TIME TO DECIDE: A STUDY OF EVALUATIVE DECISION-MAKING IN MUSIC PERFORMANCE

George Waddell

Thesis submitted for the degree of

Doctor of Philosophy

Royal College of Music, London

July 2018

ABSTRACT

This thesis considers the act of music performance quality evaluation as a performance in itself, examining the processes as well as the products of evaluative decision-making. It provides new understanding of performance evaluation through two experimental studies, two field surveys, and the development of a new mode to study and train evaluative skills. In the first study (Chapter 3), 42 musicians provided continuous quality evaluations of five piano works by Chopin and a twentieth-century composer varying by length and familiarity. Three of these pieces had been manipulated to contain performance errors in the opening material, and two of those the same error at the recapitulation. Results showed that familiarity had no effect within works of a well-known composer, but times to first and final decision were significantly extended for an unfamiliar work of an unfamiliar composer. A shorter piece led to a shorter time to first decision. An error at the beginning of a performance caused a shorter time to first decision and lower initial and final ratings, where the same error at the recapitulation did not have a significant effect on the final judgement, despite causing a temporary negative drop.

In the second study (Chapter 4), 53 musicians and 52 non-musicians gave continuous quality evaluations of one of five randomly assigned videos manipulated to include an inappropriate stage entrance, aural performance error, error with negative facial reaction, or facial reaction alone. Results showed that participants viewing the 'inappropriate' stage entrance made judgements significantly more quickly than those viewing the 'appropriate' entrance. The aural error caused an immediate drop in quality judgements that persisted to a lower final score only when accompanied by the frustrated facial expression from the pianist; the performance error alone caused a temporary drop only in the musicians' ratings, and the negative facial reaction alone caused no reaction regardless of participants' musical experience.

The two survey studies comprised custom questionnaires delivered to large audiences (300 & 433) in live professional settings. The first survey (Chapter 5) examined the relationship between self-reported mood and anxiety states before and after performance with perceived quality and enjoyment of the music. The second (Chapter 6) expanded this to incorporate individuals' perceptions of the social and physical environment. Results from both studies found high correlations of enjoyment and quality ratings, with familiarity with the music not predictive of either outcome. Mood states following the performance were more predictive of judgements than those reported prior. Seat location was not predictive of perceptions, although ratings of the building's acoustic and appropriateness were moderately predictive. Concertgoers assumed their own ratings to be marginally higher than those of their fellow audience members.

Based on the challenges faced in studying performance evaluation in ecologically valid settings, and the parallel difficulties in training the skill of performing evaluations, the principles of Immersive Virtual Environments and distributed simulation are discussed as potential solutions through the proposal of the *Evaluation Simulator* (Chapter 7). All results of the thesis are then discussed concerning their implications for musicians, teachers, and organisations, as well as domains beyond music, in executing and training effective evaluations of human performance. A new research agenda is posited that examines the act of performance evaluation with the same rigour and consideration of complexity given to the performances themselves.

ACKNOWLEDGEMENTS

There are numerous people to whom I owe my thanks and without whom this work would not have been possible. First, I thank my supervisors Aaron Williamon and Rosie Perkins. They have served not only as excellent supervisors but as mentors, coaches, inspirations, colleagues, and friends, and for that I am truly grateful. Thank you both for pushing me to see the bigger picture and for reminding me to consider the details. I also give thanks to Elisabeth Cook and Chia-Jung Tsay for serving on my supervisory team in the early stages and providing invaluable insight.

Second, I thank my Centre for Performance Science family for an engaging and inspiring environment in which to work over the years. In particular, to Tânia Lisboa, Terry Clark, Liliana Araújo, Louise Atkins, Sara Ascenso, Daisy Fancourt, Katey Warren, Sarah Yorke, Lisa Aufegger, Elena Alessandri, Vera Fonte, Pétur Jónasson, and Letizia Gionfrida. Across the wider Royal College of Music, I thank Ingrid Pearson, Natasha Loges, Matt Parkin, Richard Wistreich, Keith Bower, and Erin McHugh.

There are many who generously offered their time and insight along the way. In particular, I thank Reinhard Kopiez, Friedrich Platz, Gary McPherson, Emery Schubert, Evangelos Himonides, Arielle Bonneville-Roussy, and Sam Thompson for their expert advice and feedback on the work. My thanks as well to those who contributed directly to the preparation and development of the studies and collection of the data, including Bruno Bower, Everett Hopfner, Sogol Shirazi, Nicola Pennill, Erika Laiche, David Rees, Victoria Soames Samek, Meurig Bowen, Eric Whitacre (and the Eric Whitacre Singers), and the pianists and study participants who made the research possible.

I thank my parents, Bruce and Janice, for their unquestioning support of whatever I do and for being models of perseverance, hard work, and generosity. Finally, I thank Christine, without whom this thesis would have, quite literally, not been possible. Thank you for believing in me, and for crossing an ocean with me on this adventure. And thank you, Ellie, for being patient.

George Waddell

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

TABLES

Chapter 1

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

LIST OF MULTIMEDIA

Video 4. Aural error only (Aural)

Video 5. Facial reaction only (Facial)

Chapter 7. Recordings for the *Evaluation Simulator* (see Section 7.6.2; the files can be found in Appendix 9).

Sim Video A. Good quality Ravel with *confident* reaction and exit

Sim Video B. Poor quality Ravel with *frustrated* reaction and exit

Sim Video C. Poor quality Tchaikovsky with *distraught* reaction and exit

Sim Video D. Good quality Tchaikovsky

LIST OF PUBLICATIONS

The following publications and notable presentations are based upon the contents of this thesis.

Chapters 1 and 2

Waddell, G. & Williamon, A. (2017). Measuring the audience. In S. Lee (ed.), *Scholarly Research for Musicians* (pp. 148-155). Routledge.

Chapter 3

Waddell, G., Perkins, R., & Williamon, A. (2018). Making an impression: Error location and repertoire features affect performance quality rating processes. *Music Perception*, *36*(1), 60-76.

Waddell, G. (2015). Time to decide: A study of evaluative decision-making in music performance. *Royal College of Music Grove Lecture Series*, London, UK. (Public lecture)

Chapter 4

Waddell, G. & Williamon, A. (2017). Eye of the beholder: Stage entrance behaviour and facial expression affect continuous quality ratings in music performance. *Frontiers in Psychology, 8*(513), 1-14.

Waddell, G. & Williamon, A. (2015). Time to decide: The effects of extra-musical variables on continuous ratings of performance quality. In A. Williamon & M. Miura (eds.), *Abstracts of the International Symposium on Performance Science 2015* (p. 94). Kyoto: Ryukoku University. (Spoken presentation)

Chapters 5 and 6

Waddell, G. (2015). Time to decide: The process of evaluating a musical performance. *University of Cambridge Science and Music Seminar Series*, Cambridge, UK. (Public lecture)

Chapter 7

Waddell, G., Perkins, R., & Williamon, A. (2019). The Evaluation Simulator: A new approach to training music performance assessment. *Frontiers in Psychology, 10*(557), 1-17).

Waddell, G. & Williamon, A. (2017). Time to decide: Designing a simulated evaluation platform. In A. Williamon and P. Jónasson (eds.), *Abstracts of the International Symposium on Performance Science 2017* (pp. 77-78). Reykjavík: Iceland Academy of the Arts. (Winner of the Graduate Poster Award)

Chapter 8

Waddell, G., & Williamon, A. (2018). Time to decide: The art and science of audition evaluation. *Workshop designed and conducted for principals, members, and administrators of the London Philharmonic Orchestra*. London, UK.

