, in particular at *f* and *ff* dynamic levels.
	Variety (12)	Wide timbral range.
	Appropriateness (4)	Timbre in line with the music character.
	Control (4)	Ability to use and control timbral qualities.
Dynamics (44)	Variety (25)	Wide range of dynamic levels. Emphasis on the use of *pp* for expressive ends.
	Accuracy (8)	Precision in differentiating between dynamic levels and following score indications.
Rhythm (37)	Steadiness (13)	Steady and even pulse.
	Energy (9)	Strong and vital rhythmic impulse.
	Clarity (3)	Precision in the delivery of rhythmic patterns.
	Control (3)	Ability to use and control pulse and rhythmic changes.
Articulation (36)	Emphasis (14)	Effective use of accents and *sforzandos*.
	Clarity (8)	Precise and well-differentiated articulations.
	Variety (4)	Wide range of articulations.
	Accuracy (3)	Exactness in following score indications.
	Lightness (3)	Light touch.

Descriptor	*Value adding qualities*	*Details*
Expressive Timing (32)	Negative (18)	Expressive timing criticised for its negative influence on *fluency*.
	Positive (9)	Expressive timing praised for adding to the *affective power*.
	Accuracy (5)	Exactness in following score indication.
Energy (143)	Positive (119)	Basic value adding property of performance.
	Negative (10)	Energy is criticised when it is *inappropriate* to the musical context or it mars *variety*.
Tension (14)	Positive (13)	Basic value adding property of performance.
	Negative (1)	Tension is criticised for being *inappropriate* to the musical context.
Technique (44)	Appropriateness (31)	Technique put at the service of the music, as an aid in the realisation of other performance properties (e.g., *energy*, *variety*, *affective power*, etc.)
	Clarity (23)	Technical precision and cleanness.
	Assuredness (11)	Ability to control and master technical challenges.
	Brilliance (7)	Value adding property of ***Technique***.
	Energy (4)	Energetic, exciting technical delivery.
Virtuosity (14)	Positive (13)	Basic value adding property of performance. It can also be praised additionally for its *appropriateness* and *care* taken in the delivery of ornaments.
	Negative (1)	Virtuosity criticised for working against *control*.

Articulation (36): Here are three value adding criteria already encountered in the previous ***Musical Parameters***: *clarity*, *variety*, and *accuracy*, plus two new qualities concerning *emphasis* and *lightness*.

Emphasis was the most common evaluative reason for *Articulation* (n = 14, 38.89%). It refers to the ability of choosing what notes to emphasise. Accents should be used for expressive purposes, to add tension to the performance, generate drama and urgency or evidence structural relationships. Sforzandi in particular should be given proper importance. A heavy accentuation of up- or downbeat should be avoided, since this mars the force of the expressively significant points and leads to a 'square' and segmented performance.

> "…the development section's relentless left-hand arpeggios gain urgency through unusual accentuations" (Distler, September 2007, p. 76)
>
> "…some of his sforzandi, notably in the Scherzo, are understated to the point of inaudibility" (Fiske, October 1958, p. 65)
>
> "The most conspicuous shortcoming shows up straight in the opening bars of the Sonata, Op. 101 with … a touch of squareness resulting from inadvertent emphasis on upbeats" (Fanning, March 1990, p. 69)

Other value adding qualities of *Articulation* are *clarity* (n = 8, 22.23%), *variety* (n = 4, 11.11%), *accuracy* in following score indications (n = 3, 8.33%) and the use of a light touch (*lightness*, n = 3, 8.33%).

> "The difficult fourth movement is played with great power and absolute clarity of articulation" (Robertson, February 1948, p. 23)
>
> "For breathtaking variety of articulation, listen to Op. 2 No. 2's Largo appassionato" (Distler, December 2005, p. 97)
>
> "Dotted notes in the explosions of bars five and six get their dots as if were moved from their sides to their tops, so that they emerge as a detached kind of staccato" (Chissell, June 1971, p. 54)
>
> "Schnabel gives the ubiquitous semiquavers a light and almost fantastic touch" (Plaistow, March, 1964, p. 63)

Expressive Timing (32): The evaluative connotation of *Expressive Timing* – understood as temporal variations from the underpinning tempo on the local level – has been already discussed in Chapter 5, within the analysis of the notion of expression in critical review. The present results support the previous findings: the use of *Expressive Timing* is more often criticised (n = 18, 56.25%) than praised (n = 9, 28.13%) by critics. When

reasons are given for the critique, these focus on expressive timing as being disruptive of the flow and unity of the music (*fluency*), or of positively adding to the *affective power* of the performance.

> "…he makes a ritardando at the end of the A major's first movement … of dimensions far too large for the movement as a whole to sustain without structural unbalance" (Plaistow, December 1961, p. 57)
>
> "Many will find the first movement of the Moonlight also rather lacking in feeling – there is almost no rubato" (Fiske, July 1984, p. 41)

Another element in line with Chapter 5 is the connection between *Expressive Timing* and *Dynamics*, the latter preferred at times to the first as an expressive means.

> "On one point only do I feel inclined to disagree with him and that is over the accelerando he makes at each (immediate and higher) repetition of the second subject of the first movement. An increase in tone is certainly called for, but not, I feel, in speed" (Robertson, August 1934, p. 29)
>
> "Although Paik generates genuine excitement throughout the Appassionata, No. 23, other pianists also do so with steadier basic tempi and more pronounced dynamic contrasts" (Distler, October 2005, p. 81)

A few times *Expressive Timing* is judged for its closeness to the score indications (*accuracy*, n = 5, 15.63%).

> "The only inconsistency here is the surprisingly sudden plunge in the triumphant final return of the fugue subject (in left hand octaves) instead of the composer's prescribed poco a poco animation" (Chissell, June 1969, p. 53)

Energy (143): Both ***Energy*** and its sub-descriptor *Tension (14)* present a quite homogeneous construct in critical review: they are usually praised as basic value adding properties of performance. Almost all value judgements related to ***Energy*** or *Tension* are positive (n = 119, 92.25% and n = 13, 92.86%, respectively).

> "…in the final resort it is the voltage that counts in this eruptive fugue" (Chissell, March 1972, p. 74)
>
> "…his playing is never less than acute, his energy coursing like electricity from point to point, from pylon to pylon" (Morrison, June 2006, p. 71)
>
> "I have never heard the lead back to the recapitulation … realized with quite such heart-stopping intensity as here" (Osborne, February 1996, p. 75)

A few times though (n = 10, 7.75%), a performance is criticised for being too energetic or tense. These critiques are explained in terms of *appropriateness* to the music character or differentiation of musical patterns (*variety*).

> "Among the late sonatas, tension and severity serve Gulda's Opp. 109 and 110 less well than in his remarkably concentrated Hammerklavier" (Distler, September 2006, p. 80)
>
> "Only in the F major Sonata could there reasonably be room for some small doubt. For such an attack on the first of the two movements seems, and is, effective enough at the time; but it does create less of a contrast with the turbulent second movement than is ideal" (MacDonald, March 1965, p. 57)

Technique (101): When critics discuss the mechanics of musical delivery, a main distinction emerges between evaluations of ***Technique*** in relation to the interpretation (*appropriateness*, n = 32, 37.21%) and evaluations that focus on the performative value of ***Technique***. Assessed for its *appropriateness*, ***Technique*** is praised for being meaningful, aimed at supporting the musical message, or put at the service of the music.

> "He neither subjects the notes to his virtuosic will, nor demeans his own technique by mimetic attempts at audible disorder" (Osborne, December 1983, p. 84)
>
> "Sheppard is never less than eloquent, his outsize technique and personality always at the composer's service" (Morrison, June 2006, p. 71)

Comments mainly focus on the use of pedal (n = 10) and realisation of repeats or ornaments (n = 8).

> "But why on earth does he keep the sustaining pedal down throughout the two lovely recitatives in the middle of the movement?" (Robertson, April 1936, p. 18)
>
> "…he follows the frowned-upon practice of playing the glissando octaves from the wrist" (Distler, April 2007, p. 92)

However, ***Technique*** can be a value adding or detracting element of the performance also independently from interpretative issues. *Clarity* (n = 23, 26.74%) and *assuredness* (n = 11, 12.79%) are the qualities most often praised, the latter indicating the conveyance of a feeling of mastery and command of the technical challenges the music poses.

> "Also notice that Freire, like Claudio Arrau, takes trouble to make the first movement's rapid left-hand figurations clear and distinct" (Distler, September 2007, p. 76)

> "Watt's technique is very rarely embarrassed by Beethoven's demands" (Fanning, September 1988, p. 80)

Two further qualities also independent from interpretative issues are *brilliance* (n = 7, 8.14%) and *energy* (n = 4, 4.65%). *Brilliance* emerged as basic value adding feature, possibly related to the sub-descriptor *Virtuosity*:

"…Gilels plays the music with … great technical brilliance" (Osborne, August 1986, p. 49)

In relation to *energy* performers are praised for an "impetuous, angular fingerwork" (Distler, June 2007, p. 84) or for a "strong-fingered technique" (Fanning, March 1991, p. 85).

Virtuosity (14): The sub-descriptor of ***Technique***, *Virtuosity*, is praised most of the times (n = 13, 92.86%) as a value adding feature of the performance. Only one time it is criticised for challenging the *control* over tempo.

> "I was again much impressed … by the steady tempo he adopts for the A flat fugue, so easily spoilt by too much stress on virtuoso brilliance" (Fiske, February 1986, p. 52).

Valence of Supervenient Descriptors

Supervenient Descriptors build on **Primary Descriptors** in that they relate to the way **Primary Descriptors** are used. Most value adding qualities found in **Primary Descriptors** are relevant for **Supervenient Descriptors** as well. **Supervenient Descriptors** are, however, more varied and metaphorical in nature; the valence of these statements is strongly shaped by the use of valence laden terms in the characterisation of performance. Out of the 1,404 occurrences of **Supervenient Descriptors** found in critical review, 1,016 (72.36%) were used in valence loaded statements. Seven descriptors (*Balance*, *Emphasis*, *Performer Understanding*, *Control*, *Care*, *Sensibility*, and *Spontaneity*) emerged as value adding features on their own, almost always praised by critics in reviews. This section presents value adding qualities relevant to **Supervenient Descriptors**, Table 7.4 summarises the findings. As in Chapter 6, comments of ***Style***, ***Character***, ***Emotion***, and ***Understanding*** focusing on qualities of the performer, rather than on the performance itself, are discussed separately at the end of this section (*Performer Qualities*).

Table 7.4. Value adding qualities emerged in the analyses of **Supervenient Descriptors**. Number in parentheses show how many times a descriptor was associated with a valence loaded statement (table continues on next page).

Descriptor	*Value adding qualities*	*Details*
Style (375)	Appropriateness (39)	Style suitable to the music piece.
	Simplicity (11)	Straight and unaffected way of playing.
	Control (9)	Command over the music demands.
	Variety (8)	Way of playing emphasising richness of elements and patterns.
	Breadth (5)	Relaxed, poised, unhurried way of playing.
	Effort (4)	Vigour and determination gone into the playing.
	Lightness (3)	'Light' vs. 'heavy' performance is praised.
	Finesse (3)	Delicacy and skills in delivery the performance.
	Steadiness (3)	Firm and balanced performance.
Expressive (21)	Expressive inflections (9)	Use of dynamics, colour and structural *variety* as expressive tools (see Express (A), Chapter 5).
	Expressiveness (12)	Intense and effective communication of inner states (*affective power*, see Express (C), Chapter 5).
Historical (18)	Appropriateness (11)	Style in line with the music compositional background and performance practices.
	Romantic (3)	Romantic approach praised, as it adds to *energy* and *affective power*.
Character (193)	Appropriateness (86)	Character appropriate to the music piece.
	Energy (40)	Character conveying strength, intensity. Focus on drama and urgency.
	Mystery (28)	Character suggesting a mysterious, transcendental experience.
	Elegance (12)	Gracious, elegant, charming character.
	Poise (5)	Character suggesting calm and restraint.
	Risk (4)	Character pointing to the risk the performer takes.
	Character (5)	Performance praised for having character (no character specified).

Descriptor	*Value adding qualities*	*Details*
Structure (173)	Variety (21)	Portrayal emphasising contrasts and richness of musical details.
	Direction (8)	Portrayal conveying directionality and purpose.
	Breadth (7)	Spacious and relaxed portrayal of structure.
	Clarity (4)	Transparency in presenting the musical structure.
	Control (3)	Command over the different musical patterns.
Balance (54)	Positive (54)	Value adding property of performance. Portrayal that stresses unity and coherence in the music.
Emphasis (24)	Positive (24)	Value adding property of performance. Ability to give prominence to selected elements in an effective way.
Journey (18)	Fluency (14)	Portrayal perceived as smooth, fluid, freely flowing.
Emotion (88)	Appropriateness (32)	Emotion suitable to the music piece.
	Affective power (12)	Emotional intensity (no emotion specified).
	Poise (8)	Feeling of calm and inner balance.
Dialogue (50)	Clarity (21)	Directness and effectiveness of communication.
	Sophistication (29)	Refined and beautiful form of communication. Focus on poetry, lyricism, songfulness.
Understanding (137)	Insightfulness (17)	Intellectually stimulating performance. Focus on insights, meaningfulness, ambiguity, and fantasy.
	Thoughtfulness (10)	Rational quality of the performance, it reflects the thought with which it is instilled.
	Clarity (10)	Cogency and lucidity with which the music is portrayed.

Descriptor	*Value adding qualities*	*Details*
Performer Qualities	--	--
Performer Style (54)	Appropriateness (29) Effort (12) Dedication (11)	Performer's approach and attitude suitable to the music's demands. Rigour, work and determination gone into the preparation and delivery of the performance. Commitment and respect towards the musical piece.
Control (43)	Positive (42)	Value adding property of performance. Conveyance of a feeling of command, determination and assurance.
Care (34)	Positive (32) Negative (2)	Value adding property of performance. Attention and rigour in dealing with musical elements. Care can becomes excessive if it mars the music *flow* and *emphasis* by over-focusing on details.
Sensibility (23)	Positive (23)	Value adding property of performance. Sensitivity to the presence and importance of musical features.
Spontaneity (10)	Positive (10)	Value adding property of performance. Open, natural, instinctive approach to the music.
Intention (6)	Assuredness (6)	Determination and certainty in the delivery of the performance.
Perf. Understanding (91)	Understanding (91)	Comprehension of the music and discernment and/or imaginative power in its realisation.
Performer Emotion (33)	Affective power (22) Appropriateness (8) Poise (4)	Performer emotional involvement in the music. Performer felt emotion suitable to the music piece in terms of kind and intensity of the emotion. Emotional control and calm in delivering the performance.
Performer Character (27)	Appropriateness (18) Morality (9)	Performer's character suitable to the music piece. Performer's good moral principles.

Style (375): This large dominant theme gathered comments describing manners of execution. These comments are rather varied, and difficult to interpret given their particularly metaphorical nature.

About three fourth of comments on ***Style*** are valence loaded (72.67%) – the valence often implied in the terms used to describe the performance. Some judgements are plainly negative. Among these are expressions like "emasculated" (Fiske, October 1958, p. 65), "immature" (Fiske, August 1963, p. 31), "pedantic" (Plaistow, June 1963, p. 36; Fanning, March 1990, p. 69); the description of Gilels' Waldstein being an "'interesting corpse' of a performance" (Osborne, August 1986, p. 49) or that of Hess's Op. 109 being "like a reflection on the sonata rather than the sonata itself" (Porter, October 1954, p. 51).

Aspects of ***Style*** positively evaluated are *simplicity* (n = 11), *variety* (n = 8), *breadth* (n = 5), *effort* (n = 4), *lightness* (n = 3), *steadiness* (n = 3), and *finesse* (n = 3).

Performances are praised for being "simple" (Chissell, June 1969, p. 53), "without tricks and mannerisms" (Fiske, April 1959, p. 64), and "unaffected" (Porter, October 1954, p. 51). *Variety* is praised in terms of "subtle" (Fanning, April 1992, p. 111), "nuanced" (Distler, September 2006, p. 80) and "manicured" (Osborne, November 2004, p. 79) performances set against "coarse" (Robertson, April 1936, p. 18) and "generalised" (Distler, October 2005, p. 81) ones. *Breadth* indicates a way of presenting musical events that give them "time to smile, even to breathe" (Chissell, March 1975, p. 81), while *effort* points at the amount of work invested in the performance:

> "With Gilels the issues are brought out into the open, identified, and worked out with great rigour" (Osborne, August 1986, p. 49)
>
> "Buchbinder … is altogether too superficial to come into the reckoning" (Fanning, September 1986, p. 84)

Finesse, lightness and *steadiness* were also praised in a few cases (the latter two encountered also in the evaluation of *Articulation* and *Rhythm*):

> "Try Op. 28's finale for an ultimate pianistic and musical finesse" (Morrison, June 2008, p. 81)
>
> "The first movement of the 'Pathétique' is heavy" (Porter, February 1955, p. 46)
>
> "The steadiness of the Op. 101 fast movements certainly compels respect" (Fanning, March 1990, p. 79)

However, most often ***Style*** is evaluated for its *appropriateness* (n = 39). Thus a "muscular" pianism is said to suit middle-period Beethoven's sonatas (Distler, October 2005, p. 81) and a "woodwind-like élan" the fifth variation of Op. 109's

third movement (Distler, May 2006, p. 90). A "dynamic" approach is said to spoil the beauty of the slow movement of Op. 2/3 (Robertson, April 1936, p. 18) but it is praised in the first and last movements of Op. 7 (Chissell, December 1970, p. 86).
A last value adding quality discussed in relation to ***Style*** is *control*. *Control* – as *appropriateness* and *variety* – emerged already in the discussion of **Primary Descriptors**. Here, however, a tension surfaces between *control* and freedom. *Control* is generally discussed as positive (n = 9), performances are praised for their "restraint" (MacDonald, November 1964, p. 52; Morrison, May 1993, p. 74) or for not being "over-driven" (Plaistow, March 1964, p. 63). *Control* however can become excessive (n = 4), leading to a performance that is "disciplined out of existence" (Fanning, March 1990, p. 69) or "constrained" (Morrison, February 2002, p. 63).

Two small sub-themes of ***Style*** are *Expressive* and *Historical*.

Expressive (21): Here are the few passages suggesting styles that make use of *expressive inflections* (n = 9) or that are generally described as expressive (*expressiveness*, n = 12), according to the distinction done in Chapter 5, as to indicate an intense, convincing, and skilful form of outward expression of (unspecified) inner states. Findings support the results of the previous analysis of expression: when discussing the use of expressive inflections – as seen also in the discussion of **Primary Descriptors** in this chapter – performers tend to be criticised for the use of *Expressive Timing* as expressive means (n = 3), and praised for their focus on *Colour* and *Dynamics variety*.

> "The slow movement left me wishing that he had not relied so much for expression on rhythmic flexibility, but had sought it instead in a melodic contour shaped by subtle dynamic gradation, pure and simple" (Chissell, June 1971, p. 54)

While when they describe the performance as expressive or inexpressive in general terms, without any specifications, this is used as a value adding feature on its own.

> "…the playing is very expressive" (Fiske, August 1963, p. 31)

Historical (18): Manner of execution linked to different practices and historical periods have generally to be appropriate (*appropriateness*, n = 11, 61.11%) to the work performed, and/or coherent to the overall interpretation. The two

historically related styles usually discussed in reviews of Beethoven's sonatas are the Romantic (n = 3) and the Classical (n = 2). A Classical approach is admired for its restraint (*control*), the Romantic one for its *energy* and *affective power*. For example a "classical poise" is praised in the Adagio of Op. 31/2 (Chissell, June 1992, p. 66) while in the Allegretto of the same sonata it is said that the "blaze of romantic fire …would have won Beethoven's hearty applause and approval" (Morrison, March 2003, p. 63).

Comments on *Historical* ***Style*** may require a certain musical knowledge on the side of the reader for them to be interpreted. For instance, the assertion that Kempff's playing of Op. 111 "makes some parts of the music … sound like a Chopin nocturne rubato and all!" (Robertson, November 1936, p. 17) may be read as a negative judgement. However, "the Haydnish feeling about the exposition of the opening movement of Op. 2 No. 3" (Plaistow, August 1979, p. 69) can be taken as an appreciative statement, if it is known that this sonata was indeed dedicated to Haydn by the composer.

Character (193): A large majority of characterisations of the performance in terms of mental and moral qualities of an individual or of an atmosphere was imbued with positive or negative valence (78.46%). As for ***Style***, also the sub-category of ***Character*** entailing comments on the performer is discussed in a separate section.

Critics describe performances through a variety of characters and atmospheres, from "lugubrious" (Chissell, March 1972, p. 74) to "heroic" (Robertson, August 1950, p. 23), from "turbulent" (Porter, October 1954, p. 50) to "hesitant" (Plaistow, October 1989, p. 98). Five times performances were praised for being characterful, or criticised for their lack of character. When a character is specified, the criterion against which this is most often set is *appropriateness* (n = 86, 51.80%). In terms of what characters critics praise, the two most often recurring qualities are *energy* (n = 40, 24.10%) and states and atmospheres that suggest a transcendental experience (*mystery*, n = 28, 16.87%). Within energy critics focus particularly on the drama (n = 28) and urgency (n = 12) the performance conveys.

> "His launching of the work gives warning of its stature – the fortissimo opening chords are … majestic in their urgency" (Chissell, March 1972, p. 74)
>
> "If anything is missing, it is the sense of tragic pathos" (Osborne, November 2000, p. 86)

> "Paik imbues the transition into the recapitulation with appropriate mystery" (Distler, October 2005, p. 81)
>
> "It is afterwards, in the variations, when the light should dissolve into one that is not of this world, that chinks of common daylight reappear to disturb us" (Porter, October 1954, p. 51)

Other characters recurrently praised are *elegance* (n = 12) and those suggesting an element of *risk* (n = 4). Performances are appreciated for their "grace" (Morrison, December 2002, p. 72) and "elegance" (Distler, October 2005, p. 81); their "perilous spirit" (Fanning, November 1986, p. 78) and "reckless, all-or-nothing mood" (Osborne, December 1983, p. 84). A last criteria – *poise* (n = 5), also found in ***Emotion*** – seems to resonate with the *control* and *assuredness* criteria emerged in the analysis of **Primary Descriptors**. Performers are admired for being "unhurried" (Porter, May 1958, p. 16) and criticised for being "hectic" (Plaistow, June 1963, p. 36).

Structure (173): Comments on the way in which the performer portrays the design of the music build the second largest dominant theme within the family of **Supervenient Descriptors**. A 62.45% of these comments are valence loaded. Within the dominant theme ***Structure***, three major properties praised by critics are *variety*, *direction*, and *breadth*.

Variety (n = 21, 27.27%), already encountered in the analysis of several **Primary Descriptors**, is imbued here with a wider meaning, indicating a portrayal of the music structure that highlights contrasts and celebrates the richness of musical details. Performers are praised for presenting a "variety of perspectives, from huge vistas to tiny units" (Plaistow, August 1979, p. 69), for their "nuances of phrasing" (Robertson, October 1953, p. 22) and "multi-hued conceptions" (Distler, September 2006, p. 80). Performances rich in elements of "contrasts" (e.g., Porter, November 1956, p. 55; Distler, December 2008, p. 103) are preferred to performances characterised as "streamlined" (Chissell, March 1975, p. 81), "flat" (Porter, February 1955, p. 46), "smooth" (Plaistow, December 1961, p. 57) or "four-squared" (Distler, October 2009, p. 88).

Direction (n = 8, 10.39%) indicates a portrayal of the musical events that conveys a feeling of purpose and directionality. The performer takes a wider perspective and shows how the elements are linked together, specifically, how they follow one another and build together.

> "Schnabel's great gift … of letting us perceive the growth and design of the music stands him in good stead" (Robertson, April 1936, p. 18)
>
> "…how exciting Backhaus is as he works towards the climax" (Porter, June 1954, p. 42)
>
> "…a masterclass in steady cumulation" (Distler, September 2007, p. 76)

Breadth (n = 7, 9.10%) also found in ***Style***, includes metaphors that describe the portrayal of ***Structure*** in terms of "spaciousness" (Plaistow, March 1964, p. 63) with which the musical events are presented. Performers are praised for a "relaxed" presentation of the events (Plaistow, June 1963, p. 36) or they are criticised for "telescoping phrases" (Morrison, July 2010, p. 77), offering a "clipped statement" of the musical argument (Fanning, April 1992, p. 111), and not giving time to the elements to articulate themselves (Robertson, November 1936, p. 17).

Beside these three criteria, a few times the portrayal of ***Structure*** is praised for its *clarity* (n = 4, 5.19%), or *control* (n = 3, 3.90%).

> "…the playing of the part-writing in the prestissimo is beautifully clear" (Robertson, February 1937, p. 19)
>
> "…Gilels' mastery of the music's asymmetric lines" (Osborne, December 1983, p. 84)

Three sub-themes of ***Structure*** are *Balance*, *Emphasis*, and *Journey*.

A portrayal of the musical structure that stresses coherence and unity in the music (*Balance*, n = 54) or brings to the fore specific elements or details of the music (*Emphasis*, n = 24) is always discussed as value adding feature of the performance in reviews. *Emphasis* already emerged in the discussion of **Primary Descriptors** in relation to *Articulation*; here this concept applies in a wider perspective, to embrace *emphasis* of harmony or structural elements. Here as well, to assure that important elements are brought to the fore, a selection needs to be made regarding what details should be given priority. When no selection is done, but everything is 'emphasised', no hierarchy can be perceived anymore and paradoxically the emphasis dissolves.

> "…Gelber is again finding sunshine in every diatonic seventh, storm-clouds in every minor triad, and the broader lines of thought which distinguish Beethoven from your average Early-romantic are little in evidence" (Fanning, June 1989, p. 64)
>
> "…in getting to the heart of these matters he shows how important it is to pare away the inessentials" (Plaistow, October 1989, p. 98)

The last sub-theme of ***Structure*** is *Journey* (n =18), in which the portrayal of ***Structure*** is described as a dynamic process. These comments show a weaker connection with evaluative statements; only 23.38% of them are valence loaded. In most cases (n = 14, 77.78%) critics praise the *fluency* of the performance. Positive qualities are "momentum" (Fiske, November 1959, p. 68), "flow" (MacDonald, November 1962, p. 52), and "fluency" (Osborne, March 1993, p. 73). As discussed in the **Primary Descriptors**, *Expressive Timing* is one ***Musical Parameter*** that can easily spoil musical *fluency*.

Understanding (137): Most comments on the performance and its realisation that reflect reasoning and use of intellect are valence loaded (80.12%). A large portion of them (n = 91) concerns the *Performer Understanding*, and will be discussed separately.

Performances are appreciated mainly for their *insightfulness* (n = 17, 36.96%). This includes comments on the performance being insightful (n = 5), meaningful (n = 4), or even attractive in its ambiguity (n = 3), and element of fantasy (n = 5).