1 EVALUATING PERFORMANCE

1.1 INTRODUCTION

Quality judgements are endemic to musical practice. Through formal assessments of performance quality and their resultant grades, placements, rankings, acceptances, and rejections, the career trajectories of aspiring musicians are inevitably shaped. Formal and informal assessments of concerts and recordings then determine their popularity and success. Jurors judge competitors, teachers their students, audiences their entertainers. The very act of experiencing a musical performance in any sense is inherently evaluative in that any exposure and reaction to a musical stimulus will be shaped by the listener's experience, knowledge, culture, taste, emotional state, and physiology (Cross, 2010). In this way, musical evaluation is inherently a subjective practice. This seems obvious when speaking of the more visceral reactions to music; our emotional and physiological reactions at a conscious and subconscious level are well documented (e.g. Egermann et al., 2009a; Grewe et al., 2009; Juslin, 2009; see Chapter 5 for a more detailed discussion). These reactions, however, have been shown to be separable to some degree from one's judgement of a performance's *quality* (Thompson, 2006).

Through an examination of the relevant empirical literature, this chapter outlines the core body of work that has sought to understand and control the reliability, subjectivity, and utility of music performance quality evaluation. Most importantly, subjectivity as a result of the human aspect of evaluation is considered, with focus both on the external variables that influence and confound the evaluator's perceptions, and how their training, experience, and knowledge contribute to their consistency and reliability. Existing models for categorising these variables are overviewed, culminating in the presentation of a new process model of performance evaluation that serves to structure the presentation of the literature and the thesis as a whole. Music performance quality evaluations are then considered as a temporal process in line with the performances they seek to quantify and qualify, and an agenda to consider the act of evaluation with the same temporal specificity is established. Finally, and to facilitate the overarching theme, a statement of the research aims of the thesis follows, focussing on determining the temporal points at which musical decisions are made and how they are affected by extra-musical factors.

Following this first chapter, the second chapter of this thesis presents a summary of the tools used in previous literature to measure performance quality, as well as their development. This includes a survey of traditional written rubrics and of digitally-driven continuous measurement tools, and gives justification for the holistic approach to performance quality assessment chosen for this thesis. Four empirical studies then follow. Chapters 3 and 4 employ lab-based experimental designs with novel recorded audio (Chapter 3) and audiovisual (Chapter 4) stimuli, rated using written and continuous measures to determine the effects of repertoire features, error placement, facial expression, and stage entrances on the evaluation process. Chapters 5 and 6 describe two studies conducted within live choral concert settings in which custom surveys were distributed before and at the interval of the performances to determine how changes in affective state, evaluations of the physical and social environment, and aesthetic judgements interacted with quality ratings. Chapter 7 then considers an existing methodological gap in the study and training of music performance evaluation, and proposes a new tool developed using the principles of Immersive Virtual Environments and distributed simulation. To contextualise each of these studies, Chapters 3 through 7 commence with subject-specific reviews of the relevant literature that expand beyond the core material summarised in the present chapter. Finally, Chapter 8 considers the results across all five studies, their implications for performance evaluation in music and other domains, and outlines a line of future research that considers performance evaluation with the same attention granted to performance itself.

1.2 DEFINING MUSIC PERFORMANCE QUALITY EVALUATION

In defining *music performance quality evaluation,* one must first address the component parts. For the purpose of this thesis, *music* and *performance* refer to the distinction in the classical Western practice in which notated music exists separate to its performance and interpretation. A Beethoven sonata can be an object for evaluation without a performer; its musical and technical qualities, inherent emotional and programmatic content, and compositional quality can all be considered without reference to the interpretation of any one musician. Conversely, the quality of the performance, and by extension the performer, is considered to be separable from the qualities of the work. Were this not true, competitions allowing differing programmes between the competitors would be redundant, as one would simply be endlessly comparing the various qualities of Beethoven's works with those of Chopin's. The performer, then, can be considered separately from the composer in the evaluation. While there are many examples in the Western classical and popular genres (not to mention the myriad non-Western traditions, which go far beyond the scope of this thesis) in which the role of the performer/composer is one in the same, this brings a far greater realm of criteria to the evaluation practice and must be considered separately.

Another distinction that must be acknowledged is that between solo and ensemble performance. Ensemble performance, including that of soloist with their accompanist, comprises a great deal of music making, to the point where true solo performance could be considered the exception to the rule. However, the complexities of separating the performance qualities of one musician from their collaborator, not to mention a group of dozens or hundreds, are many, which may account for academia's preference for individual assessment (Barratt & Moore, 2005). Research has also focussed upon solo assessment, and while it is beginning to approach these challenges of ensemble performance (e.g. Hash, 2012; Harrison et al., 2013; Tsay, 2014) much more remains to be done. This chapter and thesis examine both settings, considering the evaluation of solo pianists in Chapters 3 and 4 and of choral groups in 5 and 6.

Thus, for the purposes of this thesis, music performance quality evaluation is defined by the author as follows: the act of forming a judgement concerning the quality of a musical performance, and by extension the skill of the performer (or performers), considered separately from the merits of the composition itself or one's affective or emotional response to the experience. Whether or not such distinctions between quality and emotional reactions can be made is another matter. Variations in affective responses to musical performances have already been mentioned. Is the quality of a musical performance any more objective? Levinson (1987) described a *perspective relativity of evaluation of performance* (PREP) in which:

> …there is no *single, overriding* point of view concerning performances such that whatever seems good from that point of view qualifies in effect as an absolutely good performance of the work, although there may be a *particular* point of view that is arguably most *central* to evaluative assessment, so that grading of a performance without further specification will naturally be taken to refer to that point of view. (p. 75)

Put another way, it is not enough to qualify how good any performance is; one must also consider who is making the judgement and to what that judgement is being compared – the very definition of subjectivity. Yet, the use of performance evaluation as an objective tool to judge performance quality in both music education and research is widespread (Thompson & Williamon, 2003).

1.2.1 The functions of music performance evaluation

Goolsby (1999) has delineated four categories of music performance evaluation which define the variations in structure and purpose of the act in educational settings: (1) placement, encompassing the auditions that determine a student's acceptance into and placement within an organisation, ensemble, or musical hierarchy; (2) summative, in which the results of a period of learning are demonstrated through a complete performance; (3) diagnostic, used to pinpoint learning and technical deficiencies; and (4) formative, to determine whether development has taken place. These may occur in tandem (e.g. a summative assessment may also carry diagnostic and formative elements) but they highlight the variety of roles that the assessment may take in an educational context. Performance assessment takes a particularly central role in musical practice as, while music-making itself may often be collaborative, the music industry is inherently competitive. Conservatoires, colleges, universities, competitions, agencies, and any organisation that strives to select, train, and/or feature the finest performers all require some degree of ranking one musician over another. Whether an educator or evaluator is able to reliably differentiate and apply these varied forms of evaluation is another matter. In terms of its use as a formative tool, Mitchell, Kenny, and Ryan (2010) found that a panel of expert singing pedagogues evaluating audio recordings, while able to distinguish between first- and third-year performances by the same undergraduate vocalists, could not significantly distinguish between the students' first- and second-year or second- and third-year recordings. This is an unpromising result should institutions wish to demonstrate student improvement more frequently than biennially.