> "…the slow one [movement] is meaningful" (Fiske, November 1959, p. 68)
>
> "I would cite the development section of the F major Sonata and the entire first and third movements of the D major as models of insightful Beethoven playing" (Fanning, April 1992, p. 111)
>
> "It is, in fine, an absorbing and ambiguous reading" (Osborne, December 1983, p. 84)
>
> "…but S.'s slight element of fantasy [is] exactly right" (Robertson, November 1936, p. 17)

In addition, critics admire the *thoughtfulness* (n = 10, 2.74%) with which the performance is instilled and the *clarity* of understanding it reflects (n = 10, 21.74%).

> "Yet do not think that this is less than a thoughtful and remarkable performance" (Porter, October 1954, p. 51)
>
> "At times it is a model of lucidity, arguments and textures appearing as the mechanism of a fine Swiss watch must do to a craftsman's glass" (Osborne, December 1983, p. 84)
>
> "Listeners will notice, for example, the Op. 101 first movement's cogent voice-leading" (Distler, October 2009, p. 88)

Emotion (88): Most of the times (86.27%) when critics characterise the performance in terms of affective states, they do so within a valence loaded judgement. Occasionally these characterizations hint at qualities of the performer; these will be discussed later on. Critics praise performances that are "emotionally intense"

(Distler, December 2008, p. 103), "passionate" (Chissell, February 1983, p. 52), or "charged with feeling" (Osborne, February 1996, p. 75) (*affective power*, n = 12, 21.82%).

When a specific emotion is discussed, its value depends on its *appropriateness* (n = 32) to the piece. So Backhaus is criticised in his performance of Op.111 for failing "to discover the full peace … of the final movement" (Robertson, August 1950, p. 23) and Sheppard is praised for conveying convincingly the "music's cold fury" (Morrison, June 2006, p. 71). Even emotions that are generally negatively loaded, like impatience, can be welcome in the appropriate context:

> "The Minuet is sturdily played, though not without sudden Beethovenish touches of impatience. …this is all admirable" (Porter, May 1958, p. 16)

One emotion that is recurrently praised by critics is *poise* (n = 8, 14.55%). Critics praise performers for their "poise" (Chissell, June 1969, p. 53), "inner repose" (Osborne, November 2000, p. 86), "controlled calm" (MacDonald, January 1965, p. 59) or "feeling of controlled abandon" (Distler, May 2006, p. 90).

Dialogue (50): Comments on the communicativeness of the performance that are valence loaded (54.35% of the total) focus mainly on two evaluation criteria: *clarity* and *sophistication* of the communication. Critics praise *clarity* (n = 21, 42.00%) of communication using terms like "telling" (Chissell, March 1969, p. 66), "eloquent" (Porter, May 1958, p. 16), "direct" (Fiske, August 1963, p. 31), "immediate" (Fanning, November 1992, p. 152), or "expounded" (Chissell, June 1969, p. 63).

On the other hand, a performance can be criticised for the presence of "solecisms" of for a lack of clarity in conveying the message:

> "There's a failure of communication somewhere" (Plaistow, July 1966, p. 47)
>
> "…the artistic message itself becomes blurred" (Fanning, November 1986, p. 78)

When critics evaluate the *sophistication* (n = 29, 58.00%) of communication, they praise performers for their "poetry" (Fiske, July 1955, p. 44), "lyrical" qualities and "songfulness" (Plaistow, October 1989, p. 98).

These two aspects of artistic communication, *clarity* and *sophistication*, are both essential for a performance. So Gulda is praised for conveying a sense of "musical sophistication beneath his outwardly plain-speaking surface" (Morrison, December 2002, p. 72). While a clear form of musical communication that lacks sophistication

is "rather like listening to a fine elocutionist as opposed to a fine actor" (Fanning, September 1988, p. 80).

Performer Qualities

With the exception of one sub-theme of *Performer Style*, *Intention*, all themes within *Performer Qualities* are strongly bounded with evaluative judgements. On average, 81.86% of *Performer Qualities* statements are valence loaded. This percentage reaches above 95.00% for the sub-themes *Control* and *Sensibility*.

Performer Style (54): The performer's attitude towards or approach to the work has to be *appropriate* (n = 29, 53.70%) to the musical demands. So Barenboim is said to be "always a little too ready to yield" for "a composer with a backbone like Beethoven's" (Chissell, June 1969, p. 53) and Kempff's "amiability" in his performance of Op. 111 is said to "have no proper place in a work of this calibre" thus "vitiate[ing] much of K.'s performance" (Robertson, November 1936, p. 17).

Beside *appropriateness*, two more aspects of the performer's attitude emerged as relevant in critics' evaluations: *dedication* to the music, and *effort*. Performers are criticised for their "self-indulgence" (Chissell, June 1971, 54) and praised for showing commitment and respect towards the work, and to be ready to put their resources at the music's service (*dedication*, n = 11, 20.37%).

> "Solomon played this movement with immense reverence, as though he thought it the greatest piano music in existence; his performance is an occasion" (Fiske, November 1959, p. 68)
>
> "He does moderate his approach to suite the more intimate scale of the Op. 78 Sonata" (Fanning, March 1991, p. 85)

Beside *dedication*, a further criterion is the possibility for the critic to perceive the work, rigour and seriousness the performer invested in the playing (*effort*, n = 12, 22.22%) – criterion encountered also in the evaluation of ***Style***. Critics praise the performer's "concentration" (Fiske, July 1955, p. 44; Fanning, September 1990, p. 116), "engagement" (Morrison, June 2006, p. 71), "professionalism" (Morrison, July 2010, p. 77) and "searching" attitude (Chissell, December 197, p. 86; Fanning, November 1986, p. 78). At times though, *effort* can be excessive, leading to performances that sound "forced"

(Plaistow, July 1966, p. 47), the performer giving "the impression of standing outside the piece and of pressing through it" (Plaistow, August 1979, p. 69).

Five sub-descriptors of *Performer Style* are *Control*, *Care*, *Sensibility*, *Spontaneity*, and *Intention*.

Control (43): A 97.67% (n = 42) of comments on *Control* praised the performer for his/her aesthetic and technical command of the performance (or criticised him/her for a lack of command). One single time a performer was criticised for being "too obviously masterful" (Osborne, December 1983, p. 84).

Care (34): As *Control*, also *carefulness* in dealing with aspects of the music was discussed almost exclusively as a value adding feature of the performance (94, 12%, n = 32). Only two times a performer was criticised for excessive *carefulness* that detracted from the music *fluency* and *emphasis*: so Barenboim is claimed to be "too hung up on details" (Fanning, September 1986, p. 84) and Arrau is criticised for his "anxiety that no point should be missed" that led him to over-emphasise details (Plaistow, July 1966, p. 47).

Sensibility (23): The performer sensitivity to the presence and importance of diverse musical features was always praised as value adding feature of the performance.

> "In Op. 110 he is most exquisitely sensitive to the phrases" (Porter, October 1954, p. 51)

"In the slow movement of the Hammerklavier at the end of the first section (bar 27) he writes espressivo, and two bars later at the start of the next section con grand' espressione. If such remarks produce no response in the pianist, then in my view he is wasting our time. They invariably produce a response in Alfred Brendel" (Fiske, August 1963, p. 31)

Spontaneity (10): The performer's instinct and inclination to act in a certain way emerged also as value adding feature; critics either praised the performer's spontaneity or – more often – criticised their being

"stilted" (Morrison, February 2002, p. 63), "too deliberate" (Chissell, October 1980, p. 71), "not spontaneous enough" (Distler, May 2006, p. 90).

Intention (6): Comments on the performer's intentions, preferences and decision processes were rarely attached to an evaluative judgement (13.64%). These comments seem to be used by critics to help make sense of what they hear in the performance, but without evaluating this aspect directly. A few times though, comments on *Intention* focused on *assuredness*, in terms of the determination with which the performer acted.

> "…in the finale the fugue goes particularly well, Miss Donska achieving the flow and conviction not altogether conveyed earlier on" (MacDonald, November 1964, p. 52)

Five out of six times *assuredness* was assessed as a positive aspect; one time the performer was criticised for being "more determined than graceful" (Chissell, March 1969, p. 66).

Performer Understanding (91): Performers' comprehension of the music and discernment and imaginative power in its realisation are sought and praised by critics in their evaluations. Critics praise the pianist's "imaginative penetration" (Chissell, February 1970, p. 54), "wisdom" (Chissell, March 1972, p. 74), "intellectual control" (Plaistow, October 1989, p. 98), "terrier-like grasp of argument" (Morrison, December 1983, p. 84), "overall grasp" (Fanning, April 1992, p. 111), "insight" (Morrison, July 2010, p. 77), and "stylish perception" (Distler, December 2005, p. 97).

Critiques to the performer's understanding are often expressed in terms of different (possibly equally valid) interpretations, personal preferences, or misunderstanding between performer and listener:

> "I did not always agree with his view of it" (Fiske, November 1959, p. 68)
>
> "It may be that I fail to perceive what he is trying to do and judge him unfairly; but I can only say that the record disappointed me" (Plaistow, July 1966, p. 47)

Performer Emotion (33): Comments on the performer's affective states most often do not specify any kind of feeling, but rather praise the performer for being "passionate" (Fiske, October 1958, p. 65; Distler, September 2007, p. 76), for playing "with very deep feeling" (Chissell, March 1972, p. 74) or "affectionately" (Plaistow, March 1964, p. 63; Distler, September 2006, p. 80). Or, they criticise a player for being "unmoved" (Fiske, November 1959, p. 67), "cold" (Porter, May 1956, p. 49), or "sedate" (Distler, December 2005, p. 97) (*affective power*, n = 22, 66.67%).

When a feeling is specified, this is evaluated for its *appropriateness* (n = 8, 24.24%). A feeling can be inadequate for a given work, or inappropriate because excessively intense. Thus Kempff is praised for playing "with the right tender feeling" (Robertson, February 1937, p. 19), while Goode fails to convince the critic due to his "excesses of enthusiasm" (Fanning, March 1990, p. 69).

Other times the critic focuses on the feeling of calm and emotional control the performer shows in coping with the performance (*poise*, n = 4, 12.12%):

> "In Op. 110 Serkin seems to have regained poise" (Plaistow, October 1989, p. 98)
>
> "…she becomes flustered and hysterical in the great Adagio e sostenuto from Op. 106, almost as if she had lost patience with music" (Morrison, July 2010, p. 77)

Performer Character (27): Two aspects of *Performer Character* emerged from the analysis: moral qualities that characterise the performer's approach to the music, and mental qualities with which the performer imbues the music.

The first group (*morality*, n = 9, 33.34%) entails qualities that point to good moral principles on the side of the performer:

> "Bernard Robert is a Beethoven interpreter of sterling integrity" (Osborne, November 1995, p. 146)
>
> "K. stands bravely up to the welter of notes" (Robertson, November 1936, p. 17)
>
> "…Gulda's fluence, grace and honesty" (Morrison, December 2002, p. 72)

While the character discussed in the second group of comments (n = 18, 66.67%) is evaluated for its *appropriateness* to the music. For instance, Gulda's "mix of severity and inwardness" in the Maestoso of Op.111 is said to be "enthralling" (Morrison, December 2002, p. 72) while Ogdon's severity in the Moonlight central allegretto is criticised for leading to a performance that sounds like "a clump of nettles" (Fanning, November 1986, p. 78).

Performance evaluation criteria in critical review

On completion of the 30 sub-analyses reported in the previous section, 35 recurrent value adding qualities of performance could be identified, frequently adduced in critical review as reasons to support evaluative judgements. In the final step of the analysis these 35 value adding qualities were grouped into seven higher-order properties, employed as criteria of value in critical review: **intensity**, **suitability**, **coherence**, **complexity**, **sureness**, **comprehension**, and **endeavour**. The use of the seven criteria was consistently spread among critics (Cronbach's Alpha $\alpha = .928$), Figure 7.1 shows the frequency with which each criterion was coded in the text, for each critic separately.

Three of these properties, **intensity**, **complexity** and **coherence**, are aesthetic related: they describe properties of the (perceived) musical sound and how this is organized in time. Three more criteria, **sureness**, **comprehension** and **endeavour**, are achievement related: they point at elements of the preparation and delivery of the performance that can be derived/assumed through an interpretation of what is aurally perceived, but are not a description of the sound of the performance. One more criterion, **suitability**, indicates the extent to which each of these criteria is desirable in a given musical context. Figure 7.2 summarises the emergent model.

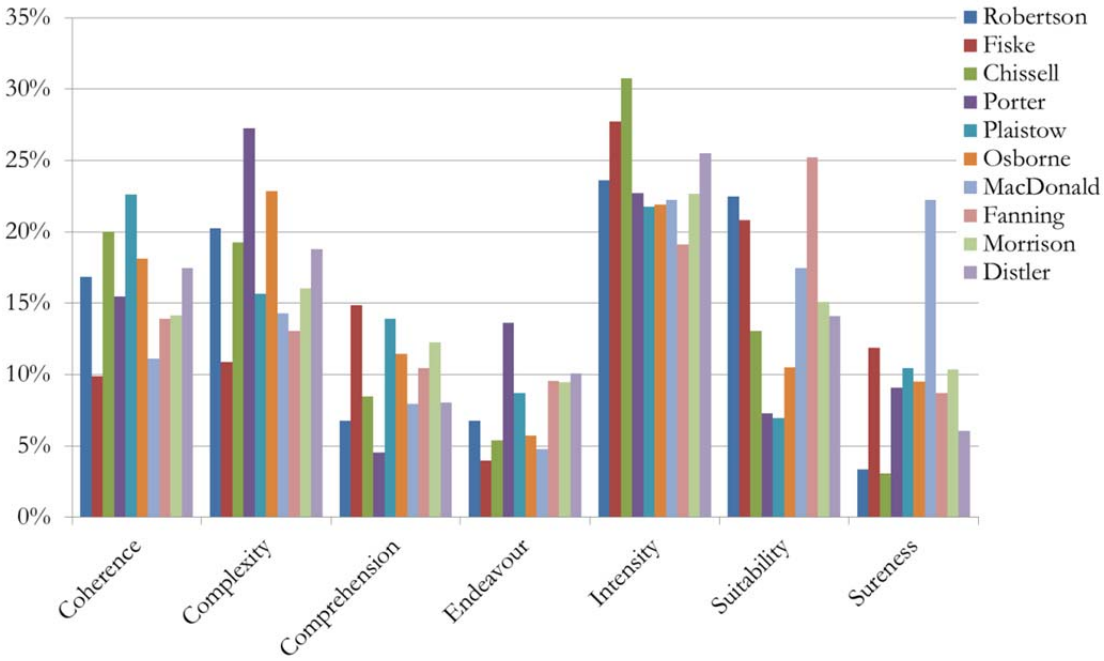


Figure 7.1. Distribution of codes across evaluation criteria for each critic. For each critic, the relative frequency is shown with which each criterion was coded in the text.

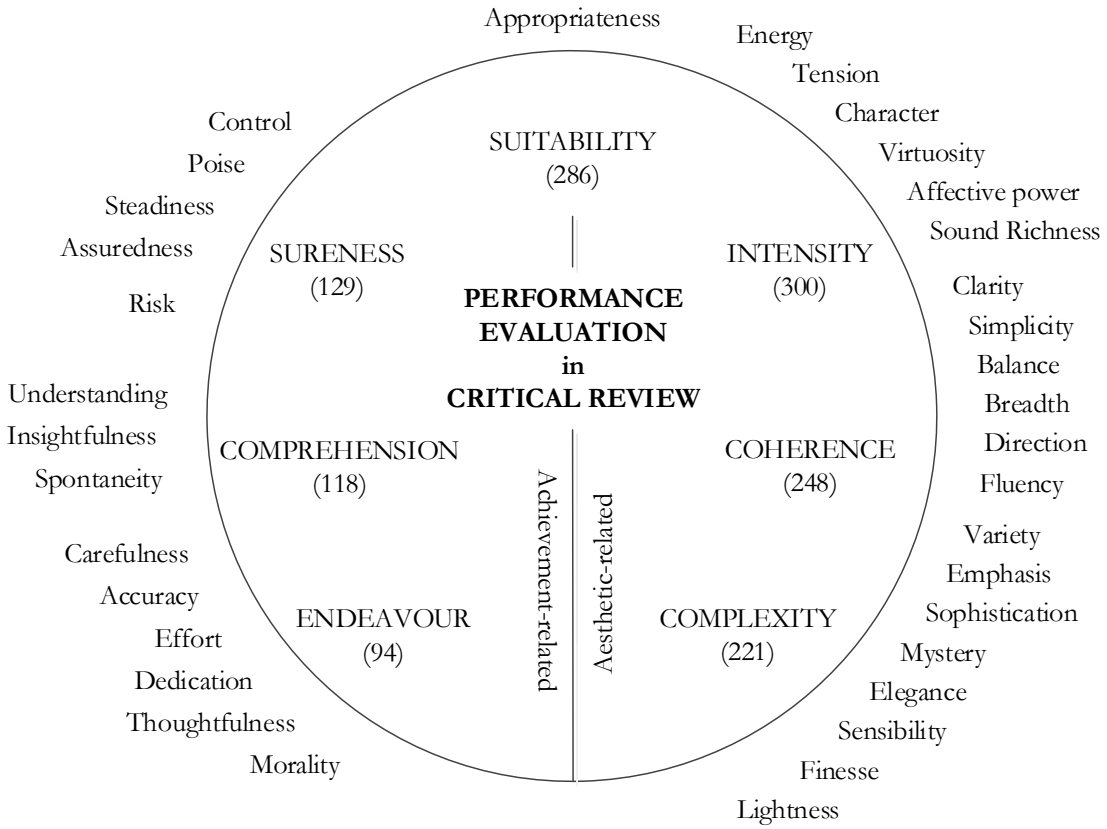


Figure 7.2. Criteria of performance evaluation emerged from the analysis of the relationship between valence and performance descriptors.

The evaluation criteria are visualised in a circular shape to emphasise a major characteristic of the model emerged from the analyses: value adding properties of performance only maintain their positive valence as long as they are counterbalanced by other properties, or they do not mar them.

Given the density and complexity of the data at hand, it was not possible to investigate systematically all combinations of properties in reviews. However a tension clearly emerged between the different evaluation criteria. During the analyses, the term **tightrope** was employed to point at these tension zones: as if walking on a thin rope, performers are required to balance the different poles within the circle of evaluation. So an increase in **complexity** in terms of *variety* and *emphasis* is highly appreciated but only insofar as it does not taint *fluency* and *direction*, essential for the **coherence** of the performance. While a high level of **intensity** – reached for instance through a fast tempo – is positive, but only as long as it does not mar **coherence**, in terms of *clarity* and *breadth*. On this line goes for

example Fiske's complaint for "Solomon's more tempestuous but less precise and considered performance" (Fiske, February 1961, p. 48). Even **suitability** does not seem to be completely immune to this kind of tension, and Badura-Skoda's attempt at a restraint presentation of the Moonlight first movement, even if in accordance with the score indications, does not win the critic's approval:

> "Badura-Skoda has planed it down in accordance with Beethoven's indications, but unfortunately all character has been lost in the process." (Porter, February 1955, p. 46)

The tension between evaluation criteria becomes palpable when a performer succeeds in 'walking on a tightrope', finding equilibrium between two or more areas of evaluation. When this happens, the balance between the different properties seems to become an added value for the performance. Among the examples of successful tightrope walkers are Wührer's, Gilels' and Gulda's balance of **intensity** and **sureness**:

> "…the prestissimo is impetuous but not undisciplined" (Porter, June 1957, p. 19)
>
> "…his extraordinary technique allows the music's evident ferocity to be tempered by Orphic assurance" (Osborne, December 1983, p. 84)
>
> "…this is, arguably anyway, exactly how Beethoven originally heard the music ringing out in his mind's ear; with restraint and energy balanced" (MacDonald, January 1970, p. 56)

Backhaus' combination of *affective power* and *simplicity*:

> "…the Arietta, too, is played with the most beautiful tone that laps the listener in loving-kindness yet it is withal simple and unaffected" (Porter, October 1954, p. 51)

Pizarro's merging of *understanding* and *sensibility*:

> "…he balances sense and sensibility to an ideal degree" (Morrison, March 2003, p. 63)

Brendel's blend of *affective power* and *understanding*:

> "The great virtue of this player seems to me his combination of emotion and intellectuality" (Fiske, August 1963, p. 31)

Solomon's marrying of *clarity* and *energy*, *virtuosity* and *sensibility*:

> "Then comes what I can only call a sensational final movement rushing along at great speed, but with perfect clarity" (Robertson, October 1945, p. 16)
>
> "This is, indeed, the real Solomon: virtuosity married to the finest sensibility" (Osborne, November 2002, p. 86)

And Richter-Haaser's balance between **intensity** and **complexity**, that wins the critic's applause:

> "The early C major … is despatched with a virtuoso address which by no means excludes delicacy in the appropriate places; and the E flat Sonata, too, gains much from the pianist's strength, agility, and poetry alike. … The virtues of the playing are very substantial indeed; these hardly could be exaggerated." (MacDonald, March 1965, p. 57)

DISCUSSION

The present chapter moved the analysis of performance judgements in critical review a step deeper, investigating the valence of review statements and its relationship with the different descriptors identified in Chapter 6. Through a large three-step qualitative analysis it tackled the main research question of this thesis, looking at the reasons critics use to support their value judgements.

The overview of valence in critical review supported and deepened the findings of Chapter 6 in respect to the importance of evaluation in music criticism. Almost all text excerpts analysed were valence loaded, even if the valence was strongly mixed within each review, with a combination of positive, negative and mixed statements.

Alongside valence loaded judgements given in the canonical form 'Performance P is good/bad because of feature F' – where an **Evaluative Judgement** is linked to a **Primary** or **Supervenient Descriptor** – numerous judgements were found in the form 'P is X', where X is a performance descriptor, usually a **Supervenient Descriptor**, that also implies an evaluation by being inherently valence loaded. A systematic analysis of both forms of judgement led to the development of a performance evaluation model entailing seven areas of value adding features.

A successful performance – as it emerges from this model – is one charged with power and technical as well as expressive intensity, rich in its complexity but unified and coherent, instilled with dedication and rigorous, thoughtful work; a performance that conveys a feeling of mastery, assurance, and conviction and a deep understanding of the music. Above all, the successful performance emerged as one that balances all these elements in a delicate equilibrium, while accounting for the musical, historical and cultural frame, thus tempering the different elements with a perceptive awareness of what is suitable and appropriate in any given context.

These findings, for the first time, offer an empirically developed model of performance evaluation in critical review and bear both pedagogical and conceptual implications.

General validity of performance evaluation

Implications of the proposed model for current theories on criticism and music performance evaluation go to the heart of the long-lasting debate on the validity of

Beardsley's aesthetic generalism (Beardsley, 1962; 1968; 1982; Walton, 1970; Dickie 1987; 2004; Bender, 1995; Connolly & Haydar, 2003; Bergqvist, 2010). Among the seven criteria of evaluation identified in reviews, three major ones – **intensity**, **complexity** and **coherence** – resonate with Beardsley's proposed triad of general principles of aesthetic values already discussed in Chapter 1 (Beardsley, 1962; 1968; 1982). According to Beardsley, intensity, complexity and unity are the only primary positive criteria of aesthetic value – that is, criteria that are universally valid, so that an increase in one of them, all the rest being equal, will always add to the value of any given artwork. Even though in the present model 'coherence', and not 'unity' was used to better capture the variety of qualities emerged in the analysis, the correspondence with Beardsley's trinity is evident.

Critics of Beardsley's theory pointed mainly to the context dependency of aesthetic properties, and the consequent impossibility to find any property that may universally act as value-making feature. In turn, these defects were used to suggest that a valid form of critical evaluation in the arts (i.e., evaluation grounded in valid or adequate reasons) is impossible (Dickie, 1987).

The present model supports the hypothesis that intensity, complexity and coherence are basic value-adding properties of music performance, related to its aesthetic value, that is, to its value as a work of art. However, findings also emphasise the context dependency of these value-making properties, both in terms of musical, historical, and performance practice background (reflected in the model's **suitability** criterion) and in terms of combinations of features within a performance (reflected by the tightrope tension between the evaluation areas). All seven criteria were found to be consistently spread among different critics, despite the different periods of publication and sonatas and performances reviewed. Thus, the present analysis shows that – at least for the corpus of review at hands – there are no generally valid aesthetic principles for supporting evaluative judgements. However, there are criteria reliably used by different listeners in different contexts. These criteria function as generally valid value adding properties under the condition of being appropriate to the given musical context, and of not impairing other value adding criteria.

These findings support Carroll's (2009) and Sibley's (in Dickie, 1987) proposed context-aware generalism, advanced in response to critiques of Beardsley's

theory. According to Carroll and Sibley, for a valid form of reasoned evaluation to be possible, it is only necessary to find aesthetic principles that are inherently positive or negative taken in isolation and general enough to be valid within a given artistic context. The proposed model identifies value-making properties that are generally valid *within the context of music performance for the chosen repertoire and cultural background*. Each criterion, taken in isolation, is a merit for the performance; its value however can be mitigated, reversed, or – as it is suggested by tightrope statements – even increased through the combination with other criteria and the contextualization in a specific music piece or section of a piece and performance culture.

In the light of this, the validity of the proposed model is necessarily bound to the context and object of the judgements, the more so the deeper the examination of judgements becomes. The prominence of **suitability** within the set of seven evaluation criteria, for example, is clearly bound to the music genre and repertoire chosen, and would not be as evident in an examination of judgements of – say – jazz performances. Similarly, criteria like *lightness*, *richness* of sound or *affective power* might not be as relevant in the critique of other pieces within the classical repertoire, characterised by a more percussive kind of writing, for example, Bartok piano sonata Sz. 80.

In line with the aforementioned theories in philosophy of art, and with results from studies on interrater consistency in performance evaluation (Kinney, 2009; Thompson, et al., 1998), the proposed model emphasises the utility of developing repertoire-tailored assessment criteria, while suggesting the possibility of relying – at least within the boundaries of a given repertoire and cultural background – on a few higher-order, inherently positive qualities that have more general applicability and intersubjective validity. The present account of evaluation criteria thus provides reference material to be used in investigations of other review corpuses, which will clarify discrepancies and commonalities in the evaluation of different repertoires and musical genres, and in different cultural contexts.

Success value

The seven criteria of evaluation in the present model emphasise a further differentiation. The three major criteria of **intensity**, **complexity** and **coherence**

identify basic properties that are relevant for the aesthetic value of the performance, partially reflecting Beardsley's trinary theory of aesthetic value. The additional criterion **suitability**, and the tension between criteria witnessed in the tightrope kind of statements and visually evidenced through the circular shape of the model, account for the context dependency requirement of those criteria, thus rejecting Beardsley's generalism, and offering empirical support to Carroll's and Sibley's theories.

In addition, three more evaluation criteria were found – **endeavour**, **sureness**, and **comprehension** – that assess the preparation and delivery of the performance beyond what is immediately perceivable through listening, appreciating the performance as product of the performer's achievement. This finding clarifies the scope of the *Performer Qualities* themes identified in Chapter 6. These qualities emerge as far-reaching and central to performance evaluation, supporting Carroll's hypothesis of the relevance of 'success value'– that is, the value we attribute to a work of art as result of us perceiving the work as the outcome of someone's achievement – for music performance appreciation.

The importance given in critical review to these achievement-related criteria raises a question concerning the extent to which the aesthetic value of a performance can be assessed in isolation. This resonates with a further debate in philosophy of art, between empiricists and contextualists, the former defending the possibility of evaluating a work of art properly relying solely on what can be perceived through the experience of the work; the latter affirming the necessity of integrating perceptual information with a series of thoughts and beliefs necessarily bounded with information or assumptions that go beyond the direct experience of the work (Beardsley, 1988; Currie, 1989; Davies, 2006; Graham, 2006).

The present research supports the contextualists view, highlighting how considerations linked to assumptions on the perceived performer behind the performance not only inform listeners' experience of the work but enter the final assessment in a substantial way, tightly bound with other aesthetic related criteria.

Performance evaluation criteria

Finally, the present findings are relevant for the academic context. The list of value adding properties developed from the single analyses of **Primary** and **Supervenient**

Descriptors (Table 7.3 and Table 7.4) and the resulting model of evaluation criteria (Figure 7.2) provide musicians and music students with evidence of what aspects of the musical sound and of the performance as a whole expert critics focus on in their assessments. This model, specifically tailored on the assessment of Beethoven's piano sonatas, offers a practical tool in the preparation and evaluation of these music pieces, which form essential part of each pianist's standard repertoire.

A comparison of the emergent model with McPherson and Schubert's (2004) list of musical parameters commonly used in performance assessment shows a partial overlapping between the criteria applied in critical review and those used in music schools. **Suitability** and **coherence** are the evaluation criteria more prominently represented in McPherson and Schubert's (2004) list, while **intensity** and **complexity** are reflected only in minor proportion in the scheme. This suggests a larger weight given to these criteria in the assessment of professional performances, possibly linked to a stronger focus on craftsmanship versus artistic value in the academic assessment.

Concerning the achievement-related criteria, elements of **comprehension**, **sureness** and **endeavour** find partial correspondence in the parameters 'confidence', 'accuracy', 'physical control' and 'understanding of style/overall structure/emotional character', spread among the evaluation areas of communication, interpretation, expression and technique. The importance these areas of evaluation are given in critical review judgements poses a question concerning how and to what extent these achievement-related criteria are or ought to be integrated in the assessment protocols. On this line, future studies may investigate the utility – in terms of assessment consistency and reliability as well as perception of fairness and validity by evaluators and musicians – of grouping these elements under a common label, thus emphasising the notion of personal achievement in performance.

A final reflection concerns the interdependency of the different criteria within the overall evaluation. The two major questions surrounding the use of segmented assessment schemes in music education institutions as well as in research relate to the nature of criteria to be used and the relative weight that these criteria should have (Mills, 1991). These schemes work under the assumption that the different properties are independent from one another so that it is ideally possible to achieve the highest mark in each of them. The results of the present investigation, however, show that

critics discuss the different value adding properties as interdependent, suggesting that an increase in any single property influences one or more other features. They account for this tension within their assessments, bestowing certain combinations of evaluation criteria with an added value, over and above the value of the single properties.

The interdependency of evaluation criteria suggested by critics' judgements of performance poses then a further challenge to future developments of segmented schemes. Further investigations will be needed to systematically examine relationships between criteria and the role this tensions play in expert listeners' conceptualization of the performance.

CONCLUSIONS

This chapter reported methods and findings of the second thematic analysis run on critics' judgements of performance. It explored the relationship between valence of judgements and descriptors used to characterise the performance, identifying the reasons critics adduce for supporting their value judgements.

Two sets of value adding qualities were developed, for **Primary** and **Supervenient Descriptors** separately. These offer concrete suggestions to musicians and music pedagogues in regard to what properties of the performance and the musical sound critics appreciate and wish for in reviewing. These qualities were summarised into a novel model of performance evaluation in critical review that identifies seven inherently positive higher-order properties of performance praised by critics. Linear relationships between descriptors could not be systematically examined due to the density of the texts nonetheless a tension between criteria emerged as the characterising element of critics' judgements. This emergent model provides empirical data that support and widen current theories on art criticism and aesthetic appreciation and is of direct interest to musicians and music educators. It also offered further evidence of the kind of multi-layered investigations to which the critical review material is open.

The main research question of this thesis, stated in Chapter 1, was: 'What reasons do critics adduce to support their evaluative judgements of recorded performance?' This and the preceding chapter together answered the question in respect to performance-related judgements. The answer can be summarised in terms

of aesthetic-related properties (**intensity**, **complexity**, **coherence**, and their sub-criteria), performer's achievement-related properties (**endeavour**, **sureness**, and **comprehension**, and their sub-criteria), the **suitability** of all of these properties to the musical and cultural context of the performance, and the ability to balance the performance between these different poles (tightrope).

The next and final step of this research completes the answer to the question by embracing a wider perspective and examining what other aspects of a recording – beside the performance – critics discuss in their reviews, what evaluation criteria they apply, and how these build together with the performance criteria to form a final, global judgement of the recorded performance. The methods and findings of this last step of thematic analysis of reviews are the object of Chapter 8.

8 BEYOND PERFORMANCE: REVIEWING RECORDINGS

Chapters 6 and 7 reported methods and findings of the first two layers of in-depth analysis of critical review, focused on the nature of critics' judgements in relation to performance. Findings of these analyses together led to a comprehensive view of the aspects of performance critics discuss in reviewing and the evaluation criteria they apply.

The present chapter completes the investigation, examining the content of critical review beyond performance and clarifying the extra-performance features critics write about and how these diverse features contribute to the final judgement of the end-product recording. In so doing, it embraces a wider perspective and extends the previously developed models to offer a map of critical review content that accounts for extra-performance elements typical of recorded performances.

METHOD

Material

The same corpus of reviews used in Chapters 6 and 7 (N = 100) was also the object of this analysis. For the previous analyses, portions of text were selected that discussed the performance, following the thick-grained categorization of text developed in Chapter 4. The analysis reported in the present chapter examined the residual – i.e., extra-performance – part of the text.

Thematic analysis

The same protocol used in Chapter 6 was applied for this analysis. First, a codebook was developed based on a double-coder examination of a selection of 10 reviews (1 review for each one of the 10 reviewers, see Appendix 9). Then the author applied the codebook to the whole corpus of reviews. Finally, lists of quotes for each theme were compared in an iterative process, to refine the model by clarifying distinctions and relationships between themes. The section of review texts analysed at this stage was less dense and varied than the performance related text. This allowed for the exploration of patterns between themes through a systematic analysis of code co-

occurrences done upon completion of the coding stage. The software Atlas.ti 6.1 was used for the whole analysis, a sample of coded material is reported in Appendix 10.

Relationship between Performance and other Recording Elements

The inductive thematic analysis led to the development of a visual descriptive model that shows what elements of recording – beside performance – critics discuss. Upon completion of this analysis, a final examination was conducted to elucidate how the present findings relate to the findings on performance judgements reported in Chapters 6 and 7. The aim of this final examination was to clarify how considerations on the different components of the end-product recording enter the final, global evaluation.

For this analysis, each review was taken as a text unit. Using the full-coded text produced in the previous analyses, the author examined the narrative of each review, answering the following two questions:

- What elements of the end-product recording (performance and extra-performance elements) are discussed?
- How are the different elements used to form the final, composite judgement of the end-product recording?

RESULTS

Review excerpts discussing issues other than the performance were less dense and less varied than the review text analysed in Chapter 6. The 100 reviews resulted in a total of 2,421 codes, with an average density of 3.03 codes per clause, compared with the 6.66 codes per clause in the performance-related text.

Upon completion of the analysis there were two families of themes: **Recording Elements** and **Critical Activities**, entailing in total 11 dominant themes with a further 11 sub-themes. Seven dominant themes were reflected in the writings of each of the ten critics; the residual four themes – ***Composer***, ***Instrument***, ***Supplementaries*** and ***Price*** – were only found in some critics' writings (see Table 8.1). Consistency in the relative use of the different themes was high (Cronbach's Alpha $\alpha = .962$). Figure 8.1 shows mean frequency with which each theme occurred in each review, for each critic separately.

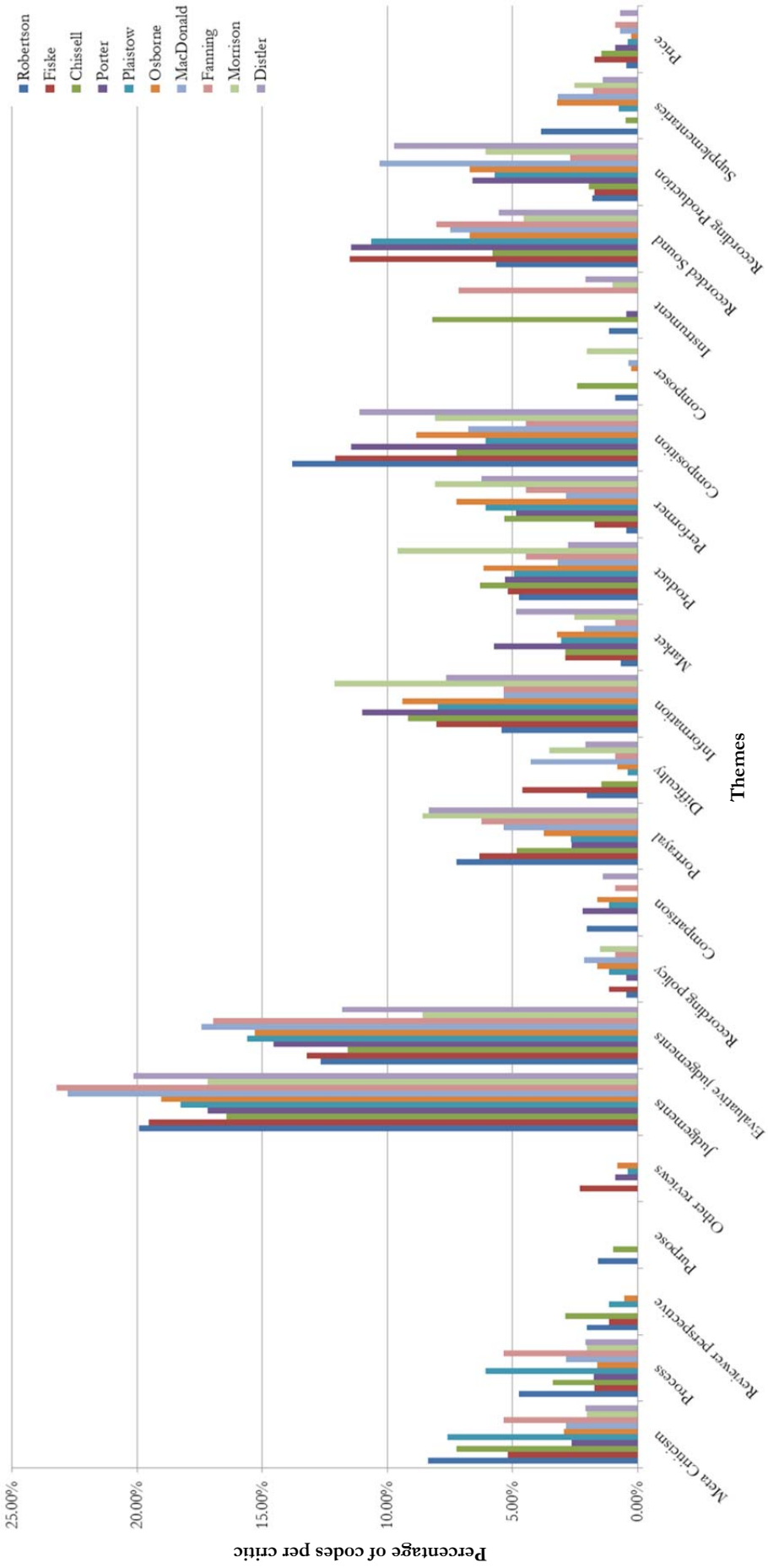


Figure 8.1. Distribution of codes across dominant and sub-themes for each critic. For each critic, the relative frequency is shown with which each theme was coded in the text.

The superordinate theme family **Recording Elements** entails eight dominant themes, which identify elements of the recordings that critics discuss. These are ***Composition***, ***Recorded Sound***, ***Recording Production***, ***Performer***, ***Supplementaries***, ***Instrument***, ***Composer*** and ***Price***. The superordinate theme family **Critical Activities** encompasses three dominant themes (***Information***, ***Judgement***, ***Meta Criticism***) and eleven sub-themes reflecting different kinds of comment that are done in relation to any, one or more of the **Recording Elements**. Figure 8.2 visualises the emergent descriptive model, with the **Recording Elements** located at the bottom, the **Critical Activities** at the top of the figure. Figure 8.3 to Figure 8.5 visualise the frequency with which different **Critical Activities** (***Information***, ***Judgement*** and ***Meta Criticism***) co-occurred with **Recording Elements**.

In the following sections a description is provided for each theme – with dominant themes in bold italic and sub-themes in italic – together with theme definitions from the codebook and examples from the text. Issue, page, and critic's name are given for each example and indentation is used to further clarify hierarchy between themes. Numbers in parentheses after theme names indicate how many times the theme was coded in the critical text. The presentation of results is organized in two parts. First, the different **Recording Elements** are briefly presented. Then, the three dominant themes ***Information***, ***Judgement*** and ***Meta Criticism*** and their sub-themes are described, and examples of their interactions (co-occurrences) with each recording element are given.

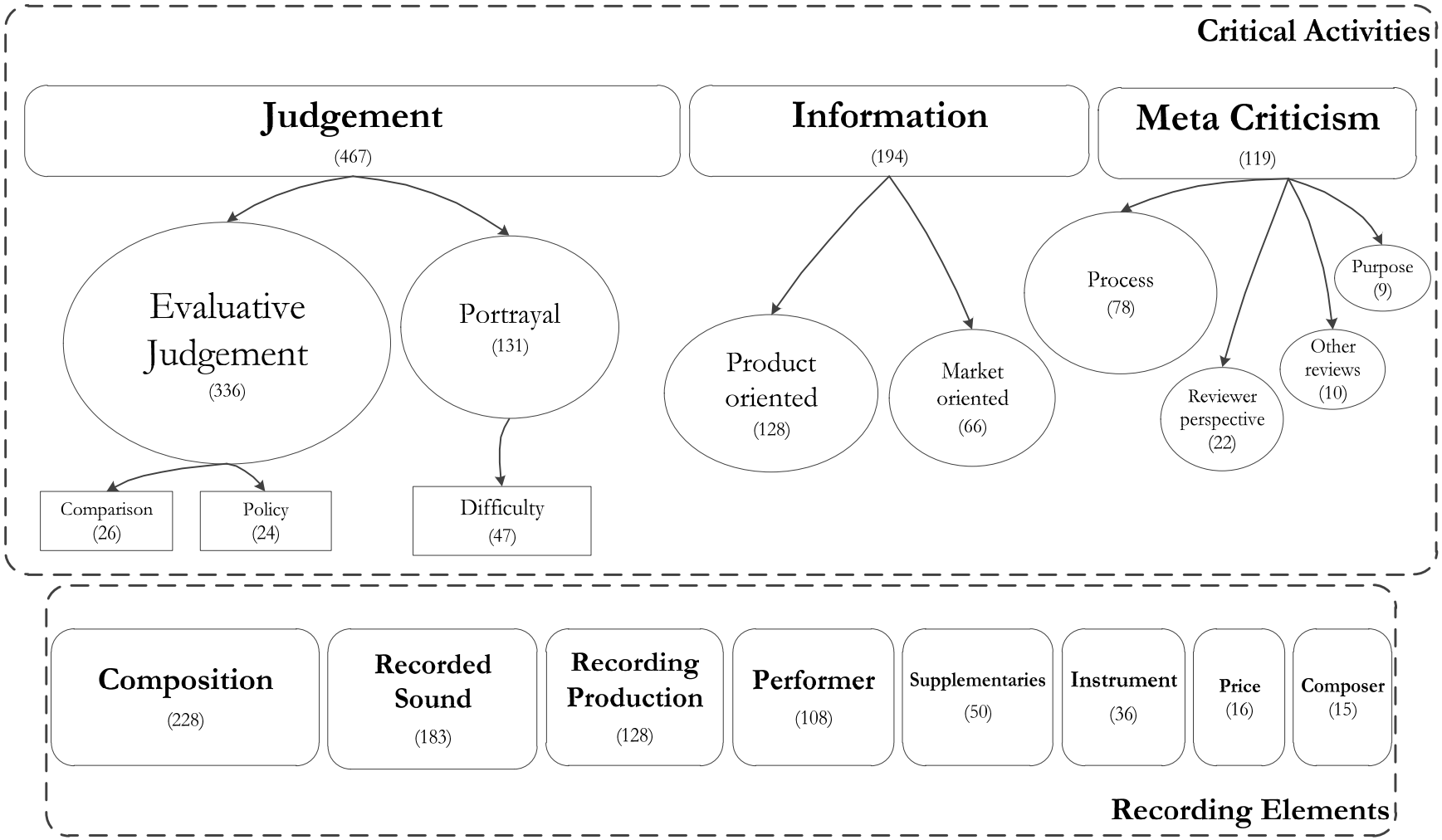


Figure 8.2. Extra-Performance related themes discussed by critics. Different **Critical Activities** are located in the top panel; **Recording Elements** in the bottom panel of the model. Themes are visualised hierarchically moving from rounded rectangles, leading to oval, and when necessary down to square shapes. Arrows reinforce the visualization of this hierarchical structure. Shape size roughly suggests the relative weight of themes, in terms of frequency of occurrence. In parentheses under each theme name is the number of times the theme was coded in the texts. Each time a sub-theme was coded, the relevant higher-order themes were coded as well.

Table 8.1. Distribution of dominant (italic bold) and sub-themes (italic) across the 100 reviews and for each critic separately (10 reviews/critic).

	Theme	*All reviews N=100*	*Robertson (n=10)*	*Fiske (n=10)*	*Chissell (n=10)*	*Porter (n=10)*	*Plaistow (n=10)*	*Osborne (n=10)*	*MacDonald (n=10)*	*Fanning (n=10)*	*Morrison (n=10)*	*Distler (n=10)*
Critical Activities	***Judgement***	**99**	**10**	**10**	**10**	**10**	**9**	**10**	**10**	**10**	**10**	**10**
	Evaluative judg.	91	10	9	10	9	9	10	10	9	9	6
	Recording policy	20	2	2	0	1	3	5	4	1	2	0
	Comparison	13	2	0	0	2	2	5	0	1	0	1
	Portrayal	65	8	8	7	3	5	6	8	5	8	7
	Difficulty	32	6	6	2	0	1	2	7	1	5	2
	Information	**73**	**6**	**6**	**8**	**8**	**9**	**8**	**6**	**5**	**10**	**7**
	Product	54	6	5	6	6	6	7	4	4	8	2
	Market	44	3	4	3	5	7	5	4	1	5	7
	Meta Criticism	**57**	**10**	**4**	**7**	**5**	**10**	**8**	**4**	**5**	**3**	**1**
	Process	43	8	2	4	3	8	5	4	5	3	1
	Perspective	17	7	2	4	0	3	1	0	0	0	0
	Reviews	9	0	3	0	2	1	3	0	0	0	0
	Purpose	4	2	0	2	0	0	0	0	0	0	0
Recording Elements	***Recorded Sound***	**87**	**9**	**9**	**9**	**9**	**10**	**10**	**9**	**7**	**8**	**7**
	Composition	**80**	**10**	**8**	**8**	**7**	**9**	**8**	**9**	**4**	**7**	**10**
	Rec. Production	**60**	**4**	**3**	**4**	**7**	**8**	**9**	**10**	**2**	**7**	**6**
	Performer	**55**	**2**	**3**	**6**	**5**	**8**	**7**	**5**	**5**	**7**	**7**
	Supplementaries	**24**	**5**	**0**	**1**	**0**	**2**	**4**	**4**	**2**	**4**	**2**
	Instrument	**16**	**2**	**0**	**4**	**1**	**0**	**0**	**0**	**5**	**1**	**3**
	Price	**13**	**1**	**2**	**3**	**2**	**1**	**1**	**1**	**1**	**0**	**1**
	Composer	**10**	**2**	**0**	**3**	**0**	**0**	**1**	**1**	**0**	**3**	**0**

Note. Themes are treated as dichotomous variable: for each review, a theme was given the value 1 if it occurred at least once in the text, a value of 0 if it did not occur in the text.

Recording Elements

These themes identify different components of the end-product recording – besides the performance – discussed in critical review.

Composition (228): This first dominant theme identifies the most frequently discussed element of the recording after performance: the work being performed (found in 80 out of 100 reviews). Here are comments that describe, analyse or contextualise the work performed, as well as considerations on the repertoire included in the recording.

> "…the whimsical Menuetto of Op. 22, where the music's skittish dancing measures are interrupted by alarm bells" (Morrison, June 2006, p. 71)
>
> "Opus 111 is coupled here (SBT1188) with two of the Op 2 sonatas" (Osborne, November 2000, p. 86)

Recorded Sound (183): The second largest dominant theme within the group of recording elements – and the one most widely spread among reviews (87 out of 100 reviews) entails comments on timbral and textural qualities of the sound of the recording, including comments on room acoustics and non-musical sounds that are captured in the final result (like sighs, applauses, or environmental sounds). Here are also comments in which the reviewer questions the influence of different elements (like the instrument or performance) on the final sound result.

> "There seems to be a trace of distortion in some of the loud bits on this new record" (Fiske, April 1959, p. 64)
>
> "At a number of moments in the mono the balance of tone struck me as quite uncharacteristic of Annie Fischer: …the sound is all middle and bass and the treble tinkles away to itself almost as if it had nothing to do with the rest of the texture. Comparison with the stereo quickly confirmed that the unnatural perspective is not due to misjudgement on the part of the pianist" (Plaistow, June 1963, p. 36)

Recording Production (128): A third element discussed by critics is the process that led to the final recorded product and the context in which it occurred. Here are comments concerning the technology used, process applied, and the way the recorded music is organized within the disc(s) (often referred to in reviews as 'spacing'). This theme also includes comments on the venue and occasion in which

the recording was realized and the people and institutions (such as labels and engineers) involved in the production.

> "…a live 1980 recital, now part of a two-disc set from Pyramid" (Fanning, April 1992, p. 111)
>
> "…disc two's more plausible running order makes better sense all around (Op. 49 No. 2, the two Op. 14 sonatas, topped off by the valedictory Op. 111" (Distler, April 2007, p. 82)

Performer (108): These are comments on the agent of the performance. They include biographical information, comments on performer qualities and attitudes, as well as comments on where the performer stays in his/her process of recording the Beethoven's cycle (these latter statements were coded under both ***Performer*** and ***Composition***).

> "…the young Swiss-American pianist, Orazio Frugoni" (Porter, June 54, p. 42)

Supplementaries (50): This theme collects comments on the presence and content of accompanying elements like booklet or disc notes.

> "The new set comes complete with notes on Gulda, on the sonatas in general, and on each sonata individually" (MacDonald, January 1970, p. 56)

Instrument (36): This includes comments on the instrument used. However, comments that generally discuss the 'piano quality' were coded under ***Recorded Sound*** unless the context clarified that the statement was about the specific instrument used.

> "The tone-quality of André Watts's instrument, a Yamaha, is mellow to the point of pluminess" (Fanning, September 1988, p. 80)

Price (16): Here are comments on the commercial cost of the recording.

> "For some people it will simply be a question of 20s. against 36s." (Robertson, November 1936, p. 17)

Composer (15): Finally, this last dominant theme within the family **Recording Elements** entails comments on the author of the work performed.

> ""I am not satisfied with my works to date, and from now on I want to take a new path", so Beethoven allegedly confessed not long before embarking on the three sonatas of Op. 31" (Chissell, June 1992, p. 66)

Critical Activities

Three main kinds of activities were found in the review text, done in relation to any, one, or several of the **Recording Elements** described above: offering factual ***Information***, making a ***Judgement***, or reflecting on criticism metacognitively (***Meta Criticism***). In this section these three dominant themes and their sub-themes are presented, together with details on how they relate to the different **Recording Elements**.

Information (194): This dominant theme entails comments on facts (historical or current) related to the **Recording Elements** that are purely descriptive. They describe, contextualise or analyse the **Recording Elements** in a factual way. Figure 8.3 visualises the portion of the model relevant to the ***Information*** theme, together with the links to the different **Recording Elements** involved.

Within the larger theme of ***Information***, two sub-themes emerged that are *Product oriented* or *Market oriented*.

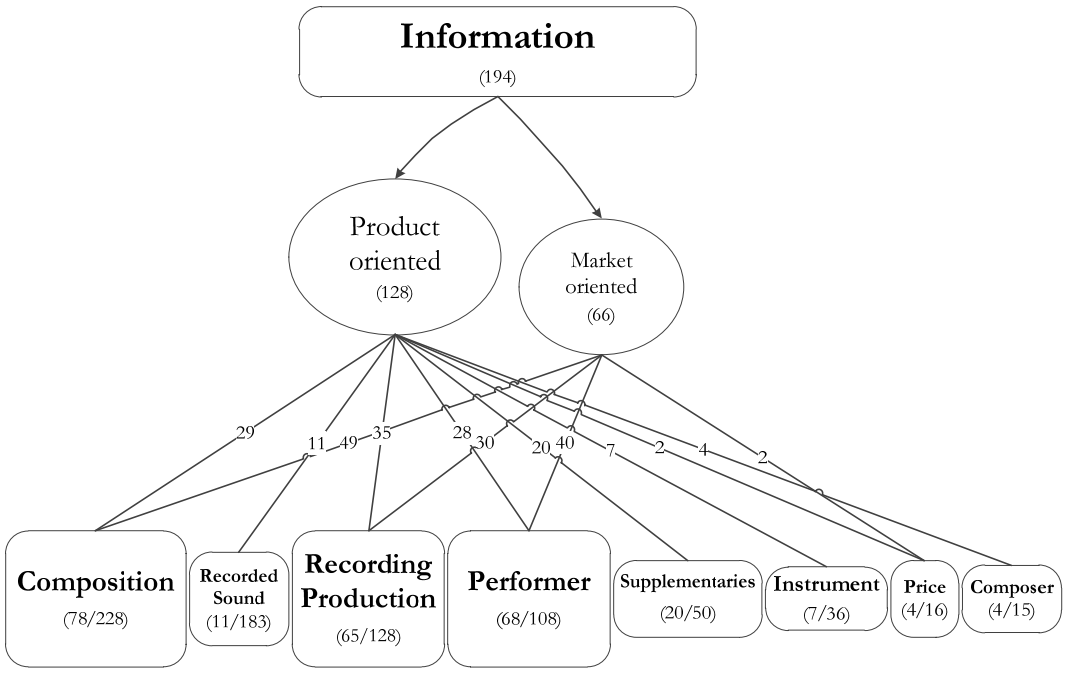


Figure 8.3. Visualisation of co-occurrences between ***Information*** statements and **Recording Elements**. Arrows indicate hierarchical relationships between themes. Straight lines visualise co-occurrences. Numbers in parentheses after the names of each recording element show how many times that element co-occurred with ***Information*** (first number) versus how many times it was coded in total within the critical text (second number). Shape size roughly indicates the number of times the element co-occurred with ***Information***.