Performance evaluation, as a tool, is also used heavily in the study of musical practice, development, and wellbeing. A researcher examining, for example, the predictive effects of self-efficacy (e.g. McPherson & McCormick, 2006; Ritchie & Williamon, 2012), anxiety levels (e.g. Kokotsaki & Davidson, 2003; Chan, 2010), or practice hours (e.g. Williamon & Valentine, 2000; Bonneville-Roussy & Bouffard, 2015) on performance quality will require some stable, objective measure of the outcome. By doing so, however, critical assumptions as to the stability and reliability of these evaluations are being made. Thompson and Williamon (2003, pp. 23 - 24) summarised these assumptions as follows:

- Musical performance quality is a dimension with a common psychological reality for experienced listeners.
- Experienced musicians are able to offer consistent judgements of music performance quality.
- Experienced musicians are able to distinguish between aspects of a performance such as technique and interpretation.

The first assumption concerns the relativity of performance. It implies that music performance quality can be defined, one performance can be objectively superior to another in terms of that definition, and that the features of that definition are perceivable by an evaluator. It also introduces the topic of expertise, which are discussed below. The second assumption questions the reliability of the evaluator: will a judge give equal ranking to the same performance twice, and are they able to evaluate different performances under the same definition when asked? The third assumption addresses the definition of quality, whether individual components (i.e. criteria: see Chapter 2) contribute to its perception, and whether these components can be perceived and considered separately. With every formal evaluation, in both research and academic contexts, these assumptions are being made. This is problematic for musical practice, as no universal assessment scheme has been identified or widely adopted; rather, each institution employs its own variation, with practices emerging "out of experience, intuition, and tradition, rather than scientific inquiry" (Davidson & Coimbra, 2001: p. 34). Furthermore, the practices used in much research differ little from the schemes used in musical institutions and have not been shown to be any more reliable (Thompson & Williamon, 2003). The fundamental issue then remains concerning the degree to which music performance quality judgements can be considered objective at all. This also begs the question of whether human decision-making in general can be considered objective, for any act of musical evaluation will be fundamentally driven by underlying processes of human cognition, reasoning, and judgement. Thus, the following section will briefly summarise efforts to understand these processes in the wider context.

1.3 FOUNDATIONS OF RESEARCH IN DECISION-MAKING

A full discussion of research in human decision-making would require a survey of cognitive and behavioural psychology the breadth of which is beyond the scope of this thesis. In examining the influence of extra-musical factors on performance quality ratings, literature from related performance and psychological domains are cited as relevant across this thesis, including studies from such applied practices as individual and team sport, medicine, law, and education. However, to give context to the present topic of the subjectivity of human decision-making in presumably objective situations, it is worth briefly touching upon the relatively recent birth of the field of behavioural economics and the pioneering work of Daniel Kahneman, Amon Tversky, and Richard Thaler.

The central theme of these researchers' contributions, conducted since the 1970s and resulting in two Nobel awards, has sought to overturn traditional economic models that assume the people involved are purely rational actors, i.e. that their purchase and exchange decisions are driven by completely objective assessments of value (Kahneman, 2003; Grüne-Yanoff, 2017). This assumption of objectivity has been found to be false, or at least complicated by psychological complexity, time and time again. People will sacrifice financial gain to increase perceived fairness among their community (Kahneman et al., 1986). They value items more if received than if given (Kahneman et al., 1990). They will seek risk to avoid losses, and not take the same risk for commensurate gains (Kahneman & Tversky, 1984; Tversky & Kahneman, 1991, 1992). They prefer maintaining the status quo, even if a slight change could result in a slight gain (Kahneman et al., 1991). They will overpredict rare outcomes if they have recently experienced them (Kahneman & Tversky, 1973; Tversky & Kahneman, 1973). They will take less risk if they evaluate the outcomes of their decisions more often (Thaler et al., 1997). They assume that money is presumed to contribute to life satisfaction and happiness, but those with greater resources do not consistently exhibit these features (Kahneman et al., 2006). And, crucially, biases of decision-making can be found in experts across fields (Kahneman & Tversky, 1977).

These findings represent but a small sample of the contributions of behavioural psychology as they continue to shift how economics, politics, and business are conducted. However, they indicate the general trend towards recognising the inherent subjectivity of judgement. The reign of *Homo economicus* – that mythical, purely rational decision-maker – in economic theory has ended (Thaler, 2016). What, then, of *Homo assessoris*, and the implications for theories of music assessment? The remainder of this chapter, and indeed this thesis, considers the rationality and objectivity of the assessor on whose decisions so much of musical practice is based.

Within the music performance evaluation literature, and due to the inherent ecological richness and complexity of a music performance and assessment thereof, there exists myriad influencing factors and innumerable interactions between them. Due to this complexity, it is unsurprising that attempts have been made to graphically summarise what is known to inform current practice and direct future work. The following section will outline two such ‘process models’ currently available in the literature. It will then outline where these models remain incomplete and can provide challenges in categorising the existing literature. Thus a novel process model is presented that will serve to structure the presentation of existing literature and the organisation of this thesis as a whole.

1.4 PROCESS MODELS OF MUSIC PERFORMANCE EVALUATION

McPherson and Thompson (1998) codified a ‘process model’ of music performance evaluation (see Figure 1.1) based upon work by Landy and Farr (1980) in the domain of management and employment. The model considered the research literature concerning both the creation of evaluative criteria as well as the characteristics of the evaluator and performer. The model was hypothetical, wherein weighting or exclusivity of the displayed relationships was not known, and directions of causality assumed. Thus, it served primarily as a demonstration of the complexity of the subject and a stimulus for future discussion, training, and research.

A simplified model (see Figure 1.2) was formed in 2004 by McPherson and Schubert, updated to reflect new research as previously unconsidered variables within the evaluation process were revealed. Rather than hypothesising the specific relationships between factors, it divided them into the ‘musical’ (e.g. technique, expressiveness, musicality), ‘extra-musical’ (e.g. the performer’s appearance and movement; venue acoustics, familiarity with the repertoire), and ‘non-musical’ (order of performance, race and gender stereotyping). The authors strove for flexibility, admitting that the extra-musical category was an “unclearly defined, fuzzy set” and that “the location of these factors are largely subjective or dependant on circumstances” (p. 65), while positing only gender and race stereotyping and