Product oriented (128): This sub-theme entails ***Information*** that describes the actual product being reviewed.

Within the 100 reviews, *Product oriented* ***Information*** was found that concerns all of the eight different ***Recording Elements***. Most often this kind of information focuses on the ***Recording Production*** (n = 35), ***Composition*** (n = 29) ***Performer*** (n = 28), or ***Supplementaries*** (n = 20).

Concerning ***Recording Production***, a large portion of statements (n =14) focus on how recorded sonatas are distributed within and between discs ('spacing'); while an equally large group of statements gives information on the context of the record production:

> "Vox presents "Moonlight" and "Pathétique" together on one side, and "Appassionata" on the other" (Porter, February 1955, p. 46)
>
> "Craig Sheppard's Beethoven piano sonata cycle was given in Seattle's Meany Theatre during 2003-2004" (Morrison, June 2006, p. 71)

Information on ***Composition*** mostly focuses on the structure and compositional context of the piece (n = 16) or on the repertoire entailed in the reviewed disc(s) (n = 6).

> "The place of a slow movement is taken by one labelled Scherzo but actually is what is known as "first-movement" form" (Robertson, February 1948, p. 23)
>
> "The second disc couples the Waldstein, the Appassionata, and Les Adieux" (Osborne, August 1986, p. 49)

Concerning the ***Performer*** critics offer mainly biographical information (n = 21) and a few considerations on his/her reception and fame.

> "Eduardo del Pueyo is the Spanish pianist now living in Belgium" (Fiske, November 1959, p. 68)
>
> "His appearances at New York's Birdland club where he was known as DeadEye Fred were as legendary as his concerts at Carnegie Hall" (Morrison, December 2002, p. 72)

Product oriented ***Information*** concerning ***Supplementaries*** mostly focuses on the content of the booklet.

> "The latest disc in the series comes with a booklet essay by William Kinderman … entitled "Intimacy and Pastoralism"" (Osborne, February 1996, p. 75)

Finally, *Product oriented* ***Information*** was linked to ***Recorded Sound*** (n =11), ***Instrument*** (n = 7), ***Composer*** (n = 4) and ***Price*** (n = 2):

> "Op. 111 is recorded at a higher dynamic level (one needs to turn down the volume control)" (Porter, October 1954, p. 51)
>
> "Binns plays the Pathétique on a Stein of c.1802" (Chissell, February 1983, p. 52)
>
> "…some scholars argue that Beethoven's not too good Italian was responsible for his frequent use of the word in the very different French sense of assez" (Chissell, March 1969, p. 66)
>
> "…limited availability of this price with a deadline at the end of February." (MacDonald, January 1970, p. 56)

Market oriented (66): Differently from *Product oriented*, *Market oriented* ***Information*** sets the reviewed recording in the context of the wider music market (for instance, commenting on what other recordings of the same piece(s) are available – in the same or different coupling/format, or what recordings by the same pianist have been produced).

Market oriented ***Information*** is mostly linked to ***Composition*** (n = 49), ***Performer*** (n = 40) and ***Recording Production*** (n = 30).

> "The LP Beethoven Sonata repertory is not extended by this record" (MacDonald, August 1954, p. 39)
>
> "Meanwhile competition comes from Backhaus" (Porter, October 1954, p. 50)
>
> "Since issuing the original review copies (none of which reached the shops) DG have re-mastered the Compact Disc" (Osborne, May 1983, p. 49)

Often the three elements are mentioned in combination:

> "Op. 101 also features on a five-disc set of Arrau recordings of Beethoven sonatas and concertos made for EMI in the 1950s" (Osborne, March 1993, p. 73)

In two occurrences *Market oriented* ***Information*** focused on the ***Price***.

> "Turnabout reissued the same two sonatas in July for only a pound from Brendel" (Chissell, December 1970, p. 86)

Judgement (467): This is the largest, most complex and most widely spread dominant theme emerged in the analysis, found in 99 out of 100 reviews. It encompasses comments that express the reviewer's opinion or subjective conclusion on the recording and its elements. It entails two large sub-themes, *Evaluative judgement* and *Portrayal*, and three further minor sub-themes, *Difficulty*, *Comparison*, and *Policy*. Each sub-theme is linked to at least four recording elements. The resulting descriptive model is visualised in Figure 8.4.

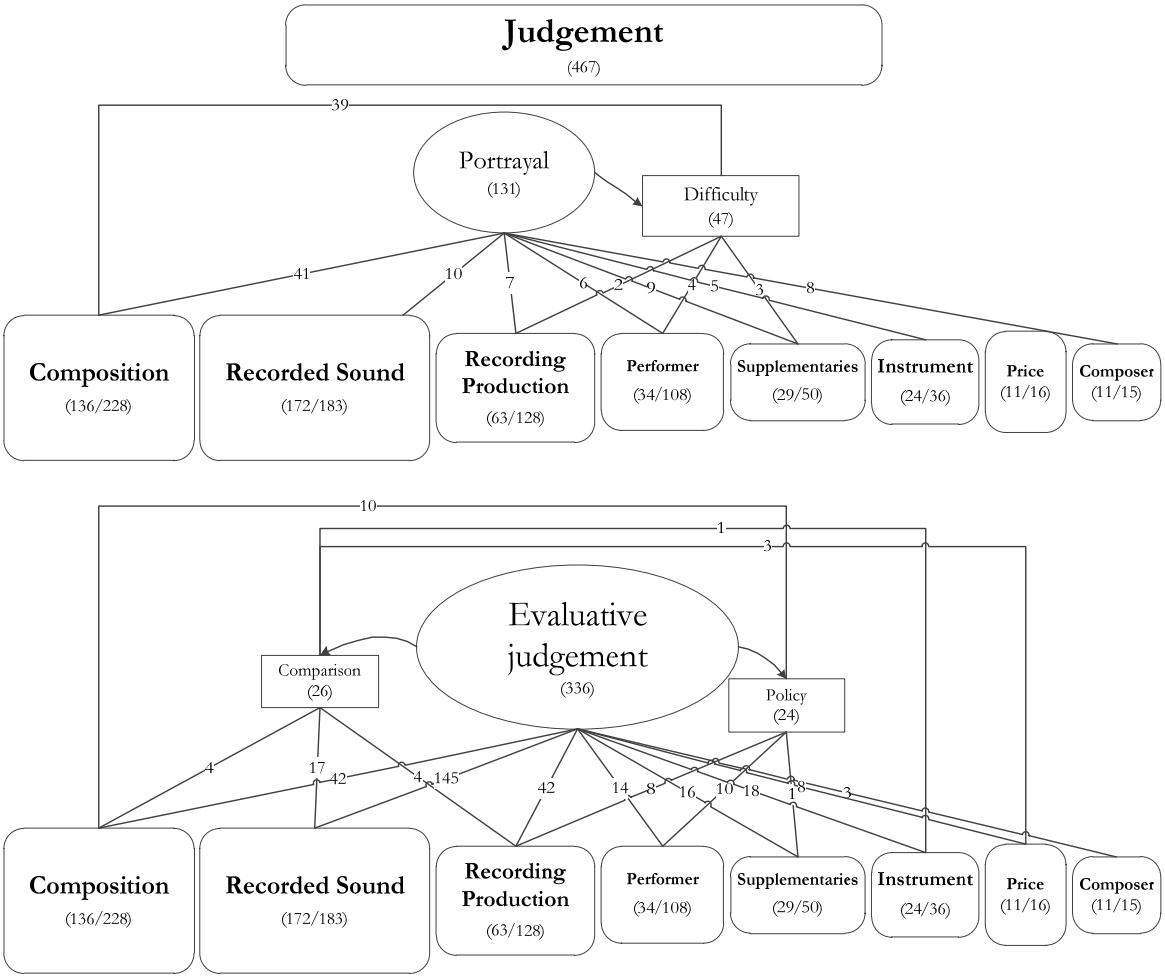


Figure 8.4. Visualisation of co-occurrences between ***Judgement*** statements and **Recording Elements**. Arrows indicate hierarchical relationships between themes. Straight lines visualise co-occurrences. Numbers in parentheses after the names of each recording element show how many times that element co-occurred with ***Judgement*** (first number) versus how many times it was coded in total within the critical text (second number). Shape size roughly indicates the number of times the element was linked to ***Judgement***. Given the large amount of co-occurrences to be visualised, the figure has been divided in two sub-panels, showing co-occurrences between **Recording Elements** and the ***Judgement*** sub-themes *Portrayal* (top panel) and *Evaluative judgements* (bottom panel) separately.

Evaluative judgement (336): The largest sub-theme to emerge from the analysis, present in 91 out of 100 reviews, encompasses opinions or conclusions about the worth, merit, importance or usefulness of the **Recording Elements**. It includes comments that point out the relative importance of different elements in the final assessment. These will be discussed in more detail in the final results section on composite judgements (p.270).

> *Comparison:* In a few cases (7.74%) the *Evaluative judgement* sets the reviewed recording and its elements against other produced recordings. These cases are grouped in this sub-theme.

Evaluative judgements – either done in isolation or in form of *Comparison* – are at times (n = 28) offered as holistic assessments of the end-product recording[25], as in:

> "Don't pass this amazing release!" (Distler, December 2008, p. 103)

Holistic evaluations also point at times at different kinds of value that a recording can possess, like historical or pedagogical value, related to the tangible, semi-permanent nature of the recording-object (collectability, n = 11):

> "All in all this well-recorded album is musical history at its liveliest and best" (Chissell, October 1980, p 71)
>
> "I pity any student who hopefully buys this recording as a model to copy. There are pianists like Schnabel and Lamond whom we feel are teachers as well as artists and can, therefore, inspire us to greater efforts, their technique not being so impossibly above our own. Ernest Newman once said 'Wagner, glorious companion can never inspire' and in the interpretative field that is true of Edwin Fischer" (Robertson, October 1935, p. 18)
>
> "But anyone who wants Beethoven's three last sonatas in his collection – and who could not? – is warmly directed to this new Vox" (Porter, June 1957, p. 19)
>
> "Those who collect multiple versions of Beethoven cycles will find much food for thought in the offer" (Distler, April 2007, p. 82)
>
> "The "Appassionata" is the sort of fairly good and fairly interesting performance that one would be quite content to hear at a recital, but which is probably not to be bought and lived with" (Porter, May 1956, p. 49)

Holistic evaluations, however, represent a minor part of *Evaluative judgement*. Most times these judgements address specific **Recording Elements**, and links were found

[25] Holistic judgements of the end-product recording are presumptively judgements of the performance as well. Since, however, these statements are not specifically focused on performance, it was decided to discuss them in this chapter.

between *Evaluative judgement* and all of the eight elements. Analysis of these interactions revealed a few recurrent features discussed as value adding qualities in the evaluation of the different **Recording Elements**. These value adding qualities are summarised in Table 8.2.

Among the eight different elements, the by far largest number of evaluations concern the ***Recorded Sound*** (n = 162, including *Comparison*). In fact, critics mentioned the *Recorded Sound* almost exclusively (93.99%) to offer an *Evaluative judgement* of it. In several cases (n = 63) the evaluation given is a pure one, devoid of descriptive content:

> "Recording quality is fine" (Fanning, April 1992, p. 111)

When the evaluation is supported by a characterization of the sound three main qualities are praised (or wished for) by critics: depth, richness and warmth of tone (n = 41); faithfulness (n = 25) and clarity (n =15).

> "…the recording, which, though serviceable enough, is not really among 1969's best. The sound lacks depth and richness" (MacDonald, January 1970, p. 56)
>
> "Decca warmly detailed engineering accurately captures Paik's sound" (Distler, October 2005, p. 81)
>
> "And here a word needs to be said about the recordings, which are thrillingly loyal to the music-making" (Osborne, February 1996, p. 75)
>
> "…it is impossible not to admire the sound, its clarity and dynamic range, its marbled splendour" (Osborne, November 1955, p. 146)

Another aspect of the recorded sound that critics often discuss is the presence of extraneous noises (n = 27). Surface noise, pre-echo and resonance, whirr of the hammers, pedal sounds, or noises that betray the cut between takes are usually perceived as disturbing factors.

> "The recording is tonally much better, but the surface is very spluttery" (Porter, June 1954, p. 42)
>
> "Pre-echo (heard in the repeat as well) spoils the silence twelve bars after the change to four flats in the first movement" (Porter, October 1954, p. 51)

Conversely, breathing and gasps by the performer or audience can be perceived as either positive or negative additions to the performance:

> "The recordings are excellent, with less fingersound and intrusive breathing than was sometimes the case with Arrau on record in the digital age" (Osborne, March 1993, p. 73)
>
> "Sometimes I thought I could just hear the artist's involved breathing – which seems to bring him into your room without it being unduly distracting" (Chissell, March 1972, p. 74)

Two more large groups of *Evaluative judgement* concern the ***Composition*** and the ***Recording Production*** (n = 46 each, including *Comparison*).

The ***Composition*** is evaluated mostly for its artistic value (n = 33):

> "…the greatest piano music in existence" (Fiske, November 1959, p. 68)
>
> "…wondrous pieces" (Plaistow, January 2002, p. 81)
>
> "…the loftiest yet at the same time physically beautiful and physically exciting music known to man" (Osborne, November 2000, p. 86)

At times, however, what is praised is the repertoire, in terms of combination of pieces that a recording offers (n = 11):

> "Wonderful value, Beethoven's last three sonatas on a single disc" (Porter, June 1957, p. 19)
>
> "ACM2015 will appeal, for with the Pathétique, the Moonlight, and the Appassionata it offers three of the most popular" (MacDonald, May 1981, p. 92)

Evaluations of the ***Recording Production*** relate mostly (n = 35) to the way pieces are distributed within the disc(s). Spacing strategies are praised that avoid unnecessary breaks, order sonatas logically and aesthetically, and make effective use of space.

> "Where so much comment each month needs to go on disc-spacing that is either unhappy or (more often than you would think possible) simply idiotic, it is a very great pleasure to be able to point for once to spacing that is absolutely first class" (MacDonald, January 1965, p. 59)
>
> "It is refreshing to see someone coupling sonata performances with the kind of care with which a gallery director would juxtapose his paintings" (Osborne, April 1982, p. 66)
>
> "The turn-over in the middle of the Moonlight is unfortunate" (Chissell, February 1983, p. 52)
>
> "Sometimes it's hard to fathom the programming logic. For instance, why follow the mighty Hammerklavier with the much slighter F major, Op. 54?" (Distler, April 2007, p. 92)

A few more comments concern the quality of engineering work like remastering and transfer, length of silence between movements or (in the CD era) the localization and number of access tracks (n = 7).

> "…superbly remastered by Mark Obert-Thorn" (Morrison, January 2005, p. 76)
>
> "…I would have liked longer scrolls between the movements – a small but by no means an unimportant point" (Plaistow, March 1964, p. 63)
>
> "I also have not-so-fond memories of Amadeo's own CD edition, remastered with fake reverberation and only one access track per sonata" (Distler, September 2006, p. 80)

Beside ***Recorded Sound***, ***Composition***, and ***Recording Production***, a few *Evaluative judgements* were found linked to the other **Recording Elements**.

Evaluative judgements of the ***Instrument*** (n = 19) are often pure ones, with no descriptive content:

> "The piano tone is good throughout" (Robertson, February 1937, p. 19)

When the judgement characterises the sound, the focus is on its richness (n = 5), the timbral and dynamic variety the instrument accommodates (n = 5), or the ability to sustain the playing without marring clarity of articulation (n = 3):

> "…the instrument itself is unpleasantly tinny in the high treble" (Fanning, November 1992, p. 152)
>
> "Inevitably both, and particularly the Heilmann restrict the music's dynamic range, those dramatic alternations of ff and pp, and so on" (Chissell, October 1980, p. 71)
>
> "…it also tends to impose a uniformity of timbre which gives diminishing returns" (Fanning, September 1988, p. 80)
>
> "…how surprisingly well the instrument sustains their song" (Chissell, October 1980, p. 71)
>
> "…once or twice had me longing for the greater clarity of a modern grand-likewise" (Chissell, June 1992, p. 66)

Evaluative judgements of ***Supplementaries*** (n = 16) most of the times (n = 9) focuses on the booklet or sleeve notes, praising notes that offer an informative and stimulating guidance to the reader, prompting critical reflection and perceptive listening:

> "…this three-disc set is crowned with a scholarly and illuminating essay by Jean-Paul Montagnier" (Morrison, June 2008, p. 81)
>
> "…Hewitt, as usual, provides her own penetrating, vividly articulated annotations" (Distler, June 2007, p. 84)

Comments on the ***Performer*** (n = 14) usually come as pure evaluations:

> "…a rising young pianist – one to watch" (Morrison, March 2003, p. 63)

In three cases, an *Evaluative judgement* concerned the ***Composer***:

> "…one can appreciate the extra mastery with which Beethoven treated the two forms" (Robertson, August 1950, p. 23)

Finally, a few *Evaluative judgements* were found linked to ***Price*** (n = 11). These comments evaluate the recording's value for money when weighed against the quality of other elements.

> "The recording is outstandingly good: it would earn high marks at any price" (Chissell, June 1969, p. 53)

> "It's brilliant, it's a bargain – welcome back Gulda's Beethoven" (Distler, September 2006, p. 80)

The importance of ***Composition***, in terms of what pieces are coupled within the recording, here is charged with a new meaning: what counts is not just what pieces are recorded, but also how many of them there are:

> "…the result is a splendid bargain in minutes per shilling" (Fiske, November 1959, p. 68)
>
> "Excellent value for money, in point of bars per penny" (Porter, February 1955, p. 46)
>
> "On whether two further points are, for him, small ones only the reader can decide: one is the price of the issue – eleven records for less than £8 is a temptation indeed; the other is the limited availability of this price with a deadline at the end of February" (MacDonald, January 1970, p. 56)

Policy: A sub-theme of *Evaluative judgement* entails statements focused on the producers' (label, performers) course of action. Characteristic of these comments is an apparent change of readership target; here reviewers seem to talk to the record producers, rather than evaluating the recording.

Policy statements are often expressed in the form of a desire or hope for a certain product to be produced or made available:

> "Schnabel LP reissues of these and the other sonatas are long overdue" (Porter, October 1954, p. 50)
>
> "I could wish that some of these major pianists, to say nothing of the recording companies, would turn their attention to some of the lesser recorded sonatas" (Plaistow, June 1963, p. 36)
>
> "Indeed I would suggest only one improvement: that the central scroll should in these circumstances be considerably wider." (MacDonald, March 1965, p. 58)
>
> "I hope this is the start of what promises to be a more than distinguished series" (Morrison, March 2003, p. 63)

Or they can express gratitude/disappointment for the existence (or lack thereof) of a certain product.

> "I will not make the obvious remarks about yet another recording of the C sharp minor Sonata" (Robertson, October 1945, p. 16)
>
> "It is a little surprising that Columbia should issue another disc of Beethoven's last two sonatas considering how very good was the Arrau version sponsored by the same firm last February" (Fiske, November 1959, p. 67)
>
> "…the famous 1956 Solomon recording, whose absence from the catalogue is much to be regretted" (Osborne, December 1983, p. 84)
>
> "I am pleased Appian has released the recital" (Osborne, November 2004, p. 79)

Portrayal (131): The second largest sub-theme emerged in the analysis, present in 65 out of 100 reviews, entails descriptive opinions or conclusions that characterise the recording or its elements.

Interactions were found between *Portrayal* and all **Recording Elements** except ***Price***. The strongest link is between *Portrayal* and ***Composition*** with 41 co-occurrences. Comments that portray the musical work being performed focus on characterisations of the music in terms of character, atmosphere, rational qualities, or reception-related qualities (n = 26); structural analysis and interpretation (n = 13); and compositional style, technique and context (n = 8):

> "…the pensive E flat Sonata" (Fanning, September 1990, p. 116)
>
> "…the famous minuet" (Fiske, October 1958, p. 65)
>
> "…this tune is implicit in the rapid first subject of the movement, showing itself as the germ idea of the whole Sonata" (Robertson, August 1934, p. 29)
>
> "…the only one of the three to betray the darkness from which it grew in 1802" (Chissell, June 1992, p. 66)

Links between *Portrayal* and other **Recording Elements** are weaker, featuring between five and ten occurrences each. Ten comments focus on the *Portrayal* of ***Recorded Sound***:

> "Held chords, with the sustaining pedal down, do not fall away in tone" (Robertson, February 1948, 23)

Comments on ***Supplementaries*** (n = 9) describe the content of sleeve notes:

> "The notes are anonymous, but even in good translation (by Richard Rickett) betray their Germanic origin from time to time" (MacDonald, January 1970, p. 56)

Portrayal of the ***Composer*** (n = 8) focuses on intellectual qualities as well as assumed intentions and thoughts:

> "Like Haydn and Mozart, only more so, he was always dreaming of resources beyond the horizon to carry the intensity of his thought" (Chissell, October 1980, p. 71)
>
> "Beethoven's pioneering and burgeoning spirit" (Morrison, March 2003, p. 63)

Table 8.2. Value adding qualities emerged through the analysis of co-occurrences between *Evaluative judgement* and the eight **Recording Elements**. Number in parentheses after element names show how many times an element co-occurred with an *Evaluative judgement*. Number in parentheses after value adding properties names show how many times the quality was mentioned in relation to the given element.

Recording Elements	*Value adding qualities*	*Explanation*
Holistic evaluations (28)	Collectability (11)	Value of the recording-object linked to its semi-permanent nature. Recording-object valuable for collectors, as pedagogical source or as historical document.
Rec. Sound (166)	Richness (41)	Rich, warm, resonant sound.
	Extra-noises (27)	Avoidance of non-musical noises linked to recording process. Noises coming from the artist may be welcome.
	Faithfulness (25)	Realistic reproduction of the performance sound.
	Clarity (15)	Sharp and defined sound.
Composition (45)	Artistic value (33)	Composition praised as artwork, mostly pure evaluations.
	Programmatic value (11)	Combination of compositions praised for their programmatic logic.
Rec. Production (45)	Distribution (35)	Distribution of pieces within disc(s) that is aesthetically meaningful and avoids unnecessary breaks.
	Technicalities (7)	Sufficient number of access tracks, appropriate silence length between tracks, quality of transfer/remastering.
Instrument (18)	Richness (6)	Rich, warm, resonant sound.
	Variety (5)	Instrument accommodating dynamic and timbre variety.
	Clarity/Sustain (3)	Instrument that sustains melody yet maintaining sharpness and definition of sound.
Supplementaries (16)	Listening guide (9)	Illuminating, informative and stimulating notes.
Performer (13)	Pure evaluations (9)	Valuable performer (no specification)
Price (11)	Value-for-money (11)	Economical value set against quality of other elements, and length of recorded music.
Composer (3)	Pure evaluations (3)	Valuable composer (no specification)

Portrayal of the ***Recording Production*** (n = 7) express opinions and ideas concerning the process of production, its surrounding and technicalities.

> "The two little Op. 49 Sonatas appear to have been recorded in different conditions" (Fiske, August 1963, p. 31)
>
> "I make the assumption that what we're given is an edited montage of long takes from perhaps two public recitals" (Plaistow, October 1989, p. 98)

Comments on the ***Performer*** (n = 6) describe mental states and artistic traits that characterise the performer independently from the specific performance reviewed:

> "It is as though Ogdon's personal experiences have put him beyond reach of some of the deadening pressures of the musical 'scene'" (Fanning, November 1986, p. 78)
>
> "A pianist who, as someone said of Liszt, 'does not just play the piano but tells at it', he was best heard in real time in a real hall" (Osborne, November 2004, p. 79)

Finally, a few comments portray the ***Instrument*** (n = 5), in terms of the qualities of the sound produced:

> "The Fazioli piano's lean bass and bright treble characterise the kind of timbral differentiation one often associates with instruments of Beethoven's time" (Distler, November 2006, p. 97)

Difficulty (47): Within the sub-theme *Portrayal*, one recurrent emergent idea is that of *Difficulty*. Here are opinions or conclusions that reflect challenges or risks with which the people involved in the different stages or the recording production [composer, performer, engineer(s)] had to cope.

The largest interaction found within *Difficulty* is with ***Composition*** (n = 39). Usually these comments point at challenges that the ***Composition*** poses to the performer (n = 37). A few times the nature of the challenge is briefly explained:

> "The 187 bars, without hardly a break, of the Adagio, impose a tremendous strain on the player which can only be supported by iron discipline and control" (Robertson, November 1936, p. 17)

Most times, however, an adjective or short expression indicating *Difficulty* is added *en passant* while mentioning a certain piece or section of a piece to introduce a performance-related judgement:

> "… Miss Donska here succeeds without a doubt, as she does indeed in the whole of that restrained and difficult E major finale" (MacDonald, November 1964, p. 52)

Comments of this kind were found copiously spread across reviews:

> "…the difficult trio section" (Fiske, November 1959, p. 67)
>
> "…an Everest among the 32" (Chissell, March 1972, p. 74)
>
> "…the problematic first movement recitatives" (Morrison, March 2003, p. 63)

Composition however can also present challenges to sound engineers (n = 2):

> "The Hammerklavier, with a long first movement, a short second, a monumental third which cannot be guaranteed to come out at much under twenty minutes, and again a long fourth has always been a difficult sonata to space on record" (MacDonald, March 1965, p. 58)

Beside ***Composition***, *Difficulty* was also linked to ***Performer***. The performer treatment of the music can be challenging for engineers, while physical limitations can represent a test for the performer him/herself (n = 4):

> "It must be said that Philips did have a rather tougher assignment than DGG in the first place: Gulda makes, with one single exception, every repeat suggested by Beethoven; Kempff, for DGG, is more selective" (MacDonald, January 1970, p. 56)
>
> "My first reaction on being swept into the A major Sonata at speed was how marvellous to have such vitality and facility at the age of (almost) eighty-five" (Chissell, March 1969, p. 66)

Finally, a few times *Difficulty* was linked to ***Supplementaries*** (n = 3) and ***Recording Production*** (n = 2). ***Supplementaries*** like booklets and disc notes can prove challenging for the reader when they lack clarity in terms of text or readability:

> "I would not claim that this has no meaning; only that if it has then I cannot myself see it, and I fancy I shall not be alone" (MacDonald, January 1970, p. 56)
>
> "There are good sleeve-notes, not especially easy to read through the superimposed reproduction of the head of the old man himself" (MacDonald, May 1981, p. 92)

The ***Recording Production*** on the other hand can prove demanding for the performer:

> "His playing is of a near flawless clarity and lucidity – a remarkable achievement given the circumstances of this recording" (Morrison, December 2002, p. 72)
>
> "Restraint must dominate much of the late A flat Sonata, too; in this the first movement seemed less rapt than can be (it is a condition of mind difficult to maintain through a recording session)" (MacDonald, November 1964, p. 52)

Difficulty coming from different sources can also add together, into an even tougher challenge:

> "…it is only fair to say that the nature of Beethoven's piano writing and Medtner's unrelenting treatment of it must have presented the engineers with a difficult problem" (Robertson, February 1947, p. 8)
>
> "For a man in his eighty-seventh year, he was remarkable, playing here often difficult music with a fullness of sound and accuracy of touch that makes late Rubinstein seem approximate, the nonagenarian Horszowski merely amateurish" (Osborne, March 1993, p. 73)

Meta Criticism (119): Finally – after the large set of themes related to ***Judgement***, this last dominant theme groups meta-reflections on the process of review writing and on the reviewer-reader interaction. These comments do not describe or express judgements on the reviewed recording, even though they can be related to a judgement or description. Here the reviewer seems to take a step back, to offer considerations relevant to the critique and its understanding, but not constituent of it. Within ***Meta Criticism*** one major sub-theme emerged, *Process*, accompanied by the three minor sub-themes *Reviewer perspective*, *Other reviews*, and *Purpose*. Given their abstract and general nature, these comments – with few exceptions – are not linked to any of the **Recording Elements**. Figure 8.5 shows the part of the model relevant to ***Meta Criticism*** statements with **Recording Elements**.

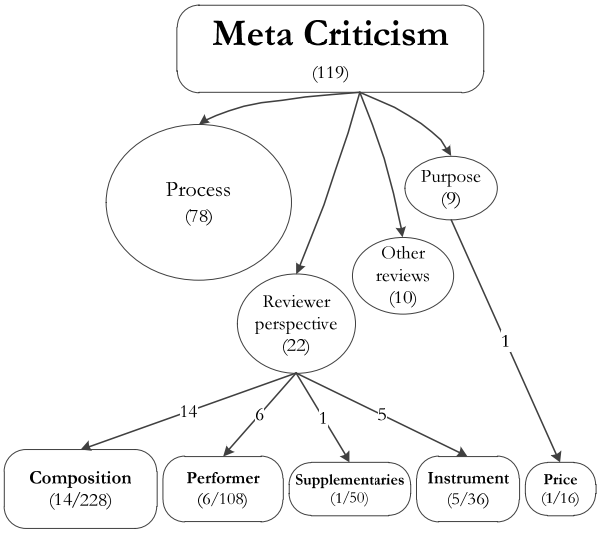


Figure 8.5. Visualisation of co-occurrences between ***Meta Criticism*** statements and **Recording Elements**. Arrows indicate hierarchical relationships between themes. Straight lines visualise co-occurrences. Numbers in parentheses after the names of each recording element show how many times that element co-occurred with ***Meta Criticism*** (first number) versus how many times it was coded in total within the critical text (second number). Shape size roughly indicates the number of times the element co-occurred with ***Meta Criticism***.

Process (78): This sub-theme entails reflections on how the reviewer came to his/her judgement as well as abstract, general comments on how critical review is or should (or should not) be done.

> "Chalking up faults is always easier than doing justice to a musician's inner interpretative vision" (Fanning, April 1992, p. 111)

Two ideas emerging from these comments concern the selective nature of review and the meaningfulness of evaluative, ranking-like comparisons between high level performances:

> "A detailed account of every performance would make dull reading even if space permitted it. So I'll just pick out a few salient points" (Chissell, October 1980, p. 71)
>
> "Only an extended essay could do justice to the fourth and final volume of Paul Lewis's Beethoven sonata cycle. But space, sometimes the critic's friend, here his enemy, forbids much beyond generalisation" (Morrison, June 2008, p. 81)
>
> "It is disagreeable to write about artists of distinction as if they were candidates at a competitive festival" (Robertson, October 1953, p. 22)
>
> "It is not at all my wish to play one master pianist off against another, far from it" (Plaistow, October 1989, p. 98)

This theme also includes comments on the reviewer-reader interaction and reflections that extend beyond the writing stage, to encompass critic and reader actions respectively before writing and after reading the review.

> "It was only recently, when I was listening to selected comparisons for a review of Stephen Kovacevich's fine new account of Beethoven's three Op. 31 Sonatas (EMI, 11/95), that I was alerted to the extraordinary qualities of Alfred Brendel's" (Osborne, February 1996, p. 75)
>
> "One week later, with no students or comparative versions to distract me, I played Uchida's disc again" (Distler, May 2006, p. 90)
>
> "The reader who has followed me this far will be more eager to share my enthusiasm, I trust, than to weigh in the balance a couple of reservations" (Plaistow, August 1979, p. 69)
>
> "…[readers] will need to hear both versions to render their verdict entirely satisfactory to themselves" (Robertson, November 1936, p. 17)

Reviewer perspective (22): Here are reflections on the reviewer's past experiences, preferences, emotions and dispositions towards **Recording Elements** prior to (and not triggered by) experiencing the reviewed product.

These statements are the ones within the dominant theme ***Meta Criticism*** that have most often as their object one of the **Recording Elements**. The largest group of statements (n = 14) concerns the ***Composition***. However, examples were also found

that discuss the reviewer's perspective towards ***Performer*** (n = 6), ***Instrument*** (n = 5), and ***Supplementaries*** (n = 1).

> "It is, to me, the least interesting movement in the sonata: and I have a rooted dislike to its second theme, the one with the Alberti bass" (Robertson, February 1948, p. 23)
>
> "Andor Foldes is a pianist I admire - particularly in Bartok" (Plaistow, December 1961, p. 57)
>
> "If, for my desert island, I had to choose between Beethoven sonatas played on instruments of their own period and the modern piano, I would choose a splendid full-size Steinway without a moment's demur" (Chissell, October 1980, p. 71)
>
> "I remember being astonished to read in Joachim Kaiser's Great Pianists of Our Time (George Allen & Unwin: 1971) that Solomon's name was scarcely known in Germany" (Osborne, November 2000, p. 86)

Other reviews (10): There are a few statements in which the reviewer comments on the existence and content of other reviews, and how this relates to the reviewer's judgement.

> "Denis Matthews on Columbia 33SX1021, reviewed by L.S. last month (a review I am entirely at one with)" (Porter, June 1954, p. 42)

Purpose (9): Finally, a few cases of ***Meta Criticism*** offer insights into the nature and role of critical review, from the perspective of the reviewer.

> "The futility of literary comment on music was brought home to me this month in reading a book by a German author on Beethoven's Piano Sonatas" (Robertson, February 1937, p. 19)

Here a focus emerged on critical review as guidance for readers to decide what product(s) to purchase.

> "Confronted with two issues of this colossal work the reader will wish to have certain material points settled straightaway" (Robertson, November 1936, p. 17)
>
> "Since each performance has some special distinction of its own to appeal to this or that Beethoven addict, a clear-cut recommendation from a reviewer gets increasingly difficult - if not impossible" (Chissell, March 1972, p. 74)

One of these statements also entailed a side-comment on price.

> "To sum up I would say that those who think first of recording should buy the Decca: the rest, if their resources allow them, may gain a little help to make a decision from what I have written" (Robertson, November 1936, p. 17)

Relationship between Performance and other Recording Elements

Following the analysis of extra-performance statements, the narrative of critical text was examined at whole-review level to clarify how comments on **Recording Elements** relate to judgements of performance. Analysis showed that comments on performance form part of every review. These are accompanied from comments on – on average – 3 to 4 other **Recording Elements**, among those identified in the present chapter (M = 3.45 elements/review, SD = 1.34, range = 1-7).

Critics discuss **Recording Elements** in relation to the final judgement of the end-product recording in two ways: *cumulatively*, by listing the different elements as separate value adding or detracting features; or *interactively*, by discussing how **Recording Elements** influence the perception and appreciation of the performance. These two kinds of composite judgements – that is, judgements based on the evaluation of two or more components of the end-product recording – often co-exist: 53% of reviews included instances of both types of judgements. The following section reports a few examples of these two forms of relationship between comments on performance and on other **Recording Elements**.

Cumulative value of recording

When critics recommend or not a recording to readers, they discuss how qualities of the different elements build together, and at times also what relative weight each element should be given. Cumulative judgements of recordings were found in 86% of reviews.

The element most often accounted for in these judgements, beside performance, is the ***Recorded Sound*** (69 reviews), followed by ***Composition*** (30 reviews, mostly related to the repertoire recorded, but also to the greatness of the performed piece), ***Recording Production*** (20 reviews, almost exclusively discussing spacing issues), ***Supplementaries*** and ***Price*** (12 reviews each), sound of the ***Instrument*** (5 reviews), fame and greatness of the ***Pianist*** (4 reviews) and *Policy* (3 reviews).

In Gulda's live recording of Sonatas Opp. 2/2, 27/2, 110 and 111 a few spurious noises in the recording are considered a negligible drawback against the quality of the performance:

> "The magnetism of the occasion is admirably caught, with enthusiastic applause between each item, audible gasps and sighs from the pianist (evidence of the immense effort of interpretation beneath a seemingly imperturbable surface) and a snapping string at 5'55'' in Op.111 that sounds like a rifle shot; marginal issues when set beside Gulda's fluence, grace and honesty" (Morrison, December 2002, p. 72)

Recording Elements like the sound of the recording and of the instrument, repertoire performed and the way this is spaced within the discs, are given an important weight in critical review composite judgements. Quality of performance emerges as the most relevant criterion to commend or not a recording; however, other **Recording Elements** play an important role in the final decision so that lack of quality in one or more of them can be discussed as reason enough to reject even a good performance:

> "The stereo and mono recordings are good, and I notice that no other disc offers these particular sonatas together, but in a field that is intensely competitive I doubt whether performances as seriously flawed as these are can claim much attention" (Plaistow, March 1965, p. 57)

> "But I cannot think that Kentner has earlier showed the music in a good light: the scherzo can with advantage be more volatile than this, and the first movement, more particularly, can be very considerably grander. There are times in Kentner's performance when the pulse seems all over the place. It is a view of the music I cannot share, but yet hope that others may; for it is a great boon to have this glorious music, decently recorded, available on a disc about which so much thought has been taken and which is on offer at such a moderate price" (MacDonald, March 1965, p. 58)

> "Edwin Fischer (ALP' og4) gives a much better performance, but it is presented in so muffled a recording that its virtues are largely negated. Walter Gieseking (33CX1073) is exceedingly accomplished, if a little lightweight in the Finale. The recording is tonally much better, but the surface very spluttery. By far the most enjoyable and attractive performance comes from the young Swiss-American pianist, Orazio Frugoni (Vox PL7t60). His freshness and his clean-cut lines make the sonata live as it does not under the fingers of the more celebrated executants. His version is also by far the best recorded (and economically coupled with both the "Moonlight and the "Appassionata" - Gieseking has the former, Fischer just the latter)" (Porter, June 1954, p. 42)

In the following example, coupling of sonatas (***Composition***) and ***Price*** are discussed as decisive factors among recordings that already offer a high quality of both performance and ***Recorded Sound***:

> "You could say (but only just) that Annie Fischer brings more poetry and flow to the opening than Richter-Haaser, that the latter is more meticulous about observing Beethoven's dynamic markings than Kempff, and that the latter is more rapt at the very end.
>
> But these differences are very slight. The Richter-Haaser is one among several good performances, mostly well recorded, and your choice will be conditioned by the backing, and also by economics. Prices scarcely vary, but the Wuehrer disc has the advantage of including three late sonatas, whereas the others find space for only two" (Fiske, February 1961, p. 48)

Finally, in this review of Gulda's recording of the complete cycle, after an intense expression of appreciation for Gulda's performance, the critic continues developing an overall judgement where performance and ***Price*** are weighted against the ***Recorded Sound***, production process in terms of spacing of sonatas (***Recording Production***) within the discs and insightfulness of the sleeve notes (***Supplementaries***):

> "Considering only Gulda's playing of the music I would have no hesitation in stopping at this point with an unqualified recommendation; but there are other factors in the situation as a whole which must be brought into account.
>
> First among these is the recording, which, though serviceable enough, is not really among 1969's best. The sound lacks depth and richness; and … it will be an exceptional reproducer which does not add some background of its own, for the recorded level is distinctly low … no fewer than six sonatas are split between two sides. It is difficult to believe that with modern techniques this is essential…
>
> The new set comes complete with notes on Gulda … The notes are anonymous, but even in good translation (by Richard Rickett) betray their Germanic origin from time to time. Thus in discussing (I think) the evolution of sonata form, we have "No longer are relations valid only within a single formal unit; they have become ambivalent and extend far beyond the passing moment, out into a region of intersection that can only be comprehended by 'listening into the distance', or by what Heinrich Schenker, scientist, artist and critic, defined as Urlinie'. I would not claim that this has no meaning; only that if it has then I cannot myself see it, and I fancy I shall not be alone. …
>
> These are of course small points in relation to the great virtue of this issue, the quality of Gulda's playing. The quality of the Philips recording is not such a small point, but I would not like to exaggerate its defects. …
>
> On whether two further points are, for him, small ones only the reader can decide: one is the price of the issue - eleven records for less than £8 is a temptation indeed; the other is the

> limited availability of this price with a deadline at the end of February. I hardly need to recommend prompt action for any readers who do in fact find the temptation of the records themselves irresistible" (MacDonald, January 1970, p. 56)

Elements influencing performance appreciation

The different **Recording Elements** are not only discussed as separate items, which add or detract from the final judgement; but also as elements that interact in more subtle ways, by influencing the perception and appreciation of the performance. Judgements built on the interaction between performance and different **Recording Elements** were found in 63 out of 100 of reviews. The **Recording Elements** most often said to influence performance are the ***Recorded Sound*** (27 reviews) and the sound of the ***Instrument*** (10 reviews).

These two elements merge and interact with the sound of the performer, facilitating or impairing a clear and unified portrayal of the music and the production of a varied, sustained and full sound, which are properties praised in the performance, as shown in Chapter 7.

> "…but the recording and his piano don't give him enough room to convey the bigness of the music" (Robertson, November 1936, p. 17)
>
> "The recording accommodates a wide dynamic range without strain" (Chissell, February 1970, p. 54)
>
> "The sound, like the playing, can be both grand and awesomely quiet. Above all, it offers a persistently clear view of the rich ensemble of inner voices that is so vital to Brendel's purpose" (Osborne, February 1996, p. 7)

The contributions of ***Performer***, ***Instrument*** and ***Recorded Sound*** to the final result cannot be taken apart clearly at the perceptive level, and critics discuss and question what contribution each element brought to the final result:

> "Comparison with the stereo quickly confirmed that the unnatural perspective is not due to misjudgement on the part of the pianist" (Plaistow, June 963, p. 36)
>
> "Perhaps M. Casadesus used the sort of piano which sounded that way, and it is a very faithful recording?" (Porter, May 1956, p. 49)

In a few cases, spacing and editing (2 reviews, ***Recording Production***), coupling of sonatas (2 reviews, ***Composition***), and content of ***Supplementaries*** (3 reviews) were said to influence performance appreciation, inviting the listener to approach the performance from new perspectives:

> "The re-coupling of Brendel's account of the Waldstein Sonata with the Pastoral throws new light on the performance, giving new validity to a reading which stresses the music's airy brightness, its alfresco colours.
> Brendel's note (which talks of "an alfresco landscape", though it is difficult to conceive of a landscape which isn't alfresco) also leads us to look at the Sonata Op. 31 No. 1 in relation to the Waldstein, but the Pastoral coupling is imaginatively right" (Osborne, April 1982, p. 66)

> "But above all I would point out, on the practical side of things, the excellent spacing of the disc … This is, for once, entirely adequate: it makes possible the full enjoyment of either sonata on its own, or, if that is the required programme, of the two sonatas in succession" (MacDonald, January 1965, p. 59)

> "The effect overall, each time, is overworked. A current doesn't run through. Not one of them gives the feeling of a performance - or as a colleague put it, of a player having hit the ground running. Too much editing? Well, I wonder, and the unconvincing timing of the many pauses in the capricious E flat Sonata, Op. 27 No. 1, may be a tell-tale sign" (Plaistow, January 2002, p. 81)

> "…here, in the dramatic rests after detached fortissimo chords, it is interesting to reflect on what the accompanying booklet (and it is very detailed and scholarly) describes as the "rather vague English damping" (actually not so very vague here) preferred by Kalkbrenner to the "dead" Viennese-type cut-out" (Chissell, October 1980, p. 71)

In addition to the influence of elements like ***Recorded Sound***, ***Instrument***, ***Recording Production***, ***Composition*** and ***Supplementaries***, critics discuss how the perception and evaluation of the performance is influenced by thoughts and information on the process of reviewing (***Meta Criticism***_*Process*, 11 reviews), previous experience with **Recording Elements** (***Meta Criticism***_*Reviewer Perspective*, 8 reviews) and on the kind and level of challenges with which the performer had to cope (***Portrayal***_*Difficulty*, 25 reviews).

Reflections on the process of reviewing and on reviewers' opinions and expectations in relation to **Recording Elements** are used to relativize a judgement or to explain how certain merits and faults should be weighted:

> "I have never cared for the C major sonata (Op. 2, No. 3) and Schnabel does nothing to make me like it more" (Robertson, April 1936, p. 18)

> "Any interpretation of a sonata must, naturally, be judged as a whole, not by the separate movements: and therefore the very slow tempo adopted here for the opening movement must be related to the movements following" (Robertson, October 1945, p. 16)

> "It will necessarily be the case that somewhere in an undertaking of this magnitude every listener will have some favourite passages he will wish Gulda had taken differently; but I fancy that few of us will find anything like agreement on which those particular passages are" (MacDonald, January 1970, p. 56)

> "If, for my desert island, I had to choose between Beethoven sonatas played on instruments of their own period and the modern piano, I would choose a splendid full-size Steinway without a moment's demur. …I also – perhaps ignobly – cherish a secret suspicion that some of our present-day antiquarians woo these old instruments because in this fiercely competitive world of great names, it is the only way they can make a modest voice heard. So how good to be able to say that no one has come nearer to making me think again about all that than Malcolm Binns" (Chissell, October 1980, p 71)

Finally, comments on *Difficulty* are used to emphasise the value of the performance as the performer's – or the engineers' – achievement.

> "S.'s amazing finger work, fine phrasing, and clarity of exposition are going to carry him as successfully through the tortuous fugue as any man can hope for" (Robertson, November 1936, p. 17)
>
> "When the most difficult things are so marvellously done it is all the more frustrating that some of the easy ones are spoiled" (Fanning, April 1992, p. 111)

DISCUSSION

In this chapter, the selected corpus of reviews ($N = 100$) was analysed to outline what elements of the recording – beside performance – critics discuss, and how and to what extent considerations on these elements enter the final evaluation of the end-product recording. Results show that along with the description and evaluation of performance, critics also comment on ***Composition***, ***Recorded Sound***, ***Recording Production***, ***Performer***, ***Supplementaries***, ***Instrument***, ***Composer***, and ***Price***. They either offer factual information about these elements, or portray and evaluate them. Furthermore, critics engage in meta-reflection on the process and purpose of reviewing, on their own experiences with the material at hand, and on other people's comments on the recording. Among these different activities, evaluation emerged by

far as the prominent type of statement in critical review in relation to recording elements beyond performance, thus supporting the results of Chapter 6 and the view of music performance criticism as an essentially evaluative activity (Calvocoressi, 1923; Newman, 1925; Walker, 1968; Cone, 1981; Carroll, 2009).

Similarly to the results of Chapter 6, findings in this chapter showed the relative weight given to each element of the recording in critical review to be highly consistent between different critics, suggesting that the emergent visual descriptive model describes traits of critical review that have general validity for the review corpus investigated, independently from the work and performance reviewed, the reviewer style and background.

These results lead to three main observations, concerning (1) the nature of value judgements of recorded performance, (2) the different values a recording may possess, and (3) the role critical review holds in the music market.

Recording evaluation criteria

The evaluation of performance has been made object of a vast corpus of research in the past decades, and several elements of performance (musical and extra-musical, McPherson & Schubert, 2004) relevant to the final judgement have been identified (see Chapter 1). These studies have almost exclusively focused on the evaluation of live performances – partly because of the educational context in which they were mostly set. When recorded performances were used, as for instance in Duerkson (1972), the focus remained on the performance-related aspects of the recording. As such, no study so far has offered insights on the criteria that are actually applied in the assessment of recorded performance.

The present research moved a step in this direction mapping, for the first time, elements of the end-product recording other than performance that are discussed and evaluated in critical review. The emergent model reveals that in the evaluation of recorded performance, as in the evaluation of live performance, there seem to be musical as well as extra-musical factors that concur to the final assessment. The nature of these extra-musical factors and the way they interact with and influence the appreciation of the performance renders recorded performance a unique product, distinct from live performance.

Critical review is interspersed with comments and reflections on different aspects of the recording, which combine and build together shaping the final, aggregate judgement. The quality of the performance, in this context, becomes but one – most important – element that prospective consumers should be taking into account in their choice. Analysis showed that even though performance quality has a privileged position, it is – taken alone – not sufficient for a critic to commend the recording to readers. The support of other elements does not merely offer added value: it is necessary for the recording to be successful.

This is true especially for those elements that directly influence the aural outcome of the recording: first of all the quality of the ***Recorded Sound***, but also the sound of the ***Instrument*** and the way performances are distributed within discs (***Production***). However, in a market burgeoning with recordings that offer plenty of musical and flawless performances with a high level of recording quality, elements like repertoire, content of sleeve notes, or price can become decisive factors in the final assessment of the product.

One point in this respect concerns the praising of accompanying notes. In the past few years, studies have begun to investigate the influence of verbal information on the enjoyment and understanding of music performance. Initial findings suggest that providing (adult) listeners with information on the performance while or prior to listening increases focused listening but lowers affective response and enjoyment of the performance (Silveira & Diaz, 2014, Margulis, 2010, Margulis, Kisida & Greene, in press). Critics discuss accompanying notes as informative and insightful guidance for listeners. However, following up on the results by Halpern (1992), Margulis (2010) and Margulis et al. (in press) – who found that programme notes' usefulness depends on listener familiarity, type of information presented (contextual information vs. analysis) and style (technical vs. metaphorical) – it may be worth further investigating what kind of text may actually offer proper guidance to specific target readerships.

Comparisons between recordings were found to be only rarely used – 13% of reviews entailed comparative judgements between **Recording Elements**. This suggests that comparative judgements are less relevant to the evaluation of **Recording Elements** than to the assessment and description of the performance.

Not surprisingly, ***Instrument*** quality is often discussed in relation to historical instruments, as it would be expected also in criticism of live performances. The quality of the ***Recorded Sound*** and the distribution of pieces within the discs on the other hand are recording-related factors. Criteria used to evaluate ***Instrument*** and ***Recorded Sound*** overlap with performance assessment criteria found in Chapter 7: the sound of the instrument and of the recording interact with performance and influence qualities like richness and warmth of tone (**intensity**), clarity (**coherence**), and dynamic and timbral variety (**complexity**). Given the primary role these properties have in the evaluation of performance, the weight accorded to ***Recorded Sound*** and ***Instrument*** in the assessment of recordings seems justified.

In addition to these evaluation criteria linked to performance, one more group of criteria used to assess ***Recorded Sound*** and ***Production*** reflects a notion of recording as 'portable concert', in line with Clarke's (2007), Philip's (2004) and Katz's (2004) discussion on the influence of recording technology on listening.

Elements inevitably linked to the process of production – including sounds and noises that reveal the mechanics of the process – should remain in the background, possibly be unnoticeable, so as to permit as natural an experience of the performance as possible, where natural is meant as resembling the experience of a live performance. The listening of a piece should be uninterrupted, free from mechanical necessities like the turning of the disc. ***Recorded Sound***, prior to being aesthetically pleasurable should be realistic; a faithful, transparent image of the sound of the performance. In addition, the ways pieces are coupled and distributed within discs are comparable to the assembling of a concert programme: critics praise couplings and distributions that not only minimise breaks, but also maximise logic and aesthetic meaningfulness, offering illuminating insights through the juxtaposition of certain pieces.

This emphasis on the fidelity of the recording medium and on the recreation of an experience as close as possible to that of a concert performance questions the notion of 'realism' and 'naturalness' in relation to a recorded performance. As Philip (2004) argues, in a recorded performance, even when the editing is kept to a minimum – like in the recordings produced by Nimbus – the end result is always significantly shaped by the interaction between performer, engineers and resources employed. The importance of this interaction has been emphasised recently by Pras

and Guastavino (2011) in their investigation of the role of musicians, music producers and sound engineers in the recording context. Through interviews with 16 musicians and 6 sound engineers, they found that 'interaction' was one of the three main themes discussed with reference to an ideal recording process; the other two main themes were 'skill' and 'mission'.

Furthermore, as Philip (2004) suggests, ***Production*** and ***Recorded Sound*** are used to add value to the end-product recording creating a product that gives the impression of a live performance experienced from the best possible seat and the ideal concert hall. Patmore and Clarke (2007), discussing the work of the record producer John Culshaw, describe this phenomenon as the tension between "capturing performances" and "creating virtual worlds" to which producers and engineers are exposed, and the way in which different producers and engineers responded to this challenge is at the core of the debate between 'minimalists' and 'interventionists' in the music recording market (Philip, 2004).

In light of this, the notion of 'realism' and the idea of recordings as portable versions of live performances – despite its weight in the way we experience and evaluate recordings, supported by the present investigation – seem paradoxical. What is strived for is the perceptual experience of a natural, realistic performance, rather than a physical notion of being close to the original product. In support of this, the importance given to the notion of realistic performance in critical review judgements is counterbalanced by the critics' appreciation for the end-product recording as collective result of several people's achievement (e.g., comments on the *Difficulty* with which engineers had to cope).

Another factor emerged from the analysis as relevant to the final assessment of recordings is the series of expectations, thoughts and beliefs critics entertain. A few studies in past decades have suggested that expectations on the quality of the performance affect musical preferences (Duerkson, 1972; Ziv & Moran, 2006-). Critics' comments on their previous experiences with and opinions about given works, pianists or instruments seem thus to warn the reader about potential biases, setting a context in which their evaluative statements can be more properly interpreted. This echoes the increasing importance given to critics' identity discussed in Chapter 3: through their meta-reflections, critics remind us that their assessment is the opinion of but one – expert – person, rather than an 'objective' judgement.

Finally, in addition to *Evaluative judgements* also *Portrayal* judgements emerged at least partly relevant to the final assessment. The notion of *Difficulty*, for instance, was spread widely across reviews, emphasising the value of the recording as expression of the performer's and the engineers' achievement.