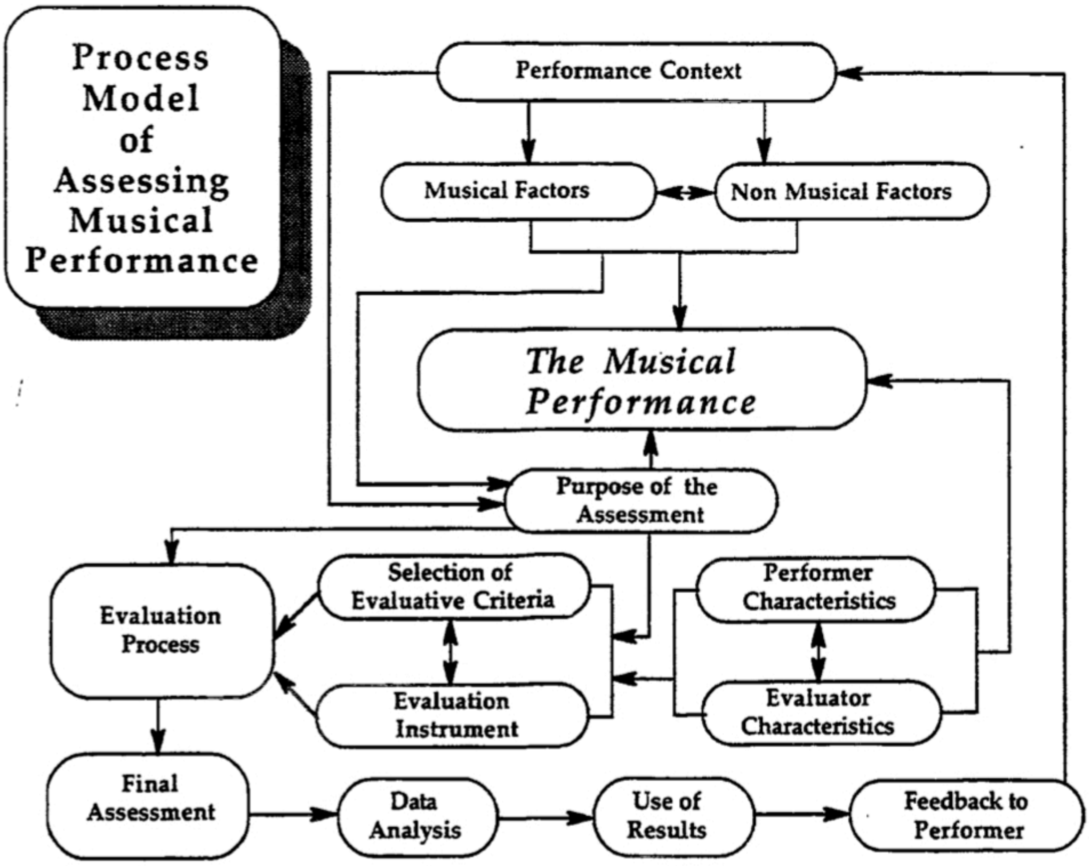


Figure 1.1. McPherson and Thompson's (1998: p.13) process model of music performance evaluation.

performance order as non-musical examples; i.e. variables that *should not* be part of the evaluative process.

This revised model was a useful advancement over the previous in that it did not assume specific interactions, although the ‘fuzzy’ distinction between musical, extra-musical, and non-musical factors complicates its generalisability. By asserting that ‘non-musical’ elements by definition should not be the subject of evaluation and extra-musical should, it implies that an aesthetic choice must be made by those using the scale in determining what qualities are deserving of value in a music judgement. This is complicated also by the artificial distinction between aural and visual

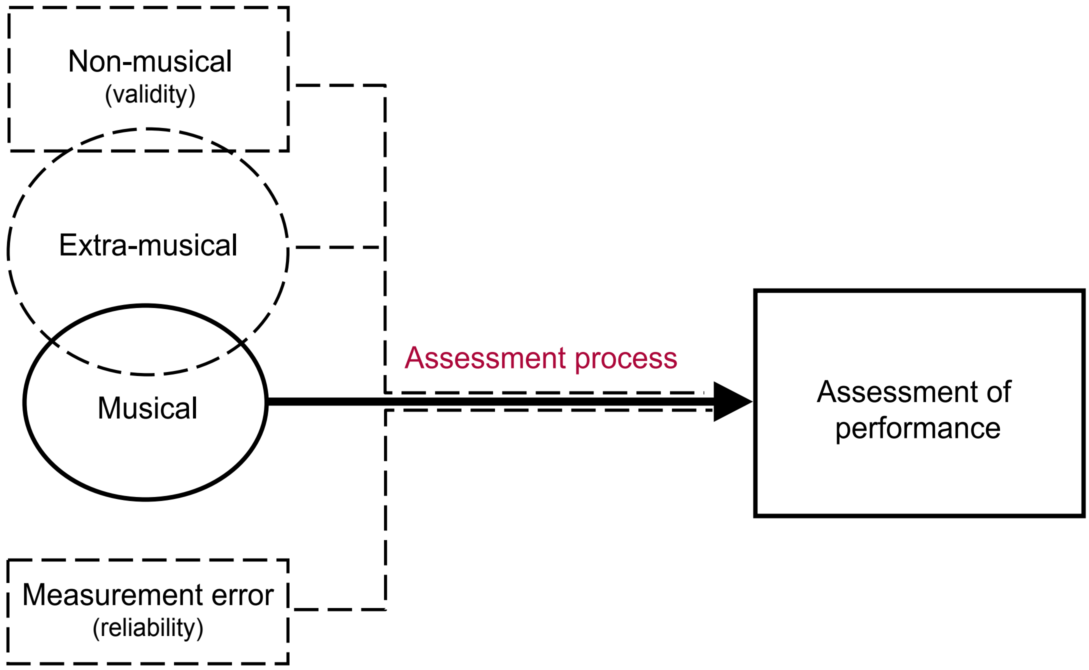


Figure 1.2. McPherson and Schubert's (2004: p. 64) process model of music performance evaluation.

components of music performance, and the degree to which visual information can be classified as ‘musical’. Davidson and Coimbra (2001) found that not only were the physical characteristics of a performer perceived by the evaluators, but that they were actively discussed during the assessment process. Furthermore, students have been found to be aware that their performances are being considered beyond the musical content, particularly concerning repertoire choice, appearance, and behaviour, and such understandings have been shown to broaden with musical education (Kokotsaki et al., 2001).

Whether or not a performance factor is ‘musical’ brings with it considerable complications. Perhaps, then, a different categorisation might be employed, one that does not consider which elements *should* be evaluated, but rather one that delineates those that *are* evaluated. Due to unclear distinctions in the existing performance models, and the immediate need for a tool to organise the consideration and testing of variables for the purpose of this thesis, a new model of music performance evaluation is proposed.

1.4.1 A new process model of music performance evaluation

Two fundamental principles are represented in this new model. The first is based upon the existing concepts of medium, genre, and mode as nested entities that filter the perception of an experience (Thompson et al., 2005). One or more modes (i.e. 'resources of expression'), may be presented as a particular genre (i.e. a 'patterned interaction') via a particular medium (i.e. a channel through which it is expressed). Each category may reciprocally affect the other, in that certain mediums will dictate the nature of possible genres or modes. For example, a collection of text (mode) may be presented within a conversation between friends (genre) as a series of handwritten letters (medium). The same conversation could instead be sent as an email, but this change in medium would alter the experience of both the text and the conversation. Each part influences the whole, so that none can truly be considered in isolation when considering the full phenomenon.