In summary, the emergent model of recording evaluation emphasises the importance of non-performative components in the perception and assessment of the end-product recording. Critics emphasise that the distinction between the different elements and their direct (impact on aural result) or indirect (information, expectations that colour perception) influence on performance cannot always be made at perceptual level and often requires a certain amount of assumption on the critic's side. Nonetheless, critics' attempt to disentangle the value and peculiarities of single elements, also accounting for the different weights readers may give to one or the other component. In his seminal book, Philip argues that to be adequate judgements of recorded performances need to account for the way in which the sound "got onto the disc" (Philip, 2004, p. 26). In support of this, the present study shows that characterisation and evaluation of elements linked to the production process and to the object-recording do indeed play a substantial part in critical review judgements.

Values of a recording

Holistic evaluations of recordings by critics evidenced the variety of values that the end-product recording can possess. Beside the artistic value – discussed mostly in the evaluation of the performance – and the success value emphasised by the *Difficulty* theme, recordings are praised for being pedagogically valuable, historically significant or even socially relevant. These are all different kinds of value that an artwork can usually possess, what Budd discussed as instrumental values of artworks (1995). In music, however, these values are imbued with a particular meaning when discussing recorded versus live performances.

Two main characteristics of music recordings, which have a strong impact on the way we listen to music, are their semi-permanency and their repeatability (Clarke, 2007). These characteristics of recordings, for the first time in history, pulled music performances out of the instant-dimension of time, to give them the status of collectable objects. Very early on in the history of recording consumers

reacted to the object-like nature of recordings, transferring attitudes and practices typical of other consumption fields like bibliophilia to music (Morgan, 2010). Recordings became more than affordable ways to listen to good music; along the century, they have come to be invested with the status of historical documents, able to witness a certain performer achievement or artistic style, a given technology or instrument outcome for generations to come (Katz, 2004).

The relevance of the collectability attribute for the overall value of the end-product recording finds support in the present investigation of critical review. Critics express their recommendations explicitly, suggesting (or not) recordings to specific interest groups, and evaluate recordings based on how interesting for collectors certain couplings of sonatas may be, or on how a disc fills a hole in the market or completes a certain performer's cycle.

Critics' role

A last reflection concerns the role of critical review and, through it, of critics in the classical music market as it emerges from the *Gramophone* text. ***Meta Criticism*** comments on the *Purpose* of critical review, together with the focus on ***Price*** and *Market* situation, evidenced the view of critical review as a form of guidance for readers, when it comes to decide what product to buy (Frith, 2009; Pollard, 1998). Throughout the critical text, critics explicitly commend or not recordings to the reader offering price for value considerations and pondering the diverse factors like repertoire recorded (length, coupling), quality of recording and performance, and price and quality of other recordings of the same pieces on the market. They also account for different readerships, suggesting a certain choice to one group of listeners but not another.

The idea of critics functioning as consumer guide is also reflected in readers' letters to the *Gramophone* editor. Readers look to critics for solid recommendations on which recordings are worth purchasing. This leads also to different requests concerning what aspects of the recording critics should focus on and in what form they should deliver their judgements. From the very beginning, *Gramophone* readers discussed the relative weight that critics should give to the performance versus other **Recording Elements**, appealing to critics with their different opinions and wishes, as it is witnessed in these few excerpts taken from letters to the editor:

> I must say, I have rather sympathised with the occasional complaints that your reviewers often tended to lose themselves in details and fine points to the detriment of the general balance of their criticism, and it has been especially hard to figure out the general estimate placed on a given recording, and the place assigned to it in relation to other recordings. (*Gramophone*, December 1934, p. 71)

> Excellence of recording, and first-class technique we should now be able to take for granted always, in recommended records. What we do most earnestly desire is the finest available interpretations. It is not easy to decide how best to lay out our hard-earned shillings when buying records, and I personally lean heavily on your reviewers then. (*Gramophone*, March 1940, p. 42)

> As a regular reader since 1953, am I alone in thinking that – particularly with the advent of Compact Disc – your reviewers are reflecting an increasing struggle, in comparative evaluations, between sound quality and performance? …As an 'established' collector, I am well aware that the records I play repeatedly of particular pieces of music are chosen almost entirely on the basis of performance – not recording. Reflecting back, I wish I had not followed the GRAMOPHONE recommendations but had followed my nose… (*Gramophone*, December 1986, p. 7)

> I believe the magazine title tells us it is about reviewing 'recordings' first and foremost. Indeed, I am often dismayed that so much space is given to analysing performances that sound quality is relegated to second place. Like many collectors, I have a fair idea of the musical qualities and interpretative styles of most of those whom one comes across on record… (*Gramophone*, February 1987, p. 5)

In addition, comments on *Recording Policy* found in the text suggest that *Gramophone* critics play an active role in the shaping of the recording market. Their reviews address not only consumers but also producers. The role of critics as mediators between producers and consumers is evidenced again in the correspondence published in the *Gramophone* pages. Along the century readers used the magazine as a channel to reach recording companies, for example voicing pleas for new recordings, as in the following excerpt:

> You have had so many letters lately about recordings (or the lack of them) of British music…that I have decided to add my cry. …For a start, I have almost worn out my record of the Elgar 'Cello Concerto (Navarra/Barbirolli). I regard this as Elgar's greatest achievement and

> never cease trying to win others over to it. How about a brilliant new recording of this masterpiece under Barbirolli, Sargent or Boult with Navarra or Tortelier as solo 'cellist. I only have one set of it. Certainly, I would not hesitate to order a new, modern recording of it. …
> I will close by requesting two other works which I treasure on cumbersome 78s: Bax's Third Symphony and E. J. Moeran's Symphony. Both would sound thrilling in stereo. But, please, the Elgar 'Cello Concerto as soon as possible and let Enigma rest for a while. I realize how superb the new Barbirolli of it must be, but my purse just refuses to let me listen to it. (not readable, probably 1950s)

In this context, discussions on the criteria critics should apply in their assessments are charged with a sense of responsibility towards the standards of the art that critics should convey:

> …it is more than ever important to study and pass judgment on the records as records. The danger of gramophones is not so much that they may encourage people to like bad music, but rather that they may set up a false standard of what recorded music ought to sound like. (October 1925, p. 44)

> In the highest interests of music we must have the very best musicianship. Marvels of technical accomplishment, celebrity-worship, and the mere fact of outstanding sound recording must be relegated to their proper levels—the music's the thing! I think the time has come to take stock of cherished recordings. …THE GRAMOPHONE alone can perform such a service with the necessary authority; your reviewers are the men for the job. More power to them, and to all of you. (March 1940, p. 42)

These *Gramophone* readers' letters resonate with the results of the present investigation and with the findings of Chapter 3, supporting the view of critics as mediators between producers and consumers in the cultural industries, able to shape the public's preferences, establish an artist's reputation, his/her career success or failure, and help the end consumer in their purchase decision (Debenedetti, 2006).

The present findings, based on an examination of critical review produced along almost 90 years, suggest that critics have played a relevant part along the century in consumers' decisions on what to buy and listen to and maybe even in labels' policy on which recordings to produce. The extent to which the recent advent of online resources as mp3 downloads, Spotify and iTunes files has influenced or is influencing the role and importance of critical review in the classical music market is still unknown. Further research will be needed to address these questions and

investigate – at present day – the impact of music critics' judgement on consumer choices and on the establishment of a canon of master performances.

CONCLUSIONS

The present chapter reported methods and findings of the third and last layer of thematic analyses run on the corpus of 100 reviews. The results extend the previously developed models (Chapters 6 and 7) by illustrating how considerations on all components of the end-product recording enter the critical review judgement. In so doing, findings of this chapter conclude the present research completing the answer to the initial question – what reasons do critics bring in support of their value judgements – in relation to elements of the recordings beyond the performance.

The answer can be summarised in terms of quality of elements common to any performance event (i.e., including live performances: ***Composition***, ***Composer***, ***Performer***, and ***Instrument***) and of elements specific to the recording product (***Recorded Sound***, ***Recording Production***, ***Supplementaries***, ***Price***). In terms of evaluation criteria, intersections were found with evaluation areas emerged in the investigation of performance judgements (**intensity**, **complexity** and **coherence** for the assessment of ***Instrument*** and ***Recorded Sound***; **endeavour** for the discussion of *Difficulty*; **comprehension** in relation to ***Supplementaries***, ***Production*** – distribution – and ***Composition*** – coupling, cf. Table 8.2). In addition, one more criterion emerged linked specifically to the recording medium: the extent to which the recording affords a live performance-like listening experience.

Results of Chapters 6 to 8 emerged from separate analyses, hence cannot be directly compared. However, they are complementary in that they focus on different sections of the critical review text and, taken together, offer a comprehensive and detailed map of review content that offer a conceptual basis for future investigations of other corpuses of critical writing.

This chapter concludes the presentation of empirical outcomes of the present research. In the next and final chapter, methods and findings of the whole investigation of *Gramophone* reviews reported throughout the thesis are summarised, their limitations as well as theoretical and practical implications are discussed, and suggestions for future research directions are given.

9 GENERAL DISCUSSION AND CONCLUSIONS

In the present research, a vast sample of recorded performance critical review was assembled in order to examine its content. The collected texts encompassed 845 reviews published in the *Gramophone* magazine over almost 90 years (1923-2010). These totalled 334,210 words of critical text, discussing 640 different performances plus 205 re-issues performed by 216 pianists and reviewed by 52 critics.

A novel combination of data reduction and thematic analysis techniques were employed to describe and categorize the text corpus. The outcomes of this investigation offer – for the first time – empirical evidence of the content of music critics' writings and open a new perspective on the broader performance evaluation discourse, shedding light on the richness and potential of this common professional and commercial form of music written response. This final chapter draws together the research presented throughout the thesis. It summarises methods and outcomes of the investigation, reviewing the efficacy and limitations of the applied design, and then discusses theoretical, practical and empirical implications of the emergent findings.

SUMMARY OF RESEARCH AND OUTCOMES

The main research question of this thesis was: What reasons do critics adduce to support their evaluative judgements of recorded performance? Table 9.1 summarises the studies that have been carried out to answer this question. Following, the main emergent ideas are presented, and the research design and methods applied are critically discussed.

Table 9.1. Synopsis of methods and findings for the six studies reported in the thesis (Chapters 3 to 8). As in Chapters 6, 7 and 8, words in the right-hand column that appear in bold, bold italics and italics (capitalized) correspond respectively to theme families, main and sub-themes. Terms non-capitalized in bold correspond to evaluation criteria.

Chapter 3: Gramophone reviews I: An overview	
Method and analyses	*Main findings and outcomes*
Extraction of review texts from the online *Gramophone* archive (1,050 magazine issues).	*N* = 845 reviews of recordings of Beethoven's piano sonatas were collected, published in the *Gramophone* between April 1923 and September 2010.
Creation of a searchable database.	Collected reviews encompass 640 different performances, 205 re-issues; 216 pianists and 52 critics.
Descriptive and inferential statistics on critical review metadata.	Reviews are strongly polarized around a few highly expert critics (62.72% of reviews written by 10 critics with a mean period of activity of 21.32 years) and a few performers (51.95% of reviews concern 17 out of 216 pianists).
	Critics' identity becomes more explicitly defined over the course of the century (from unsigned reviews, to initials, to full names).
	Comparisons between performances are found in 53% of reviews.
Chapter 4: Gramophone reviews II: Turning to the text	
Method and analyses	*Main findings*
Five-step data reduction procedure, including qualitative/quantitative analysis of vocabulary and word stem patterns and comparisons between critics in different periods (using ReadMe and LIWC applications, Hopkins & King, 2010; Tauchszik & Pennebarker, 2010).	*Outcome*: Selected corpus of 100 reviews, to be used in the subsequent thematic analyses.
	The discussion of the performance covers about half of the critical review text (estimation for the whole dataset: 53.50%). This percentage increases over the course of the century, from 36.38% to 60.17%.
	An important portion of text is also devoted to the discussion of the recording (16.73%) and the composition (9.09%).
	Use of different semantic categories (in words per review) vary more strongly between critics than between periods of publication (averaged Kruskal-Wallis across 13 semantic categories: $H_{10} = 52.30$, $p = .037$ between critics, $H_6 = 25.56$, $p = .086$ between decades).
	The categorization of critics' vocabulary in music-related semantic categories evidenced a need for clarification of the notion of 'expression' used by critics.

Chapter 5: A complex notion: Expression in music criticism

Method and analyses	*Main findings*
Qualitative analysis of statements entailing the term 'express' and variants of it, using a *keyword-in-context* procedure (Namey et al., 2008).	The term 'expression' and its variations are sporadically used in critical review (18.36% of reviews). 'Express' and correlated terms are used at least with four different meanings, indicating (A-statements) the way performers use performance options (e.g., timing or dynamic); (B) the portrayal of music structural patterns; (C) the outer manifestation of (defined or undefined) inner states; and (D) properties of the music composition. The prevalent uses of 'express' (A and C-statements) evidence the bi-dimensionality (physical and psychological) of the notion of expression. Intransitive (C)-use of 'express' is inherently positively loaded, used as an independent value-adding property in performance. (A)-use of 'express' is valence-neutral, its positive/negative value depending on the combination of performance acts and the musical context. Results emphasise the ease with which the musical discourse slides (often unnoticed) from one dimension to the other.

Chapter 6: Critics' judgements of performance

Method and analyses	*Main findings*
Inductive thematic analysis of critics' judgements of performance as they are stated in their published reviews, run on the corpus of 100 reviews. A three-step analysis protocol was developed based on Williamson et al. (2011). This employed double-coder development of codebook followed by iterative process of statement comparisons and code revisions.	*Outcome*: Visual descriptive model of performance judgement. **Evaluative Judgements** are the major component of critical review of performance (n = 1,502). Characterizations of performance are divided in **Primary Descriptors** (n = 719, properties of musical sound, level of energy, and mechanics of delivery) and **Supervenient Descriptors** (n = 1,404, higher-order impressions of the performance). Among **Supervenient Descriptors**, emphasis is given to presumed qualities of the performer (understanding, affective states, mental and moral qualities, intentionality, control, care, sensibility and spontaneity). Absolute and relative (taste-dependent) evaluative judgements co-exist in critical review, with taste-dependent judgements present in 47% of reviews. One in every five evaluations in critical review (n = 220) is a comparative judgement, in which the reviewed performance is evaluated against one or more other

	performances. The relative weight given to the different themes in reviews is highly consistent between critics (Cronbach's α = .986).

Chapter 7: Valence of performance judgements

Method and analyses	*Main findings*
Thematic analysis of the relationship between valence expressed in performance judgements and performance descriptors, run on the corpus of 100 reviews. Two-step analysis protocol was developed, with double-coder analysis of valence followed by 30 single-coder sub-analyses.	*Outcome*: Model of performance evaluation criteria in critical review. Thirty-five value adding qualities were identified used in critical review to support value judgements. These were grouped into seven areas of evaluation, linked to the aesthetic value of the performance (**intensity**, **complexity** and **coherence**), to its achievement related value (**understanding**, **sureness**, and **endeavour**), and to the appropriateness of each quality to the given music context (**suitability**). Evaluation criteria were reliably used across critics (Cronbach's α = .928). A tension (tightrope) characterises the relationship between evaluation criteria. Performance properties emerged as interdependent, so that an increase in one of them may neutralise, decrease or increase the value assigned to other qualities.

Chapter 8: Beyond performance: Reviewing recordings

Method and analyses	*Main findings*
Inductive thematic analysis of critics' judgements of recordings (extra-performance statements) as they are expressed in their published reviews, run on the corpus of 100 reviews. The same analysis protocol employed in Chapter 6 was used, followed by a systematic analysis of code co-occurrences between **Critical Activities** and **Recording Elements**. Upon completion of the analysis, an examination of critical composite judgements at whole-review level was run to clarify how comments on extra-performance elements relate to judgements of performance.	*Outcome*: Visual descriptive model of critical review of extra-performance recording elements. Eight different components of recordings were identified (beside performance) discussed by critics and weighted in the final judgement: four elements common to any performance event (***Composition***, ***Performer***, ***Instrument***, and ***Composer***) and four specifically linked to the nature of recorded performance (***Recorded Sound***, ***Recording Production***, ***Price***, and ***Supplementaries***). Three kinds of **Critical Activities** emerged, with eleven sub-activities, in relation to one or several of the **Recording Elements**. **Judgement** (n = 467) was the prominent activity in critical review of extra-performance features of recording: its sub-theme ***Evaluative Judgement*** was the largest single theme emerged in the analysis (n = 336). Its second sub-theme offered a ***Portrayal*** (n = 131) of the different **Recording Elements.**

Other activities were: offering factual **Information** (n = 194) or meta-reflecting on the process and purposes of reviewing, on the critic's previous experiences and other critics' judgements (**Meta-Criticism**, n = 119).

Through the analysis of ***Evaluative Judgements*** 16 value adding properties of extra-performance elements of recording were identified, used to support value judgements. These qualities partially overlap with the performance evaluation criteria developed in Chapter 7. In addition, two criteria emerged linked to the **live-performance impact** the recording offers and to the value of the recording-object as **collectable**.

Comments on *Purpose*, ***Price***, *Market* and recording *Policy* emphasise the role of critics as mediators between producers and consumers, and as guidance for consumers to decide what product to buy.

The narrative of critical review judgements at whole-review level showed that the value of the different **Recording Elements** is accounted for in the final, composite judgement (*M* = 3.45 **Recording Elements** discussed in each review beside performance).

Quality of **Recording Elements** is described as cumulatively or interactively adding to the quality of performance. **Recording Elements** described as interacting with performance qualities are ***Recorded Sound*** (n = 27 reviews), ***Instrument*** (n = 10 reviews) ***Supplementaries*** (n = 3 reviews), ***Recording Production*** and ***Composition*** (n = 2 reviews each).

Main empirical findings

Critical review as a rich source of data

Music critical review emerged from this research as a vast and rich source material, pliable to systematic investigation and open to a large variety of analytical approaches. Different analysis methods were successfully employed to investigate the material, from metadata analysis (Chapter 3) to word-stem and *word-in-context* analysis (Chapters 4 and 5) to thematic analysis (Chapters 6, 7 and 8). Even at metadata level, the examination produced relevant main and secondary findings, one example of the latter being the observation of historical trends in the repertoire reviewed (see Chapter 3).

The notion of expression

One of the most often employed and discussed terms in music research and in the broader musical parlance is 'expression'. Results of the present research show that this term has been used in very different ways in the critical discourse, to describe characteristics of the musical sound (use of timing or dynamics, for instance) that can add to or detract from the value of the performance; or to indicate the outer manifestation of emotions, characters, or other abstract concepts (like the revolutionary spirit) that are usually positively loaded but may be inappropriate in certain contexts. 'Expression' is also used in an intransitive way as a form of pure evaluation of performances, as a synonym of 'beautiful' or 'musical'. Moreover, it can indicate features of the work performed, that may or may not be related to the performance itself (Chapter 5). The ambiguity of the term 'expression' and the ease with which the critical discourse slides from one dimension of expression to the other, may explain why critics use the term rarely in reviewing, even though the diverse components of expression identified are found copiously spread in performance judgements (Chapter 6 and 7).

The nature of recordings

This research focused on critical review of recorded performance. In the course of the investigation the nature of recorded performance as a unique artistic product, distinct from live performance in its components and relevant evaluation criteria, became increasingly clear.

The large quantity of reviews of re-issues found in the *Gramophone* suggested that recordings are more than reproducible pocket versions of the performance they carry, and that the choice of what to review may be based on criteria other than the nature and quality of performance (Chapter 3). The thick-grained content analysis run in Chapter 4 corroborated this observation, showing that the discussion of performance accounted for about a half of the review text and that other aspects of the end-product recording, most noticeably the recording medium, had also a fixed part in reviews. These findings were then deepened in Chapter 8, where results revealed that four recording specific elements – ***Recorded Sound***, ***Recording Production***, ***Supplementaries***, and ***Price*** – form an essential part of critical review and enter the final judgement of the end-product recording.

Critics' role

This research showed that critics enact the role of critical review as guidance for listeners and mediator between producers and consumers. The metadata analysis showed that information on the identity of the critics gained increasing importance over the course of the century (Chapter 3). Critics emerged as highly expert listeners, with long-lasting careers and experience in evaluating and comparing performances and a well-defined writing style, in terms of vocabulary used and semantic categories applied (Chapters 3 and 4).

Comments on the *Purpose* of critical review, on ***Price*** and *Market* situation, and on recording *Policy* found in Chapter 8 clarified and reinforced these observations, emphasising the notion of critics as filter of choices and mediators in the music market.

Evaluation of recorded performances

The principal purpose of this research was to add relevant insights to the current discourse on music performance evaluation through the investigation of critical review. Indeed, evaluation emerged as the characterising activity in recorded performance review, being by far the most frequent and pervasive form of statement found (Chapters 6 and 8). Value judgements are expressed explicitly as well as implicitly through the use of value laden terms as descriptors. As result, about 90% of all critical statements in the present corpus are value laden (Chapter 7).

Thematic analyses allowed for the development of visual descriptive models of review content and the extraction of basic evaluation criteria used to assess the end-product recording. The overall emergent picture describes the content of critical review of recorded performance in terms of evaluative judgements, descriptive judgements (***Portrayal*** in Chapter 8, corresponding to **Primary** and **Supervenient Descriptors** in Chapter 6), factual information and meta-criticism. These diverse activities have as their object the performance (usually linked to evaluative and descriptive judgements) and one or more other **Recording Elements**. Drawing from the findings of Chapters 6 to 8, Figure 9.1 visualises the emergent synoptic model of critical review of recorded performance.

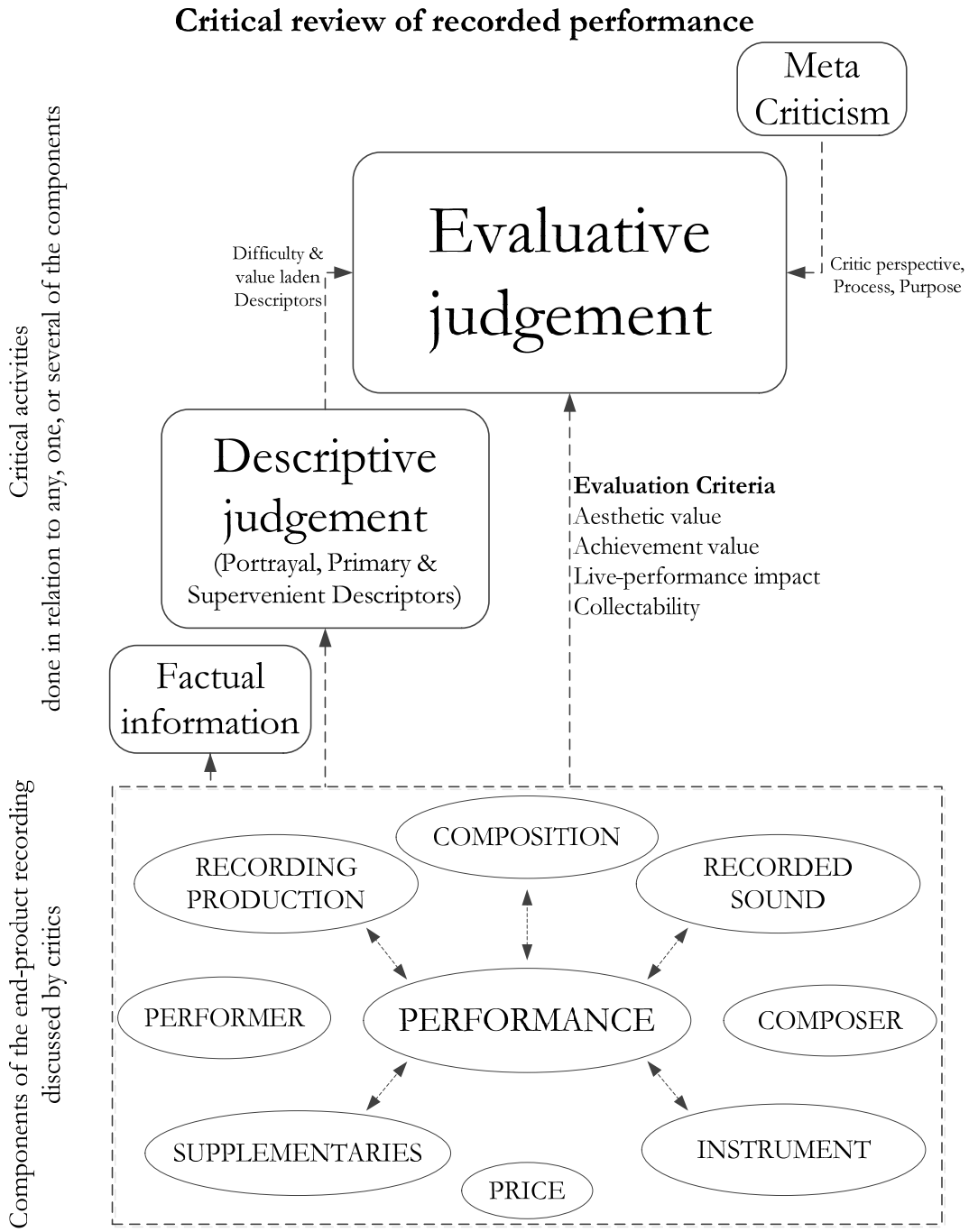


Figure 9.1. Descriptive model of critical review content, drawn from findings of Chapters 6, 7 and 8.

The nine different elements discussed by critics – that is, performance plus the eight recording elements – are visualised in oval shapes and localised in the bottom half of the model: Performance is given a central role, its importance evidenced in the reviews through the amount of text devoted to it (Chapter 4), but also through the density and variety of themes employed in its discussion (Chapters 6 and 7) and the

weight given to it in the final overall judgement (composite judgements, Chapter 8). The double-arrows between performance and five other **Recording Elements** indicate interactions described by critics as relevant for the final judgement (Chapter 8).

Rounded rectangles in the top half of the model visualise critical activities, with activities organized from left to right and from bottom to the top according to their level of abstraction from the elements discussed: starting with factual information, followed by descriptive and evaluative judgements, and finally by meta criticism reflections. All elements of the recording were linked to factual statements, which offer information without expressing any opinion or evaluation. In criticism of performance, factual information did not emerge as a theme on its own. Almost all statements about performance can indeed be considered judgements – that is, opinions or perceptions that imply some level of subjective interpretation. However, in retrospect, a few performance-related statements could be considered as information rather than judgements – for instance, those commenting on the realization or not of repeats.

Also, all nine elements were linked to judgements, either descriptive or evaluative. Elements can be directly linked to evaluative judgements, assessed as being good or bad, or better or worse than another. Beside the line of evaluative judgement the emergent evaluation criteria are summarised in terms of the values to which they relate (Chapters 7 and 8). Among these, aesthetic and achievement related values have a prominent role. In particular, this research revealed the importance of thoughts and assumptions about the person beyond the performance for listeners' appreciation. Descriptive judgements can also be used to trigger or influence an evaluation: this is the case for instance of comments on difficulty (Chapter 8) or statements using value laden descriptors to characterise performance (Chapter 7).

In addition, critical review presents a certain amount of meta-reflection on the process of reviewing itself. These considerations as well enter the evaluative judgements – expectations about the quality of a recording and considerations on the

reviewing process and purpose colouring the critic's appreciation.[26] Even though links were found in the text between meta-criticism and some of the elements, these are not shown in the model, since these comments referred to the critic's knowledge or opinion about a certain element prior to listening to the recording, and were not triggered by the actual experience of it.

Importance of comparison in performance evaluation

Comparative judgements emerged as a basic component of critical review, present in more than half of the reviews (Chapter 3). This form of judgement was used only marginally in the discussion of extra-performance elements of recording (Chapter 8), but it had an important role in the evaluation of performance: indeed, one in every five evaluations of performance in reviews was expressed in the form of comparison (Chapter 6).

Validity and interpretation of value judgements of recorded performance

One major point that emerged from this research concerns the boundaries set to the validity of value judgements in critical review. The research identified the major content components in reviews and the criteria underlining critical evaluation. These components and criteria emerged as highly reliable across critics, suggesting that the resulting model is representative of the review corpus analysed.

Some aspects of the model, however, indicate that the notion of value of recorded performance is listener and context dependent. Chapter 6 demonstrated the co-existence in critical review of absolute and taste-dependent judgements of performance, with taste-dependent judgements present in almost every second review. The results of Chapter 8 emphasised the relevance of personal experiences and potential biases towards elements of the recording in the interpretation of critics' judgements. The model of evaluation criteria in Chapter 7 was characterised by the suitability criterion, against which all other criteria have to be set, and Chapters 7 and 8 together offered examples of the different kinds of values a recording may possess raising a question on the weight given to the aesthetic value in the global appreciation and assessment of a recording.

[26] This process is indeed iterative, in that the actual experience of the recording flows back to inform and re-shape expectations. However, reviews did not offer evidence of this iterative relationship, therefore the arrow in the present model is unidirectional.

In the light of this, a given recorded performance may be valuable for one listener but not for another, for one purpose (e.g., offering a rewarding aesthetic experience) but not for another (e.g., being taken as reference for pedagogical purposes). Aesthetic and achievement related performance properties can also be valuable only in given musical contexts, or when correctly balanced by other properties. These results do not negate the possibility of having reliable judgements of performance, but they do point to the necessity of embracing a notion of value of recorded performance that is context and listener related. Within those boundaries though, the consistency with which critics were found to use different criteria – even critics born generations apart and who reviewed different recordings – suggests that even within this complex evaluation framework reliable evaluations may be possible.

One consequence of this relativity of recorded performance judgement is that the critic-listener communication becomes an essential element of critical review. In order to interpret this complex form of written response to music properly, readers need to account for a critic's preferences, expectations and beliefs (Chapters 6 and 8) and personal 'assessment style', in terms of use of certain vocabulary, weight given to different criteria or tendency to point out or not minor defects (Chapters 4 and 8). The increasing importance given to critics' identity noticed in the changes in review signing policy observed along the century (Chapter 3) seems justified. Critical review seldom offers ready-to-use recommendations on what to buy; most commonly, it invites the reader to engage critically with the text and to reflect on his/her past experiences, preferences and needs, in order to develop his/her own picture of and opinion about the reviewed recording, backed up by a partial awareness of how the recording is set within the broader market.

Suitability and limitations of applied methods

A challenging aspect of this research, and one of the objectives of the investigation, was to understand what analytical stance and what methods could be used to extract relevant and rich information from the critical review material systematically. The methodological considerations and study of extant literature on the analysis of unstructured texts reported in Chapter 2 led to the conclusion that an investigation of critical review would require quantitative and qualitative methods flexible enough to allow on-going development of analysis protocols.

The Applied Thematic Analysis (ATA, Guest et al., 2012) reflected this methodological approach and was then chosen for this research. As explained in Chapter 2, ATA is a practice-based, positivist/interpretative approach to qualitative analysis. In-depth thematic analysis is at the core of ATA, but this is accompanied by the conviction that data must be paramount in deciding at any stage what analytical method to use, without excluding *a priori* any theoretical and epistemological approach.

Based on the ATA paradigm, a new hybrid qualitative/quantitative analysis design was developed in this thesis, which included successive steps of quantification and data reduction techniques, as well as focused and inductive thematic analyses. This analysis protocol allowed a thorough examination of the review material, from the metadata, through the word-stem and *word-in-context*, to the clause and multiple clause analysis level. The process was semi-structured, in that results of previous analyses informed and shaped the subsequent ones. Hence, findings of Chapter 4 revealed the necessity of adding a focused analysis to clarify the notion of expression in criticism before moving on with the thematic analyses. The same findings also informed the structuring of the thematic analysis into two distinct layers, to investigate performance and extra-performance related statements separately. Results of Chapter 6 concerning the density of data prompted an additional development in the analysis protocol, and the insertion of a further, valence-focused examination of performance judgements (Chapter 7). Figure 9.2 schematises the applied design.

In phase one review texts were collected and organized in a searchable database. This database was used in phase two for a series of metadata and data reduction analyses that led to the selection of a representative corpus of material. At this stage the need for further preliminary analyses was established, and this led to the focused ‘expression’ analysis. Finally, in the third stage, thematic analyses were run on the selected material. A circular shape for the thematic analyses indicates the iterative nature of this process. The final outcome of the research was a novel descriptive model of critical review content.

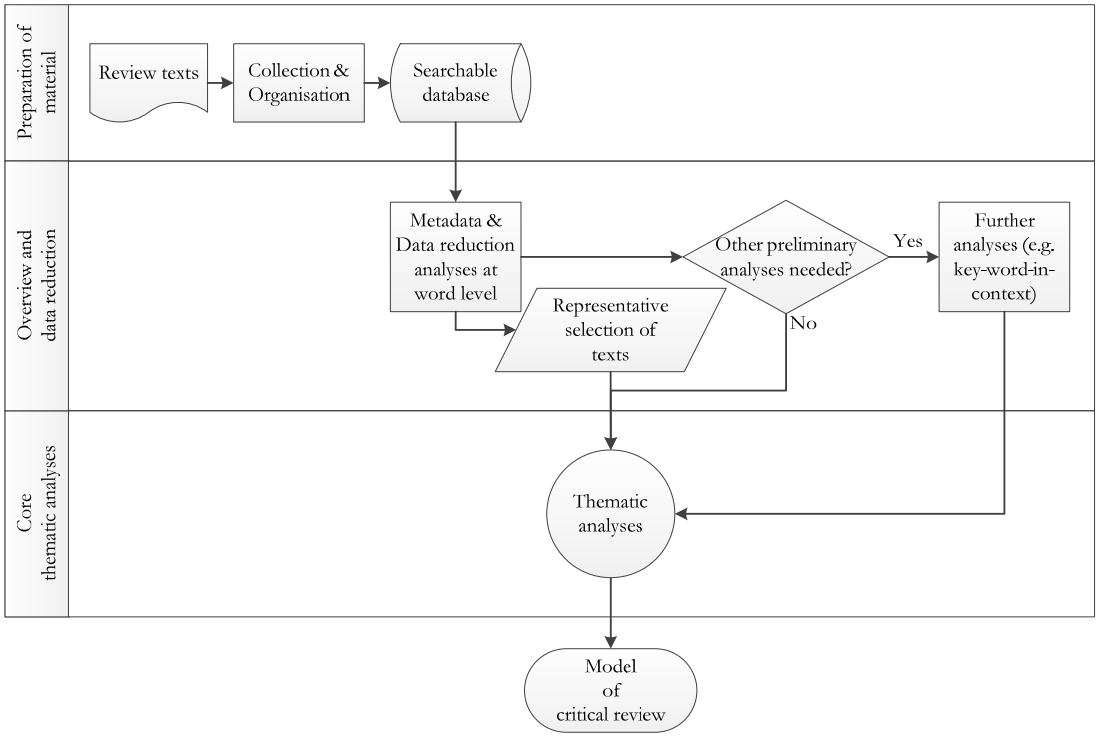


Figure 9.2. Schematic flowchart representation of the analysis protocol applied in the present research.

The combination of quantitative and qualitative methods rendered the large and complex corpus of material manageable and further allowed for cross-validation of findings. It also allowed for a comprehensive analysis of review content at different levels of textual- and meta-data. At the textual level, one further choice that had to be taken concerned what reader perspective to adopt in the interpretation of the text. Although the analysis was run inductively, without applying any *a priori* categories, the conceptualization and organization of review statements was necessarily coloured and shaped by the researchers' own knowledge and pre-conceptions about music and musical value in general, and about the chosen repertoire and relevant performance practices in particular. Indeed, an attempt at an analysis void of any researcher biases would have been not just impossible, but also lacking in validity, unable to offer an understanding of the text that would be meaningful in the real world critical practice (Mantzoukas, 2005). Instead, an effort was done in the present research to produce a model informed by the perspectives of two common types of review reader: the competent music amateur and the music professional. The results of the double-coder analysis and the ways in which the following discussions between researchers enriched the development of the code scheme supported the importance of this

perspective variety for the strengthening of the validity and robustness of the emergent model.

However, concerning the extent to which this investigation of critical review may add to our understanding of the performance evaluation process, three major limitations need to be pointed out, which relate to (1) the collected review corpus, (2) the analytical approach, and (3) the kind of data that critical review offer.

Firstly, restricting the investigation to *Gramophone* reviews of Beethoven's piano sonatas set limits in terms of cultural context and repertoire reviewed. This decision was driven by the necessity of assembling a corpus of material apt to systematic and detailed investigation and by the assumption that cultural background and even more repertoire may reasonably affect the review content. This restriction indeed allowed for a more in-depth examination of the material, free from possible confounding effects of these variables. Retrospectively, given the large number of reviews collected and the density of the data, this decision seems recommendable also for future studies. Nonetheless, it should be taken into account that the magazine and repertoire chosen for this research represent one of the world leading institutions for reviews of classical music recordings and one of the pillars of every pianist's classic repertoire. Other sources and repertoires may offer less and less dense material, thus potentially allowing for multiple-sources and/or multiple-repertoires investigations. Similarly, the focus on recorded performance represented at the same time a limitation and a gain: if on the one hand this meant the loss of essential elements of performance like context and visual components, on the other it provided an opportunity to develop, for the first time, a model of performance judgement tailored on the peculiar nature of recordings. This, as will be discussed later in this chapter, also has practical repercussions in terms of relevance for the music market.

Furthermore, a major limitation of the research was the narrowing of the focus on the textual content of reviews, without embedding in the analyses evidence from the larger critical discourse such as readers' letters, market data, or aural content of recordings. This decision was taken out of the necessity of generating an understanding of the content of the published reviews before embarking on a larger (and highly resource-consuming) investigation, and in line with the research questions expounded in Chapter 1. This decision allowed indeed a comprehensive and in-depth examination of critics' writings and the development of a detailed

model of performance critical review that would not have been feasible otherwise. Future investigations could then take the findings of this thesis as a starting point for a discourse analysis aimed at interpreting the content of critics' writings from a context-aware perspective. Based on the groundwork offered in this thesis, such investigations could move beyond the relevance of critics' writings for the music performance evaluation discourse to explore the practice of critical review in its socio-cultural matrix. Secondly, concerning the analytical approach adopted in this research, all studies presented were observational, and the approach used was (almost always) inductive. Consequently, the results – although high in validity – cannot be generalised. The aim of this research was to offer a first exploration of the intrinsic potential of music critical review as source of material for research. The inductive approach allowed for a comprehensive map of the content of review, which had not been attempted in any previous study. This model has to be understood as a description of critical review content and not as a normative model. It is the hope of the author that findings of this investigation will lead to hypotheses that will be tested in future research.

Finally, a reflection is required on the kind of conclusions that can and cannot be drawn from an examination of music critical review. An analysis of critical review aimed at generating insights on the way experts listen to and evaluate musical performances works on the assumption that critical texts are the verbal expression of critics' impression of the performance. To what extent and how the actual words relate to what critics experience remains, however, unknown. This limitation, which is common to all forms of self-response research, should be taken into account when interpreting the results.

What is also not possible to say from this kind of investigation is how the judgement developed over time. Given the statement 'Performance P is good because of feature F', we can thus not know if the judgement of goodness was deduced from the presence of feature F (and the underlining criterion that F is a desirable feature in a performance) or if feature F is given as possible explanation of the fact that performance P is instinctively perceived as good. It is true that the choice of professional criticism as material for investigation – also based on findings by recent studies on verbal overshadowing effect (Melcher and Schooler, 1996; Flegal and Anderson, 2008) – minimises the relevance of this distinction for the interpretation of

results. In fact, even if the process of appreciation naturally occurred in a non-inductive way – starting with an immediate impression of the performance, followed by a search for explanations for this impression – it could be expected that expert professional critics like the ones studied in this thesis are able to move backwards from the first impression to the reasons for it in a thorough and reliable way.[27] However, this very fascinating distinction remains nonetheless untackled in this research. An attempt in this direction would need an experimental setting that accounts for the temporal dimension of critical judgements (see Thompson et al., 2007).

Another important point that should be underlined is that reviewing music performance – unlike, say, tasting wine – is a highly selective activity. If by tasting wine it is arguably possible to deliver a rather comprehensive description and evaluation of the tasting experience, in music reviewing what critics do is to select but a few features relevant for their overall impression. As such, reasons critics bring in support of their judgements should be read as partial explanations, and not as sufficient conditions for the goodness of a performance. In the statement 'the fast tempo makes the performance exciting' what is meant is not that the fast tempo alone is sufficient for the performance to be exciting. But rather that the fast tempo contributes to this outcome, in the given musical context and in combination with all the other musical parameters and qualities that occur in the performance (including what happens prior to and after the given passage). That said, that the critic sought out tempo as the feature to highlight in the text, may suggest the salience of this feature in the critic's conceptualization of the listening experience.[28]

Finally, it should be kept in mind that what we can observe in critical review are the recorded performance features that critics discuss and the criteria they use to support their judgements. The present research found a high level of consistency between critics in the relative weight given to one or the other performance feature and criterion. To what extent, however, any single feature represents a common psychological reality for critics remains open. Critics may adduce emotional intensity or understanding as criteria of evaluation. To what extent though one and the same

[27] This is not to imply that the process is unidirectional. Indeed, it may well be that critics move in a variety of ways simultaneously during the listening process.

[28] Perhaps the critic felt the mentioning of this feature as relevant for prospective listeners or important in guiding their listening experience.

performance is perceived as emotionally intense or as expressive of deep understanding by different critics is a question that only a study using an experimental design may attempt to answer. The consistency found in this research thus suggests that there is a set of criteria that are reliably used by critics in their evaluation. However, this does not tell us how reliable the implementation of those criteria is.

In the light of these limitations, an investigation of music critical review may offer important insights that may lead to hypotheses to be tested in future studies, thus complementing the extant literature. It cannot on its own be taken as a definitive account of how critical judgement is or ought to be done. To what extent insights emergent from this kind of investigation are of interest for research and for the musical practice is discussed in the last half of this chapter.

IMPLICATIONS AND FUTURE DIRECTIONS

The question of what makes a good, bad or great performance has attracted the interest of musicians, philosophers and researchers for centuries and great efforts have been made to clarify the phenomena underpinning the listening experience. In particular, much attention has been devoted to the study of written response to music.

The tradition of music written response has a long history. Examples include examination reports and competition rankings, booklet and concert notes, and reviews by professional musicians and critics. These writings have an impact on musicians' lives and careers and offer a direct source of feedback for performers throughout their musical development.

As discussed in Chapter 1, a conspicuous corpus of research in past decades has improved our understanding of one form of music written response commonly used in educational and competitive contexts: the grading of performance, done either holistically or through a segmented, pre-defined scheme. Studies have improved our understanding of associated feedback, of its reliability and consistency and of the different performance elements, including the non-musical, which affect evaluative judgements (McPherson & Schubert, 2004; Kinney, 2009).

By comparison, little is still known about critical review. Critical review that focuses on performance, rather than on the work performed, has been the fashion since the turn of the twentieth century (Monelle, 2002) and is today one of the most

common professional and commercial forms of music written response. Despite the availability of representative material and its impact on musicians' careers, there has been little structured enquiry of the way expert music critics make sense of their experience of performances, and no studies to date have broached the key question of *how* music performance is reviewed by experts.

The research reported in this thesis moved a step in this direction, providing a detailed investigation of a vast corpus of music critical review of recorded performance. Despite the focused nature of the data source (*Gramophone* magazine), musical format (recorded performances) and repertoire (Beethoven's piano sonatas), the corpus was dense in information content and has offered new insights relevant to our understanding of expert performance evaluation and art criticism in general.

Implications for research

Although theoretical and empirical implications have been previously discussed in individual chapters, the main points are summarised and pulled together here. A first point is the relevance of the present findings for current discourses in aesthetics and philosophy of art. Insights from this research, although not directly transferable, have implications for criticism in domains other than music. Results of Chapters 6 to 8, corroborated by observations drawn from the metadata analysis in Chapter 3, offer support to a set of propositions that are at the centre of long lasting discussions on the nature of art criticism:

- Music criticism is essentially evaluative in nature, value judgements being expressed both explicitly and implicitly through value laden performance descriptors (supporting Carroll, 2009);
- The basic criteria intensity, complexity and coherence underpin critics' aesthetic judgements (in line with Beardsley, 1968);
- These criteria are not universally valid, their positive valence depending on the musical context and the balance between/combination of performance qualities (supporting a context-aware generalism like the one defended by Sibley, in Dickie 1987, and Carroll, 2009);
- At a certain level, judgements of value discerning good from less good performances no longer make sense; differences in evaluation become

qualitative and taste-dependent (in line with Levinson's, 2010, discussion and interpretation of Hume, 1757);

- Besides perceptual properties, thoughts about the person behind the performance, his/her intentions, emotions, mental and moral qualities inform and enter critics' judgements (supporting the intentionalist versus empiricist view of art appreciation; Davies, 2006; Graham, 2006);
- In particular, the perception of the recorded performance as the joint outcome of the performer's, composer's and engineers' achievement plays an important part in the final evaluative judgement (supporting Carroll's, 2009, account of 'success value').

Although results of this investigation do not bear normative power (given the observational and explorative approaches employed), they do offer new empirical evidence of the actual content of expert critics' writings, drawn from a representative corpus of critical review. The extent to which the validity of these propositions extends beyond the critical texts examined in this research will have to be explored through the investigation of other review corpuses.

The insights that emerged throughout this thesis also bear implications for empirical music research. Findings from Chapters 3, 5 and 6 suggest that the comparative element in performance evaluation and the use of the term 'expression' are two aspects that will require attention in future performance evaluation studies. As discussed in Chapters 3 and 6, the importance given in critical review to comparisons between performances in the construction of value judgements poses the question of to what extent evaluative judgements can be made, or are actually done, in a criterion based way. Studies so far have usually treated performance assessments as separate items, in which each performance is set against pre-defined criteria in isolation. The weight of the comparative judgement has been accounted for by counterbalancing the order of stimuli and controlling for order-effect (e.g., see Wapnick et al., 1993). However, the use of comparisons in reviews does not emerge as a possible bias, but rather as an important characteristic of the way we listen to and make sense of what we hear. If this is the case, the introduction of the comparative element within evaluation studies – acknowledging it its essential role in listening instead of controlling for it – could lead to more ecological investigations.

Future research should explore how this comparative element could be introduced in controlled studies in a structured and systematic way, as well as its influence on assessment reliability and consistency.

Furthermore, results from Chapter 5 concerning the different meanings that can be attached to the term 'expression' in musical parlance suggest that the use of this term in evaluation studies, if not accompanied by further specifications, may be unclear or even misleading. This is true especially for responses by participants with different levels of musical experience and formal musical training; musicians could in fact more easily tend to embrace a physical, technical notion of expression (value-independent) while non-musically-trained listeners may relate more or exclusively to its psychological notion, closer to the everyday usage of the word and intrinsically value laden. Eventually, the use of alternative terms that specify which component of expression is meant to be observed/discussed could be taken into consideration to avoid overlapping and shifts between the physical and the psychological dimensions.

In addition to specific insights on expression and comparative judgement, the descriptive model developed in this thesis offers a practical tool for music research that can be used to map other forms of performance evaluation, such as examinations, concert reports, or listener descriptions, or to inform the preparation of stimuli in a more experimental setting. Along these lines, results of this investigation will be used by the author in the context of two forthcoming studies, for the design of interview schedules and the preparation of review-like stimuli.

Besides its relevance for philosophy of art and music psychology research, the value of the present work lies first and foremost in it offering new empirical evidence on the content of critical review, as well as first conceptual and methodological grounds for the investigation of further corpuses of this well-established, authoritative and highly relevant form of written music response.

Building on the proposed method and model, further studies should examine other forms, corpuses, and aspects of critical review (historical, cultural), thus further adding to our understanding of this practice. It is the hope of the author that future investigations will also serve to develop the proposed analysis protocol and to test and enhance the present model, investing more resources on the examination of critical judgement narratives. The model developed in this thesis offers mainly an overview of the ingredients present in critical review, with a few insights on the way

these can be used to form complex judgements (e.g., tightrope and tension between evaluation criteria, composite judgements, relationship between single descriptors and valence). From this point on, an analysis focused on the large-picture narrative of reviews could help shed light on other important aspects of the interaction between criteria – for instance, clarifying delicate relationships between global and local judgements.

In sum, the present thesis offers insights and propositions relevant to philosophy of art, music psychology and empirical musicology research. However, through the development of a novel model of critical review content and the delivery of tools generally useful for the examination of written response to music, where this work really offers ground for implications is in the practical domain.

Implications for musical practice

Performance evaluation in general is a complex and often unclear terrain. Expert critics have developed a certain kind of currency of terms they use to navigate on this ground. Teaching students and musicians how critics write and what they look for in performances can help to pass this knowledge on to them, giving them new vocabularies and conceptual tools to be used in their preparation for performance and reflections upon their own practice.

The present work thus benefits conservatoires and music schools, besides being of direct interest for institutions offering targeted programmes in criticism or music journalism, like those currently available at McMaster University, the Juilliard School or Cardiff University. The findings need of course to be read in the context of the critical review practice and with the awareness that, in an educational context, aspects of the performance like craftsmanship and personal development may be given a greater weight than they have in the assessment of professional recordings. However, considerations emerged in this thesis, especially those concerning aesthetic and achievement related criteria, may be used to stimulate discourse on assessment related topics like the importance of comparison, questions of taste, and different kinds of value in performance. In so doing, it will help students, teachers and administrators gain awareness of the factors influencing our perception and appreciation of music and the challenges implied in the evaluation process.

This in turn has a direct relevance for organisations concerned with the continuous development of assessment and examination protocols, such as the Associated Board of the Royal Schools of Music and similar such bodies. In this respect, findings of the research will be used at the author's home institution in the context of a study on assessment processes and criteria.[29] Results from the present investigation will inform the design of semi-structured interviews to be run with teachers at the School of Music, Lucerne University of Applied Sciences and Arts, to better understand how different evaluation criteria are perceived by the teaching staff. In turn, the study aims to promote awareness on the nature and complexity of performance evaluation and to offer insights relevant to the development of refined assessment schemes and policies. In addition, the model proposed here is currently been used at the same institution within a teaching module with third year undergraduates. The module aims to increase awareness of major issues surrounding the musical career among students who are approaching the end of their Bachelors programme. First feedback and outcomes from this module have confirmed the utility of the model in helping students develop a critical attitude towards their own performance practice.

More generally, the findings of this thesis are relevant for the classical music market. Answering Gabrielsson's (2003) call for performance evaluation research to be run outside the boundaries of the educational setting, this work has focused on real world performances, and in particular on musical recordings, offering an empirically developed model of evaluation of recorded performance.

Insights from the model could be of interest to record companies and producers – arguably the first ones to make some value judgement and decision about what should be produced – offering them better comprehension of what critics and, by implication, their readership find valuable and pleasurable. On this point, the research raises questions as to the role of critics as mediators in the classical music scenario. Critics emerged from this investigation as possessing an extraordinarily rich experience in listening to and critically comparing high-level professional performances as well as intimate knowledge of the standard repertoire and the large variety of its interpretation. Along the course of the century, and in a burgeoning

[29] Project financed by the Swiss State Secretariat for Education, Research and Innovation (SERI), 2012-2015.

market, critics offered guidance on what to buy and listen to, polarizing their reviews around a small numbers of interpreters and products and acting *de facto* as filters of choice for consumers.

But to what extent was this intentional? What impact has it had on consumer choices? How has it shaped the establishment of a canon of master performances? And finally, what impact does criticism of recorded performance have today on consumers, in a market that increasingly encourages quick choices based on first-hand experience through low-price electronic downloads (MP3s) and free 30-second music excerpts?

Choices based on first-hand experience may appeal as the most valid; on the other hand, the average listener usually lacks the kind of knowledge of the repertoire and its different interpretations that critics possess, and his/her choices may thus diverge from what critics would recommend. To test this hypothesis, a pilot study was run by the author employing *Gramophone* excerpts collected in this research (Alessandri, Eiholzer and Williamon, 2013). The results suggested that the gap in expertise between the general audience and critics may plausibly lead critics to dislike performances that music students find pleasurable. This in turn poses the question of what selection process – direct experience versus choice mediated by critics' judgements – may offer the highest benefit for consumer choice satisfaction. Together, results from the present research and the aforementioned pilot study suggest that a better understanding of critical review impact on consumer choice would benefit music critics, music media outlets and music consumers, deepening our understanding of the psychology behind review responses and informing a more targeted and effective system of access to music evaluation.

To start addressing these questions, a follow-up of the present research has been planned that investigates the role of critical practice from the psychological perspective of the two key stake-holder groups: critics and music consumers. Through a series of interviews, a large-scale survey, and experiments employing excerpts of *Gramophone* reviews, the project will shed light on the way professional criticism is viewed by critics and consumers, and examine how these views relate to the direct influence of critical judgements on consumers' attitudes towards music

performances.[30] In the context of this project the material collected and model developed in this research will be used in the preparation of textual stimuli to test the influence of valence and style of critical judgements on consumers' attitudes.

CONCLUSIONS

In conclusion, the results presented here depict music critical review as rich source material that has been barely touched upon until now. Insights gained through the investigation of this material bear theoretical and practical implications and provide solid groundwork for the development of future testable hypotheses relevant for training critics, training performers, and furthering our understanding of this field more generally.

Ultimately, upon completion of this challenging journey through recorded performance critical review, what we can learn most is that the judgement of musical performance is far from simple and straightforward. It is complex, contextual and listener specific even for professionals who do this on a regular basis. Current trends in music consumption seem to move reviews towards star and thumbs-up/down rating system. Music critics, however, ask us to take a step back. They do not give us definitive judgements and do not try to simplify the process of listening in this way: three out of five stars does not apply here. They rather create a menu, and discuss how this can suit one or the other listener. Among the reviews analysed in this research, merely seven entailed solely positive judgements and just one was purely negative. But almost half of them entailed comments on taste and preferences, and more than half engaged in qualitative comparisons between interpretations.