Music performance may then be applied to this framework to form a new process model of music performance evaluation (see Figure 1.3). Musical *repertoire* becomes the resource being expressed (mode), which is presented via a *performer* through the pattern of that individual's particular interpretations and idiosyncrasies (genre), all within a particular *environment* (medium), e.g. a live performance with audience versus a recorded performance versus a closed audition panel. This totality is then processed by the *evaluator*. Taking these four categories, the known influencers of performance assessment can then be grouped within them. Thus, *repertoire* can signify not only the nature of the piece (its length, its genre, its date of composition, etc.) but also its relation to the evaluator (familiarity, likeability, etc.). The repertoire is then filtered through the *performer*, adding to it a particular interpretation unique to that individual. The performer also alters the experience via their nature (appearance, behaviour, etc.) and their relation to the evaluator (e.g. student, unknown, same gender, etc.). This performance is itself situated within the *environment* of both the performer (concert hall, acoustic, purpose of performance, audience size, etc.) and the

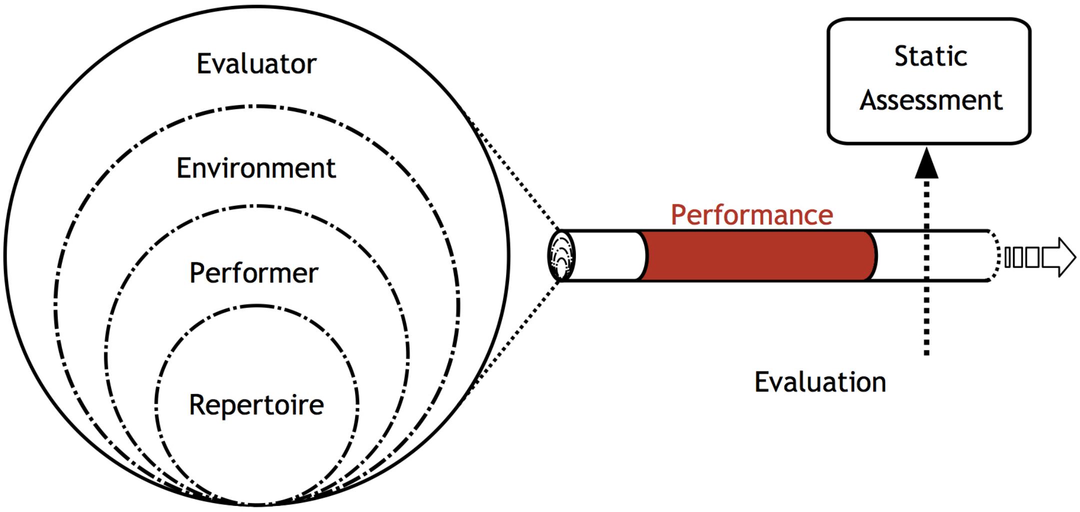


Figure 1.3. A new process model of music performance evaluation.

environment of the evaluator (live versus recording, panel versus solo assessment, place in sequence of performances, time of day, etc.). This information will finally be filtered through the *evaluator*'s own experience, knowledge, expectations, and state into a final judgement.

As in the previous process models, this is a hypothetical representation intended to group existing research into a cohesive conceptualisation of the music performance assessment. The categories of *repertoire*, *performer*, *environment*, and *evaluator* thus inform the selection and organisation of specific variables examined in this thesis.

The second fundamental principle in this model is the framing of evaluation as a *process*. This is adapted from both the 1998 model, in which "Evaluation Process" leading to "Final Assessment" is represented in the lower left portion of the diagram (see Figure 1.1), and more explicitly in the 2004 (see Figure 1.2) model, where the "Assessment process" is given a central role. This is represented in the new model as "Evaluation", with several critical alterations: (1) the processes of evaluation and of performance are presented as taking place *simultaneously*; (2) the process of evaluation is shown to begin before and continue after the musical performance itself;

and (3) the ‘overall’ or ‘final’ assessment becomes a *static assessment,* one that is representative of an evaluator’s self-reported opinion at a specific time during the process of forming their judgement, and indicating that a judgement may continue to evolve after any one report is made. With these distinctions comes the central theme of this thesis: the *temporal* process of music performance evaluation.

This chapter will now summarise the existing literature regarding music performance evaluation and the factors which influence it. Following a review of the earliest work in the field, the category of the *evaluator*, or the expert musical judge, will be first considered due to the foundational nature of the research. The factors of *repertoire*, *performer*, and *environment* will then be used to categorise the relevant literature. Finally, the temporal nature of the evaluation process will be discussed.

1.5 ORIGINS OF RESEARCH IN MUSIC EVALUATION

Research on aesthetic and emotional reactions to music dates back as early as the nineteenth century, notably Gilman’s (1892a, 1892b) descriptions of an ‘experimental concert’ in the fledgling *American Journal of Psychology* in which he provided some of the earliest academic documentation of listeners’ affective responses to a concert performance. Further studies followed and extended this line of questioning, documenting, for example, lists of musical adjectives (Hevner, 1936) and determining listeners’ ability to discriminate between them (Brantley, 1942). During this time, researchers began exploring tools to measure musical ability in areas such as sight reading, singing, aural training, rhythm, and performance (e.g. Seashore, 1939; Wing, 1947; Gutsch, 1964, 1965). The literature concerning tools used to measure performance, ability, and quality are examined further in Chapter 2. For the present discussion, the act of music performance evaluation saw a significant expansion in the Summer 1972 volume of the *Journal of Research in Music Education*. Therein, three articles were published that set the stage for an examination of the evaluator’s role in performance evaluation. These papers addressed three critical topics: the attributes of the performer the evaluator perceives (Moore, 1972), differences in the evaluators’ experience and knowledge (Duerksen, 1972), and the process of developing a scale to overcome these obstacles for the sake of reliability (Schmalstieg, 1972).

Moore (1972) approached performance evaluation from the perspective of communications engineering, in which environmental stimuli (training, the repertoire) are considered the ‘input’ and behaviours (the performance) the ‘output’, thus any difference between the two constitutes the result of an internal act of data processing the observer wishes to understand and classify. She established a hierarchy of musical processes that may be inferred by behaviours (see Table 1.1), each process dependant on the previous to take place. While this did not take into account the evaluator’s own perception and internal processing of these behaviours (wherein the performer’s ‘output’ may become the evaluator’s ‘input’ to continue the analogy of serial data processing) it framed aspects of musical technicality and aptitude as separable from higher-order, less clearly defined aspects of performance such as sensitivity, expression, and artistry. It, in essence, questioned what was being measured in an act of performance assessment, highlighting the complex interaction of technical and artistic elements that contribute to an overall impression of performance; a relationship that much subsequent research has sought to understand.

Duerksen (1972) provided an early quantitative examination of the role of the music evaluator, moving them from a passive recipient of the performance to one where their experience and expectations play an active role in altering their perceptions. He asked whether listeners would rate a performance differently if told that it was by a professional versus a student, and whether music majors and non-musicians would respond differently under these conditions. Music majors (175) and non-majors (264) evaluated a recording of Beethoven’s *Piano Sonata No. 6* performed by Wilhelm Backhaus. The recording was played for each participant twice: once with the knowledge that they were listening to a professional recording by Backhaus, and once after being told that the performance was by a student auditioning for a graduate program. The test order was randomised, with the second presentation (of the same audio material) almost immediately following the first. A control group of 78 students

Table 1.1. Moore's (1972) "Order of Sensory-Dependant Behaviors" (p. 275).

Behavior	*Definition of Behavior*	*Levels of Behavior in Order of Ascending Complexity*
1. Sensation	Behavior demonstrating awareness of informational aspects of the stimulus; detection of change.	Ability to specify the attribute that has changed. Ability to specify direction and degree of change.
2. Figure Perception	Behavior demonstrating awareness of entity; ability to separate a figure from its background	Resolution of detail. Awareness of relationships of parts to each other, to background, and to the whole.
3. Symbol Perception	Behavior demonstrating awareness of form or pattern and ability to arrange discrete information into auditory forms; naming and classifying forms, patterns.	Ability to distinguish tones in a chord. Ability to abstract melody line from its variations.
4. Perception of Meaning	Behavior demonstrating awareness of significance commonly associated with musical patterns; ability to assign personal significance to them.	Ability to reproduce musical patterns by memory. Ability to interpret musical patterns. Ability to complete phrases with musical understanding.
5. Perceptive Performance	Behavior demonstrating ability to make musical decisions in complex situations, to respond to sensory feedback from instrument and audience, and to interpret music with sensitivity and expression. Demonstration of artistry to satisfaction of competent judges.	

were simply told that they would be hearing two separate recordings, then played the Backhaus recording twice (with a visual deception involving the removal and reinsertion of the same tape). Performances were rated on seven-point scales in terms of rhythmic accuracy, pitch accuracy, appropriateness of tempo, appropriateness of accent, dynamic contrast, tone quality, interpretation, and overall quality. Duerksen found that the 'professional' recording was rated as significantly better than that of the ‘student’ across every category, and the experience of the participant (music versus non-music) did not affect their perception of the recordings. Interestingly, the control group also showed significant preference for one recording over the other regardless of experience, although in their case the effect was serial; the second recording was consistently rated as better than the first.

In the third paper, Emily Schmalstieg (1972) drew attention to the requirements of reliability in psychological rating scales, and the lack of such rigour in music evaluation practices beyond basic technical rudiments. She outlined the process of developing a rating scale for the correct production of vowels in vocalists, taking into account basic testing principles: the behaviour was defined (“a homogeneous vowel produced with correct posture, resonance, breathing, and articulation”, p. 281); a ‘degrees of correctness series’ was established taking into consideration an optimal number of differences (five in this case: superior, above average, average, below average, inferior) with concise examples of each; the order of presentation to the judges was randomised; a pilot study was conducted to measure reliability; and final judges were given explicit training on the use of the rating system. While this method only systematised a very specific aspect of one form of musical performance, it provided an early example of the application of psychological rigour to as open-ended and multidimensional a musical concept as ‘correct vocal production’.

These three studies provided a concise overview of the questions subsequent research in music performance quality evaluation has sought to answer: what is in fact being measured, what of the person measuring it, and what of the tools they are being asked to use? The literature concerning the first of these two areas is now considered, beginning with the role of expertise in evaluation, followed by the factors (*repertoire*, *performer*, and *evaluator*) influencing the assessment. The creation, testing, and implementation of rubrics by which performance quality can be measured is addressed in Chapter 2.

1.6 THE 'EXPERT' MUSIC PERFORMANCE EVALUATOR

In addressing the issue of reliability in music performance, the use of an expert assessor would seem to be a logical choice. Surely, if the act of evaluating a performance consists of comparing the presented material to an internal standard built upon experience and knowledge, gathering experts with a shared background in the techniques, styles, and ideals of the genres, instruments, and institutions for which the evaluation is to take place would ensure greater cohesion in the standard for comparison, and thus provide more reliable assessments. This may be especially true in the attempted act of assessing a performance separately from the composition itself. Musical expertise is synonymous with experience, training, and success in the field; a familiarity with the repertoire, techniques, and traditions of the act being judged (Papageorgi et al., 2010). However, one can imagine a hypothetical scenario in which excessive familiarity may inhibit the act of musical appreciation and evaluation. Levinson (1987) describes the case of the 'jaded listener', the one so familiar with the work that all musical "implications and realizations … have been fully absorbed and internalized. For such a listener, a 'standard' performance can verge on sleep-inducing…" (p. 78).

In any case, the 'expert' assessor remains the keystone in the act of formal music performance assessment and understanding their role in the act remains a key goal of research. It is perhaps not surprising, then, that they were the focus of some of the earliest work in the topic. As discussed above, Duerksen (1972) provided an early example with his demonstration that music majors will succumb to false perceptions of aural performance quality based on provided knowledge as readily as non-musicians. Fiske (1975, 1977, 1979) continued this work in a series of studies that examined the reliability of experts as they rated the performances of trumpet students. In the first (1975), he distinguished between *experts*: those who had a great deal of experience in music performance, evaluation, and teaching; and *specialists*: those whose expertise was on the same instrument as the performer they evaluated. The study then examined whether the instrument specialism affected the ratings of 64 recordings of 32 high-school students performing two excerpts in an audition, although the judges (seven-member panels of specialist and non-specialist experts) were informed they were in fact hearing 64 unique musicians. This allowed for an examination of test-retest reliability for each participant. Ratings were collected on a five-point scale across five categories: intonation, rhythm, technique, interpretation, and overall judgement. While inter-judge reliability was shown at a moderately reliable value, there was no significant difference in the mean ratings for any category between the specialists and non-specialists, neither when defining the specialists as trumpet performers or more broadly as wind performers.

A second study (1977) examined this further, taking into consideration the evaluators' own performance ability and music knowledge ('non-performance music achievement'). Thirty-three recent music education graduates rated 40 performances, 20 of which were by unique musicians and 20 of which were test-retest duplicates, unknown to the evaluators. As in the previous study, a five-point, five-item performance scale was used, although 'intonation' was replaced by 'phrasing'. Performance ability and music knowledge were ascertained from the graduates' grades in applied music, music history, and music theory. Judge stability (as an average reliability coefficient between the test-retest comparisons) was found to range from .32 to .82. In contrast to the previous study, a significant difference in group reliability was found when comparing the brass to the non-brass specialist groups. However, these were students with little experience in teaching and adjudicating (i.e. non-expert specialists), thus it was unclear whether that general experience may have reduced that difference over time. Concerning evaluator traits, multiple regression analyses showed no relationship between reliability and performing ability based upon the applied music scores, nor between applied scores and non-performance music achievement (music theory and history grades). There was, however, a significant inverse relationship between non-performance grades and evaluator reliability.

This unexpected final result prompted a third study (Fiske, 1979). Fiske hypothesised that the inverse relationship was the result of a clash between two problem-solving strategies: a "search for the 'right' answer" strategy and a "weighing-comparing" strategy. The former would increase the ability to identify and retain specific information (more useful for academic testing) while the latter would favour a more flexible form of problem-solving that would lead to greater judge reliability. He then theorised that this dichotomy would manifest physiologically as differences between brain hemispheres. This theory was tested with a pair of dichotic listening tests in which listeners were asked to rate two simultaneously presented excerpts that differed only in phrasing or intonation, examining test-retest reliability for each ear. Only low correlations and primarily non-significant results were found.

Winter (1993) examined the role of experience directly, dividing 33 qualified musicians and music educators into four groups of examiners: untrained and inexperienced, trained and experienced, untrained and experienced, and trained and inexperienced. Experience represented previous involvement with music performance assessment, and training referred to a five-part course administered for the purpose of the study that covered the reasons for and practicalities of assessment, as well as the intricacies of the specific scheme to be used. Three videotaped performances were evaluated along 33 statements in five categories (technical, pitch, time, interpretation, overall) on six-point Likert scales. Multivariate analysis of variance revealed that, while both experience and training affected the ratings, training had the stronger influence, although specific results were not presented.

Studies by Bergee (1993, 1997), Hewitt (2005), Thompson (2006), and Kinney (2009) have supported Fiske's findings of an influence of musical training and experience on evaluations, and Thompson and Williamon (2003) found some effect of the instrument played. Musical expertise undoubtedly improves music-specific perceptual skills, with studies demonstrating how musical expertise increased participant's ability to discriminate rhythm (Wallerstedt et al., 2014), to notice subtle differences between short musical phrases (Bugos et al., 2014), to detect timbre-induced pitch shift (Vurma, 2014), and to match composers with complex micro-rhythms associated with their styles (Clynes, 1995). None, however, have found the 'expert' evaluator to be infallible, and recent study continues to find significant variability in raters' internal consistency and the degree to which different evaluators implement the same rating scale (Wesolowski et al., 2016). In many cases, the experience, knowledge, and nature of the evaluators has manifested in much more complex interactions with other variables influencing their perception of the performance. Thus, the literature in recent years has shifted focus to examine these variables and interactions in more detail, often considering as a covariate the role of musical experience. This literature will now be summarised.