In a musical practice often focused on a quantitative notion of value, music critics remind us of its qualitative aspect, its multifaceted nature, and take care to emphasise that, at a certain level, talking of better or worse simply does not make sense. This is a message of hope, suggesting that even in a market that increasingly suffers from paralysis of choice there is still scope for production of other ground breaking performances: if qualitative and not quantitative value is the focus, we are not ever going to hit the top.

[30] Project currently under evaluation by the Swiss National Science Foundation (September 2014).

This also means that no ready-made recommendation may ever satisfy the discerning listener: critics offer guidance to consumers, but they do so by asking them to engage with their judgements, engage with their own previous experiences, knowledge and expectations, reflect and judge themselves to come to embrace the complexity of the musical experience on their own. In a world requiring quick moves and clear-cut decisions, critics' pleading for a deeper engagement with texts and music is indeed challenging, but it is also an honest reflection of the beautiful process of listening to great music.

In the light of this, and with the hope that this work may serve to inspire future investigations to embrace and celebrate the complexity of listening, this thesis is perhaps best ended on the words of *Gramophone* critic Alec Robertson (February 1937, p. 16):

> One might write of the thematic affinity between the three movements and of a hundred other details, but in a review such as this the best thing is to beg people to buy the work, live with it, and make it meaningful to themselves.

References

Ait-Said, E. D., Maquestiaux, F. & Didierjean, A. (2014). Verbal overshadowing of memories for fencing movements is mediated by expertise. *PLOS ONE 9*(2): e89276. doi:10.1371/journal.pone.0089276.

Alessandri, E. (2011). Discography or what analysts of recordings do before analyzing. In C. Emmenegger & O. Senn (Eds.), *Five perspectives on "Body and Soul" and other contributions to music performance studies*. Zurich: Chronos Verlag.

Aschenbrenner, K. (1981). Music criticism: Practice and malpractice. In K. Price (Ed.), *On criticizing music. Five philosophical perspectives*. Baltimore: Johns Hopkins University Press.

Baumann, S. (2001). Intellectualization and Art World development: Film in the United States. *American Sociological Review*, *66*, 404-426.

Beardsley, M. C. (1962). On the generality of critical reason. *Journal of Philosophy*, *59*, 477-486.

Beardsley, M. C. (1965). Intrinsic value. *Philosophy and Phenomenological Research 26*(1), 1-17.

Beardsley, M. C. (1968). The classification of critical reasons. *Journal of Aesthetic Education*, *2*(3), 55-63.

Beardsley, M. C. (1982). The relevance or reasons in art criticism. In M. J. Wreen & D. M. Callen (Eds.), *The Aesthetic Point of View, Selected Essays*. Ithaca: Cornell University Press.

Beardsley, M. C. (1988). The refutation of relativism. *The Journal of Aesthetics and Art Criticism*, *41*, 265-270.

Bergee, M. J. (1997). Relationships among faculty, peer, and self-evaluations of applied performances. *Journal of Research in Music Education*, *45*, 601-612.

Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education*, *51*, 137-150. doi: 10.2307/3345847.

Bergqvist, A. (2010). Why Sibley is not a generalist after all. *British Journal of Aesthetics, 50*(1), 1-14. doi: 10.1093/aesthj/ayp037.

Bonzon, R. (2009). Thick aesthetic concepts. *The Journal of Aesthetics and Art Criticism, 67*, 191-199.

Brandimonte, M. A., Schooler, J. W., & Gabbino, P. (1997). Attenuating verbal overshadowing through color retrieval cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 915-931.

Bresin, R., and Friberg, A. (2013). Evaluation of computer systems for expressive music performance. In A. Kirke & E. R. Miranda (Eds.) *Guide to Computing for Expressive Music Performance*. London: Springer, 181-203.

Budd, M. (1995). *Values of Art: Pictures, Poetry and Music*. London: Allen Lane.

Budd, M. (2007). The intersubjective validity of aesthetic judgements. *British Journal of Aesthetics*, *47*, 333-371. doi: 10.1093/aesthj/aym021

Bujić, B. (n.d.) criticism of music. *The Oxford Companion to Music.* Retrieved May 11, 2012, from <http://www.oxfordmusiconline.com/subscriber/article/opr/t114/e1716>.

Burnham, S. (2000). *Beethoven hero*. New Jersey and West Sussex: Princeton University Press.

Calvocoressi, M. D. (1923). *The principles and methods of musical criticism*. London: Oxford University Press.

Carroll, N. (2009). *On criticism*. New York: Routledge.

Charmaz, K. (1995). Grounded Theory. In J. A. Smith, R. Harré & L. Van Langenhove (Eds.), *Rethinking Methods in Psychology*. London: SAGE Publications.

Clarke, E. F. (1991). Expression and communication in musical performance. In J. Sundberg, L. Nord & R. Carlson (Eds.), *Music, language, speech and brain*. London: Macmillan, 184-193.

Clarke, E. F. (2002). Understanding the psychology of performance. In J. Rink (Ed.), *Musical Performance: A guide to understanding*. Cambridge: Cambridge University Press.

Clarke, E. F. (2007). The impact of recording on listening. *Twentieth-century music*, *4*(1), 47-70.

Cone, E. T. (1981). The authority of music criticism. *Journal of American Musicological Society*, *34*(1), 1-18.

Conolly, O. & Haydar, B. (2003). Aesthetic principles. *British Journal of Aesthetics, 43*(2), 114-125. doi:10.1093/bjaesthetics/43.2.114.

Conrad, W. J., McGill, L., Rosenberg, D., Szántó, A., Young, R., & Chieun, K.-B. (2005). *The classical music critic: A survey of music critics at general-interest and specialized news publications in America. A collaborative project of the Music Critics Association of North America and the National Arts Journalism Program at Columbia University*. Baltimore, Maryland and New York City: Music Critics Association of North America and National Arts Journalism Program, Columbia University.

Cooper, G. & Meyer, L. B. (1960). *The rhythmic structure of music*. Chicago: University of Chicago Press.

Cowart, G. (1981). *The origins of modern musical criticism: French and Italian music, 1600-1750* : UMI Research Press.

Cunningham, H. (2005). Information extraction, automatic. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd edition ed.), 665-677.

Currie, G. (1989). *An ontology of art.* New York: St. Martin's Press.

Currie, G. (1993). Interpretation and objectivity. *Mind, New Series, 102/407*, 413-428.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicinas. *Psychology of Music*, *21*, 103-113.

Davidson, J. W., & Coimbra, D. D. C. (2001). Investigating performance evaluation by assessors of singers in a music college setting. *Musicae Scientiae*, *5*, 33-50.

Davidson, J. W., & Edgar, R. (2003). Gender and race bias in the judgement of Western art music performance. *Music Education Research*, *5*, 169-181.

Davies, D. (2006). Against enlightened empiricism. In M. Kieran (Ed.), *Contemporary debates in aesthetics and the philosophy of art*. Oxford: Blackwell.

Davies, S. (2011). Philosophical perspectives. In P. N. Juslin & J. Sloboda (Eds.), *Music and emotion: Theory and research*. Oxford: Oxford University Press, 23-44.

Debenedetti, S. (2006). The role of media critics in the cultural industries. *International Journal of Arts Management*, *8*, 30-42, doi: 10.2307/41064885.

Dickie, G. (1987). Beardsley, Sibley and critical principles. *The Journal of Aesthetics and Art Criticism*, *46*, 229-237.

Dickie, G. (2000). Art and value. *British Journal of Aesthetics*, *40*, 228-241. doi: 10.1093/bjaesthetics/40.2.228.

Dickie, G. (2004). Reading Sibley. *British Journal of Aesthetics*, *44*, 408-412. doi:10.1093/bjaesthetics/44.4.408.

Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education*, *20*, 268-272.

Dutton, D. (2007, February 26). Shoot the piano player, *The New York Times*. Retrieved from <http://www.nytimes.com/2007/02/26/opinion/26dutton.html>.

Eliashberg, J., & Shugan, S. M. (1997). Films critics: Influencers or predictors? *Journal of Marketing*, *61*, 68-78.

Elkins, J. (2003). *What happened to art criticism?* Chicago: Prickly Paradigm Press.

Elliot, C. A. (1995/6). Race and gender as factors in judgements of musical performance. *Bulletin of the Council for Research in Music Education*, *127*, 50-56.

Ellis, K. (1995). *Music Criticism in Nineteenth-Century France: La Revue et Gazette musicale de Paris, 1834-1880*. New York: Cambridge University Press.

Elste, M. (1989) *Kleines Tonträger Lexikon: von der Walze zur Compact Disc*. Kassel & Basel: Bärenreiter.

Elstein, D. Y., & Hurka, T. (2009). From thick to thin: two moral reduction plans. *Canadian Journal of Philosophy*, *39*, 515-536.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. New York: Cambridge University Press.

Filipello, F. (1956). Factors in the analysis of mass panel wine-preference data. *Food Technology*, *10*, 321-326.

Fischer, E. (1956) *Ludwig van Beethovens Klaviersonaten: Ein Begleiter für Studierende und Liebhaber*. Wiesbaden: Insel-Verlag.

Fiske, H. E. Jr. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education*, *25*, 256-263.

Flegal, K. E., & Anderson, M. C. (2008). Overthinking skilled moto performance: or why those who teach can't do. *Psychonomic Bulletin & Review*, *15*, 927-932.

Flores, R. G. Jr., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking. *Journal of the Royal Statistical Society*, *45*, 97-104.

Flynn, T. (1997). *A study in music criticism and historiography: sacred music journals in France, 1848 to 1879.* Northwerstern University.

Frith, S. (2009). Going critical. Writing about recordings. In N. Cook, E. F. Clarke, D. Leech-Wilkinson & J. Rink (Eds.), *The Cambridge Companion to Recorded Music*. New York: Cambridge University Press, 267-282.

Gabrielsson, A. (1999). Music performance. In D. Deutsch (Ed.), *The psychology of music*. New York: Academic Press, 501-602.

Gabrielsson, A. (2003). Music performance research at the millennium. *Psychology of Music*, *31*, 221-272.

Glejser, H., & Heyndels, B. (2001). Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth music contest. *Journal of Cultural Economics*, *25*, 109-129.

Godlovitch, S. (1998). *Musical Performance*. London and New York: Routledge.

Goldman, A. H. (2005). Beardsley's legacy: The theory of aesthetic value. *The Journal of Aesthetics and Art Criticism*, *63*, 185-189.

Goldstein, A. (1980). Thrills in response to music and other stimuli. *Physiological Psychology, 8*(1), 126-129.

Gorn, G., Pham, M. T., & Sin, L. Y. (2001). When arousal influences ad evaluation and valence does not (and vice versa). *Journal of Consumer Psychology, 11*(1), 43-55. doi: 10.1207/15327660152054030.

Gracyk, T., & Kania, A. (2011). *The Routledge companion to philosophy and music*. New York and London: Routledge.

Graham, G. (2006). Aesthetic empiricism and the challenge of fakes and ready-mades. In M. Kieran (Ed.), *Contemporary Debates in Aesthetics and the Philosophy of Art*. Oxford: Blackwell.

Grant, J. (2010). Metaphor and criticism BSA Prize Essay 2010. *British Journal of Aesthetics*, *51*, 237-257.

Greifeneder, R., Bless, H., & Pham, M. T. (2011). When do people rely on affective and cognitive feelings in judgement? A review. *Personality and Social Psychology Review*, *15*(2), 107-141.

Griffiths, N. K. (2008). The effects of concert dress and physical appearance on perceptions of female solo performers. *Musicae Scientiae*, *12*, 273-290.

Griffiths, N. K. (2010). Posh music should equal posh dress: an investigation into the concert dress and physical appearance of female soloists. *Psychology of Music*, *38*, 159-177.

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences, 108*, 2643-2650. doi: 10.1073/pnas.1018067108.

Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis*. California: SAGE Publications.

Halpern, J. (1992). Effects of historical and analytical teaching approaches on music appreciation. *Journal of Research in Music Education, 40*, 39-46, doi: 10.2307/3345773.

Haskell, H. (1996). *The attentive listener: Three centuries of music criticism*. Princeton: Princeton University Press.

Henderson, W. J. (1915). The function of musical criticism. *The Musical Quarterly*, *1*(1), 69-82.

Hodges, S. D., & Wilson, T. D. (1993). Effects of analyzing reasons on attitude change: The moderating role of attitude accessibility. *Social Cognition, 11*, 353-366.

Holland, B. (1996). Classical view: Colleagues, critique thyselves. *The New York Times.* Retrieved July 12, 2012, from <http://www.nytimes.com/1996/07/21/ARTS/classical-view-colleagues-critique-thyselves.html>.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229-247.

Hopkins, R. (2006). Critical reasoning and critical perception. In M. Kieran & D. Lopes (Eds.), *Knowing Art: Essays in Aesthetics and Epistemology*. Dordrecht: Springer, 137-153.

Hosoda, M., Stone-Romer, E. F., & Coats, G. (2003). The effects of physical attractiveness on job-related outcomes: a meta-analysis of experimental studies. *Personnel Psychology*, *56*, 431-462.

Hu, M., & Liu, B. (2004). *Mining opinion features in customer reviews.* Paper presented at the AAAI Conference on Artificial Intelligence, San Jose, California.

Hu, X., Downie, J. S., West, K., & Ehmann, A. (2005). *Mining music reviews: Promising preliminary results.* Paper presented at the International Conference of Music Information Retrieval ISMIR.

Hume, D. (1757). Of the standard of taste. In *Four Dissertation*. London. Retrieved September 26, 2014 from <http://www.davidhume.org/texts/fd.html>.

Janasik, N., Honkela, T., & Bruun, H. (2008). Text mining in qualitative research: application of an unsupervised learning method. *Organizational Research Methods*, *12*, 436-460.

Jenkins, J. S. (2001). The Mozart effect. *Journal of the Royal Society of Medicine*, *94(4)*, 170-172.

Juslin, P. N. (2003). Five facets of musical expression: a psychologist's perspective on music performance. *Psychology of Music*, *31*, 273-302.

Kaiser, J. (1975). *Beethovens 32 Klaviersonaten und ihre Interpreten*. Frankfurt: Fischer Verlag.

Katz, M. (2004). *Capturing sound: How technology has changed music*. Berkeley and Los Angeles: University of California Press.

Kelly, G. (1955). *The psychology of personal constructs*. New York: Norton.

Kinney, D. W. (2009). Internal consistency of performance evaluations as a function of music expertise and excerpt familiarity. *Journal of Research in Music Education*, *56*, 322-337. doi: 10.1177/0022429408328934.

Kirke, A., and Miranda, E. (2013). An overview of computer systems for expressive music performance. In A. Kirke and E. R. Miranda (Eds.) *Guide to computing for expressive music performance*. London: Springer, 1–47.

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology*, *51*, 336-353.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Leech-Wilkinson, D. (2009a). *The changing sound of music: Approaches to studying recorded musical performances*. Retrieved from <www.charm.kcl.ac.uk/studies/chapters/intro.html>.

Leech-Wilkinson, D. (2009b). Recordings and histories of performance style. In N. Cook, E. F. Clarke, D. Leech-Wilkinson & J. Rink (Eds.), *The Cambridge*

Companion to Recorded Music. Cambridge: Cambridge University Press, 246-262.

Levinson, J. (1987). Evaluating musical performance. *Journal of Aesthetic Education*, *21*(1), 75-88.

Levinson, J. (1996). Performative versus critical interpretation in music, in *The Pleasures of Aesthetics: Philosophical Essays*. New York: Cornell University Press, 60-89.

Levinson, J. (2002). Hume's standard of taste. *Journal of Aesthetics and Art Criticism*, *60*, 227-238.

Levinson, J. (2004). Intrinsic value and the notion of a life. *The Journal of Aesthetics and Art Criticism*, *62*, 319-329.

Levinson, J. (2009). Aesthetic appreciation. *British Journal of Aesthetics*, *49*, 415–425. doi: 10.1093/aesthj/ayp043.

Levinson, J. (2010). Artistic worth and personal taste. *The Journal of Aesthetics and Art Criticism*, *68*, 225-233.

Lindström, E., Juslin, P. N., Bresin, R., & Williamon, A. (2003). "Expressivity comes from within your soul": A questionnaire study of music students' perspectives on expressivity. *Research Studies in Music Education, 20*, 23-47.

Lundqvist, L.-O., Carlsson, F., Hilmersson, P., & Juslin, P. N. (2009). Emotional responses to music: experience, expression, and physiology. *Psychology of Music, 37*, 61-90.

Mantzoukas, S. (2005). The inclusion of bias in reflective and reflexive research: A necessary prerequisite for securing validity. *Journal of Research in Nursing*, *10*, 279-295.

Margulis, E. H. (2010). When program notes don't help: Music descriptions and enjoyment. *Psychology of Music, 38*, 285-302, doi: 10.1177/030573560935192.

Margulis, E. H., Kisida, B., & Greene, P. J. (in press). A knowing ear: The effect of explicit information on children's experience of a musical performance. *Psychology of Music*, doi: 10.1177/0305735613510343.

Matravers, D. (2007). Musical Expressiveness. *Philosophy Compass*, *2*, 373-379.

Maus, F. E., Stanley, G., Ellis, K., Langley, L., Scaife, N., Conati, M., . . . Rothstein, E. (n.d.). Criticism. *Grove Music Online.* Retrieved May 12, 2012, from <http://www.oxfordmusiconline.com/subscriber/article/grove/music/40589>

McColl, S. (1996). *Music Criticism in Vienna, 1896-1897: Critically Moving Forms*. Oxford and New York: Clarendon Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

McPherson, G., & Schubert, E. (2004). Measuring performance enhancement in music. In A. Williamon (Ed.), *Musical Excellence*. New York: Oxford University Press.

McPherson, G. & Thompson, F. W. (1998). Assessing Music Performance: Issues and Influences. *Research Studies in Music Education*, *10*, 12-24.

Melcher, J. M., & Schooler, J. W. (1996). The misremembrance of wines past: verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language*, *35*, 231-245.

Mills, J. (1991). Assessing Musical Performance Musically. *Educational Studies*, *17*, 173-181.

Mitchell, H. F., & MacDonald, R. A. R. (2011). Remembering, recognizing and describing singers' sound identities. *Journal of New Music Research*, *40*, 75-80.

Monelle, R. (2002). The criticism of musical performance. In J. Rink (Ed.), *Musical Performance: A Guide to Understanding*. Cambridge: Cambridge University Press.

Morgan, N. (2010). "A new pleasure": listening to National Gramophonic Society records, 1924-1931. *Musicae Scientiae*, *16*, 139-164.

Morrow, M. S. (1990). Of unity and passion: The aesthetics of concert criticism in early nineteenth-century Vienna. *19th-Century Music*, *13*(3), 193-206.

Morrow, M. S. (1997). *German Music Criticism in the Late Eighteenth Century: Aesthetic issues in instrumental music*. Cambridge: Cambridge University Press.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185-200.

Namey, E. E., Guest, G., Thairu, L., & Johnson, L. (2008). Data reduction techniques for large qualitative data sets. In G. Guest & K. M. MacQueen (Eds.), *Handbook for team-based qualitative research*. Maryland: Altamira Press.

Nelson, P. (1970). Information and consumer behaviour. *Journal of Political Economy, 78*, 311-329.

Newman, E. (1925). *A musical critic's holyday*. New York: Alfred A. Knopf.

Patmore, D. N. C. & Clarke, E. F. (2007). Making and hearing virtual worlds: John Culshaw and the art of record production, *Musicae Scientiae, 11*, 269-293, doi: 10.1177/102986490701100206.

Philip, R. (2004). *Performing music in the age of recording*. New Haven and London: Yale University Press.

Plassmann, H., O'Doherty, J., Shiv, B., & Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences*, *102*, 1050-1054.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport & Exercise Psychology*, *12*(2), 131-144.

Pollard, A. (1998). *Gramophone: The first 75 years*: Gramophone Publications Limited.

Pras, A. & Guastavino, C. (2011). The role of music producers and sound engineers in the current recording context, as perceived by young professionals, *Musicae Scientiae*, *15*, 73-95, doi: 10.1177/1029864910393407.

Regev, M. (1994). Producing artistic value: The case of rock music. *Sociological Quarterly, 35*, 85-102.

Reid, C. (1984). *The music monster: a biography of James William Davison, music critic of The Times of London, 1846-78, with excerpts from his critical writings*. London and New York: Quartet Books.

Rickard, N. S. (2004). Intense emotional responses to music: a test of the physiological arousal hypothesis. *Psychology of Music*, *32*, 371-388.

Robinson, J. (2007). Expression and expressiveness in Art. *Postgraduate Journal of Aesthetics*, *4*(2), 20-41.

Rosenblum, L. D., & Fowler, C. A. (1991). Audiovisual investigation of the loudness-effort effect for speech and nonspeech events. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 976-985.

Rubinstein, R. (Ed.). (2006). *Critical mess: Art critics on the state of their practice* Stockbridge: Hard Press Edition.

Ryan, C., & Costa-Giomi, E. (2004). Attractiveness bias in the evaluation of young pianists' performances. *Journal of Research in Music Education*, *52*, 141-154.

Ryan, C., Wapnick, J., Lacaille, N., & Darrow, A. A. (2006). The effects of various physical characteristics of high-level performers on adjudicators' performance ratings. *Psychology of Music*, *34*, 559-572.

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgements. *Perception & Psychophysics*, *54*, 406-416.

Scheer, J. K., & Ansorge, C. J. (1975). Effects of naturally induced judges' expectations on the ratings of physical performances. *Research Quarterly*, *46*, 463-470.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*, 695-729.

Schick, R. D. (1996). *Classical music criticism: With a chapter on reviewing ethnic music*. New York and London: Routledge.

Schmutz, V., Van Venrooij, A., Janssen, S., & Verboord, M. (2010). Change and continuity in newspaper coverage of popular music since 1955: Evidence from the United States, France, Germany, and the Netherlands. *Popular Music and Society*, *33*, 501-515.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology*, *22*, 36-71.

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception*, *36*, 888-897.

Schwartz, B. (2008). Can there ever be too many flowers blooming? In S. J. Tepper & W. Ivey (Eds.), *Engaging art: The next great transformation of America's cultural life*. London, UK: Routledge.

Sibley, F. (1959). Aesthetic concepts. *The Philosophical Review*, *68*(4), 421-450.

Silveira, J. M., & Diaz, F. M. (2014). The effect of subtitles on listeners' perception of expressivity. *Psychology of Music*, *42*, 233-250, doi: 10.1177/03057 35612463951.

Siegrist, M., & Cousin, M.. (2009). Expectations influence sensory experience in a wine tasting. *Appetite*, *52*(3), 762-765.

Simonton, D. K. (2004). Film awards as indicators of cinematic creativity and achievement: A quantitative comparison of the Oscars and six alternatives. *Creativity Research Journal*, *16*(2 & 3), 163-172.

Sloboda, J. (1991). Music structure and emotional response: Some empirical findings. *Psychology of Music*, *19*, 110-120.

Smith, J. A. (2008). *Qualitative psychology: A practical guide to research methods* (2nd ed.). London: SAGE Publications.

Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance. *Research Studies in Music Education*, *18*, 46-55.

Szántó, A., Simon, J., McGill, L., Janeway, M., Ko, C., Chapman, P., . . . Maisto, M. (2002). *The visual art critic: A survey of art critics at general-interest news publications in America*: National Arts Journalism Program, Columbia University.

Tan, A-W. (1999). *Text mining: the state of the art and the challenges.* Paper presented at the PAKDD Workshop on Knowledge Discovery from Advanced Databases, Beijing.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

Thompson, W. F. (2009). *Music, thought and feeling*. New York: Oxford University Press.

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

Thompson, S. (2007). Determinants of listeners' enjoyment of a performance. *Psychology of Music, 35*, 20-36.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: musical performance assessment as a research tool. *Music Perception*, *21*, 21-41.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Perception*, *25*, 13-29.

Timmers, R. & Sadakata, M. (2014). Training expressive performance by means of visual feedback: existing and potential applications of performance measurements techniques. In D. Fabian, R. Timmers, & E. Schubert (Eds.)

Expressiveness in music performance: Empirical approaches across styles and cultures. Oxford: Oxford University Press, 304-327.

Van den Berg, A. E., Vlek, C. A. G., & Coeterier, J. F. (1998). Group differences in aesthetic evaluation of nature development plans: a multilevel approach. *Journal of Environmental Psychology*, *18*, 141-157.

Van Venrooij, A., & Schmutz, V. (2010). The evaluation of popular music in the United States, Germany and the Netherlands: A comparison of the use of high art and popular aesthetic criteria. *Cultural Sociology*, *4*, 395-421.

Västfjäll, D. (2001-2002). Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientiae*, *Spec Issue 2001-2002*, 173-211.

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

Walker, A. (1968). *An anatomy of musical criticism*. Philadelphia: Chilton Book Company.

Walker, A. (2010). *Hans von Bülow: A life and times*. New York: Oxford University Press.

Wallace, R. (1986). *Beethoven's critics: aesthetic dilemmas and resolutions during the composer's lifetime*. Cambridge: Cambridge University Press.

Walton, K. L. (1970). Categories of Art. *Philosophical Review*, *79*, 334-367.

Walton, K. L. (1988). The presentation and portrayal of sound patterns. In J. Dancy, J. M. E. Moravcesik & C. C. W. Taylor (Eds.), *Human Agency: Language, Duty and Value*. Stanford: Stanford University Press, 237-257.

Wapnick, J., Darrow, A. A., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education*, *45*, 470-479.

Wapnick, J., Flowers, P. J., Alegant, M., & Jasinskas, L. (1993). Consistency in piano performance evaluation. *Journal of Research in Music Education*, *41*, 282-292.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance evaluation. *Journal of Research in Music Education*, *46*, 510-521.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (2000). Effects of performer attractiveness, stage behavior, and dress on evaluation of children's piano performances. *Journal of Research in Music Education*, *48*(4), 323-335.

Widmer, G., & Goebl, W. (2004). Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, *33*, 203-216.

Williamson, V. J., Jilka, S. R., Fry, J., Finkel, S., Müllensiefen, D., & Stewart, L. (2012). How do „earworms“ start? Classifying the everyday circumstances of involuntary musical imagery. *Psychology of Music*, *40*, 259-284. doi: 10.1177/0305735611418553.

Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J., & LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin, 19*, 331-339.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*, 181-192.

Woody, R. H., & McPherson, G. E. (2010). Emotion and motivation in the lives of performers. In P. N. Juslin & J. Sloboda (Eds.), *Handbook of Music and Emotions. Theory, Research, Applications.* Oxford: Oxford University Press.

Ziv, N. & Moran, O. (2006). Human versus computer: The effect of a statement concerning a musical performance’s source on the evaluation of its quality and expressivity. *Empirical Studies of the Arts*, *24*,177-191, doi: 10.2190/E4EN-1X32-KUU1-LDHT.