1.7 VARIABLES INFLUENCING PERFORMANCE EVALUATION

The literature examined so far in this chapter has primarily considered evaluation to be the judgement of aesthetic, musical, and ‘aural’ aspects of the performance. This approach has been continued in attempts to quantify the purely aural components of music, such as its loudness, pitch, and rhythm (see Chapter 2). The focus upon the purely acoustic qualities of music performance, however, may be an artificial schism, perhaps exacerbated by the advent of recording technology in the late nineteenth century which, for the first time, removed the visual component of what was traditionally a multimodal experience (Thompson et al., 2005). To this day, evaluators consider sound to be more important than the visuals in determining music performance quality, regardless of their experience (Tsay, 2013).

However, unless they are confined to the recording studio, the performing musician is acutely aware that there is more to performance than simply the presentation of an aural stimulus. From the moment a performer walks on stage they are the subject of attention, and this continues to the moment they leave. Until the 1990’s this knowledge was rarely reflected in the research. An early reference took place in Mills’ (1987) first studies on holistic performance evaluation (see Chapter 2), in which she used video-recorded performance examples to “increase the comparability to a live performance” (p. 120).

In terms of musical attributes, expressiveness is now understood to be perceived not only by our expectations of features in aural information (e.g. Woody, 2002) but to contain a strong visual component via the performer’s behaviour (e.g. Juslin et al., 2002). Davidson (1993) provided a first examination of the visual aspects of performance as they contribute to and affect the perception of expressiveness. Drawing from research in human motion, gymnastics, dance, and acting, the study isolated the effects of movement using *point-light technique*, in which reflective tape is placed on the body joints and a spotlight placed adjacent to a camera lens so that only the movement of the tape can be seen. Four solo violinists performed excerpts of their own choice in three conditions: deadpan (little to no expression in the performance), projected (reflecting a standard performance), and exaggerated (overstating the expressive aspects). Twenty-one undergraduate students then evaluated the expressivity of each performance based on the 36 point-light displays (four performers, each playing the three presentation types, each presented as sound only, visual only, and sound with visuals). The study found that participants could not only identify the differences in expressive intension by movement information alone, but rated a stronger difference between the most- and least-expressive performances when presented visual-only information than with audio-only. The audio-video condition ratings were in the middle, indicating that the audio information may have been tempering the reaction to the visual information.

Similar results have been shown in the perception of violin vibrato (Gillespie, 1997), tone duration (Schutz & Lipscomb, 2007), tension, phrasing, and emotional content (Vines et al., 2006), phrasing, dynamics, and rubato (Juchniewicz, 2008), expressive intention and emotional intensity (Chapados & Levitin, 2008; Broughton & Stevens, 2009; Vines et al., 2011; Thompson & Luck, 2012), and musical dissonance, intervallic distance, and emotional valence (Thompson et al., 2005). One can also consider the different way in which aural information is remembered when a visual reinforcement is added. Students have been shown to do better on cognitive memory tasks concerning musical information when the visual component is added (Geringer et al., 1997).

Studies such as these confirm that the perception of music performance is not restricted to the aural sense and purely musical sensibility. With this in mind, what factors, visual or otherwise, might influence one's evaluation of music performance quality? Research to date has found numerous possibilities relating to the repertoire chosen, the performer being judged, and the environment in which the performance and evaluation take place.

1.7.1 The repertoire: Nature, familiarity, and likeability

While studies of performance evaluation often involve comparisons of performers approaching the same piece of repertoire, in naturalistic settings this is not always the case. Jurors in competitions, auditions, and examinations are often asked to rate musicians across a variety of repertoire, differing in genre, difficulty, and length. It is thus assumed that such cross-comparisons are possible, and that differences between the chosen works can be separated from the quality of their interpretation. The literature is discussed to a greater extent in Chapter 3, but for context is briefly summarised here.

Assumptions that assessors can approach varied repertoire without bias have not held up to empirical examinations, as the nature of the repertoire has also been shown to affect evaluator ratings. Glejser and Heyndels (2001) found that pianists and violinists performing in the Queen Elisabeth Competition received higher rankings when playing a more recently composed concerto, and playing a popular work (i.e. one that had been performed often within the competition) correlated with a slightly lower ranking. Research by Wapnick and colleagues (2005, 2009) showed that the tempo (slow versus fast) and duration (20 - 115 seconds) influenced the way in which performance quality was rated, both in terms of reliability and mean score, with relation to the experience of the performer. This is in contrast to the oft-cited work of Vasil (1973) which found that excerpt duration did not affect performance rating reliability and has been used to justify the use of excerpts versus complete performances of repertoire in music evaluation experiments. Wapnick and colleagues (2009) even found that excerpt length mediated the effects of visual attributes of the performer (examples of which are discussed below).

Separate but related to the nature of the work is its relation to the evaluator. Familiarity with the work in question has been shown to have a moderate effect size concerning internal consistency in groups of experienced and non-musicians, although not as strong as the effect of experience in general (Kinney, 2009). Thompson (2007) found that liking of the piece predicted enjoyment and perceived quality more than familiarity. Knowledge at hand may also have an effect. Wapnick and colleagues

(1993) found that providing the evaluator with a musical score did nothing to affect the consistency of their judgements when they were simply asked to provide a preference between two performances, but diminished their consistency when provided in tandem with a rating scheme. Presenting program notes with a structural description of a work has also been shown to reduce enjoyment of the performance in non-musicians (Margulis, 2010). Notes providing information for unfamiliar repertoire have been found to elicit similarly negative receptions (Bennett & Ginsborg, 2018), although research with schoolchildren has demonstrated a positive effect of such information on enjoyment ratings among those for whom the performance was a new experience (Margulis et al., 2015).

1.7.2 Attributes of the performer

1.7.2.1 Gender, race, and nationality

Even the most basic, descriptive qualities of the performer have been shown to influence the evaluation. Race stereotyping is a well-documented issue, with consequences as far-reaching as disparities on opioid prescription (Singhal et al., 2016) and criminal sentencing (Everett & Wojtkiewicz, 2002). Citing research that found gender and race stereotypes associated with specific musical instruments and genres, Elliott (1995) examined whether these features would influence the evaluations of experienced musicians. Four trumpeters (an instrument carrying masculine associations) and four flautists (i.e. feminine associations) were video recorded. Each group comprised a black male, a white male, a black female, and a white female. Separately recorded audio tracks were dubbed over each video to ensure consistent audio quality. Eighty-eight music education majors evaluated the tapes, with the performance order randomised and the participants allowed to delay evaluation until performances of each instrument were viewed.

Race was found to be a significant variable, with white performers scoring significantly higher. The gender/race interaction was significant, with black males scoring lower than black females, and white females scoring lower than white males, as was the instrument/race interaction, with trumpets scoring lower than flutes among black performers and vice versa among white performers. While gender was not significant as a main effect, it interacted not only with race but also with instrument, with female trumpeters scoring significantly lower than female flautists while male trumpeters and flautists received similar scores.

Davidson and Edgar (2003) carried out a follow-up study addressing several of Elliott's (1995) methodological issues concerning participant numbers and controls for instrument-gender associations and performer behaviour. Nine pianists (two each of black males and females, two each white of males and females, one Indian Asian male: a foil whose performances were not included in the analyses) were recorded using both regular video and the point-light technique described above. Each video was then presented normally and with a dubbed audio track. A sound-only condition provided 45 total performances. Thirty-six judges, divided evenly between gender, race (black versus white) and instrument (piano versus other), rated the videos on a seven-point scale of combined artistic and technical merit. Contrary to previous research, the final analyses showed a significant effect of gender (females rated higher) and no main effect of race, although several complex in-group interactions between the gender and race of both the performers and evaluators was noted. Wapnick and colleagues found that female vocalists (1997) and child pianists (2000) were given higher ratings than their male counterpoints, while male violinists (1998) received better grades. An anti-female bias has also been found in the judgement of New Age music when the gender of the composer as told to the participants was manipulated (Colley et al., 2003). Women were also found to give overall higher ratings of the performances.

Context seems to play a role as well. Ensemble performances under black conductors were rated higher than those by their white counterparts when only the visual component of the recording was altered (VanWeelden, 2004). However, the repertoire used in the study was a spiritual, specifically chosen as it was rooted in African-American culture during the time of slavery. Rather than basic stereotypical assumptions, it was hypothesised that the conductor's race simply matched the schema set up by the chosen repertoire. Gender has also been shown to interact strongly with attractiveness and dress of the performer, as described below.

Finally, music competitions have a notorious history of alleged and demonstrated nationalistic bias (McCormick, 2014, 2015). While research is lacking in the musical domain, such trends have been found in evaluations of areas including Olympic diving (Emerson et al., 2000) and ski jumping (Zitzewitz, 2006).

1.7.2.2 Appearance and dress

Attractiveness is a key human bias, found to inflate assessments of the presentation and content of written work (Landy & Sigall, 1974), quality of interviewees (Shahani et al., 1993), and perceptions of social competence, favourable personality traits, and successful life outcomes (Eagly et al., 1991). A series of studies by Wapnick and colleagues (1997, 1998, 2000) examined the influence of the attractiveness and dress of performers on evaluations. In the first (1997), 82 musicians ranging from undergraduate music majors to university music faculty rated 14 performances by unfamiliar singers presenting classical repertoire in formal dress. Participants were divided between audio only, video only, and audio-video conditions. Those in the video-only condition rated attractiveness, while those in the audio only and audio-video groups rated performance quality on both a segmented and a holistic scale. Based on the resulting scores, singers were divided into the more-attractive and less-attractive groups for both male and female performers, with significant differences between each. The results showed that more attractive vocalists were given higher ratings, but, interestingly, more attractive female vocalists were rated as better even in the audio-only condition. This implied that attractiveness, at least in females, had led to genuinely improved performance abilities, possibly as a result of differences in training and treatment. This higher rating of the audio condition did not manifest with undergraduate students; only among the more experienced graduate and faculty evaluators was this shown.

The second study (Wapnick et al., 1998) used a similar methodology, with recorded performances by 12 violinists. Seventy-two participants, either graduate music students or university music faculty members, evaluated the recordings in the same three conditions, although those in the video-only condition now specified the appropriateness of dress and behaviour along with the attractiveness of the performer.

Attractiveness was again shown to influence both the audiovisual as well as the audio-only conditions, with mixed results concerning dress and behaviour. The third study (Wapnick et al., 2000) replicated this procedure with children (20 sixth-grade pianists) and 123 evaluators of varying musical experience. All three visual attributes contributed to better ratings, but as there was no significant overall difference between audio and audiovisual rankings in either gender it could not be ascertained how much was due to an effect of judges' perceptions and how much an effect of training and attention on the performer.

Ryan and Costa-Giomi (2004) examined the effect of attractiveness on adjudication in novice pianists. An altered methodology was used here: all judges rated performers under all three conditions (audio, video, audio and video), reporting both attractiveness and performance quality along several criteria so that repeated-measures comparisons could be made. Attractiveness was found to increase the ratings of females and those whose ratings were already high via the audio-only condition.

Ryan and colleagues (2006) applied the original separate-groups methodology to a naturalistic setting in a study of the Eleventh Van Cliburn International Piano Competition. The authors divided 227 trained evaluators into the three recording conditions, who then evaluated one-minute excerpts of 18 competitors. An interaction between attractiveness and evaluator gender was found: males rated high attractiveness lower and females rated it higher. The performers' dress had a negative main effect on rhythmic accuracy and expressiveness, with male evaluators rating pianists with high dress scores more critically.

These studies by Wapnick, Ryan, and colleagues highlighted the complex interaction of performer attributes revealed through visual information. The contradictions between studies are apparent, and may be due to the varying age and experience of the performers in each study or the complex interaction of appearance, dress, behaviour, gender, and evaluation condition. Individual components of performer appearance have been picked up in other research. In a series of studies by Griffiths (2008, 2010, 2011) the clothing worn by female soloists (ranging from jeans to formal concert attire) was shown to significantly influence raters' perceptions of performance quality in the evaluation of dubbed video recordings. Both evaluators and performers were shown to have pre-existing conceptions of what comprised appropriate attire, and performers were highly aware of how clothing choices reflected the persona and qualities they wished to impart on stage.

Closely related to the appearance of the performer is the equipment with which they use to perform, and the information it visually connotes. Williamon (1999) demonstrated that an audience's preference for the musical qualities of a memorised performance can manipulated by the simple placement of an empty music stand in front of the performer. Evaluators preferred an apparently memorised performance. Furthermore, the partial visual obstruction of the performer (a cellist) by the stand seemed to inhibit the evaluators' ability to pick up on communicative and expressive aspects of the performance, leading to lower ratings of the performances in any situation where a stand was placed in front of the performer. A 2017 study by Kopiez and colleagues, in which the performance quality between presentation modes was controlled via the use of audio/video juxtapositions, replicated the finding with a small but significant effect, also finding no effect of the evaluators' musical experience.

1.7.2.3 Stage behaviour

The role of visually-conveyed stage behaviour in affecting musical and expressive perception has already been discussed (see Section 1.6 above). It has also been shown to influence performance evaluation on a more fundamental level. In vocalists, bodily communication, and specifically eye contact and facial expression, was found to prominently enter the discussion among evaluators in higher education settings (Davidson & Coimbra, 2001). The same research found that the manner in which vocalists addressed the audience between works, and specifically via repertoire introductions, altered their perception of the performer. This combined with behaviour during the performances and the musical quality to form an overall 'personality' judgement of the performer (e.g. charming, sweet, engaging). This effect may be emphasised in vocalists as compared with instrumentalists, as "the singing performance, possibly more than any other performance, involves a direct relationship between the performer and the audience" (Kokotsaki et al., 2001: p. 15). Thus, the presentation of such a ‘personality’ in vocalists may serve a more direct connection with the music being performed.

Rodger, Craig, and O’Modhrain (2012) demonstrated that both musicians and non-musicians could perceive a clarinettist’s experience level from point-light behavioural information alone, and that the ratings of novice performers’ audio recordings could be improved by superimposing the point-light videos of expert performers. The effect was not reversible, however: superimposing novice videos on expert performances did not diminish their ratings, implying that visual information may only enhance the perception of an aural stimulus.

The effects of expressive behaviour have not always been found to be positive, however. Ryan and colleagues’ (2006) study of the Van Cliburn International Piano Competition found that ‘high-behaviour’ pianists were given lower ratings than ‘low-behaviour’ performers. Interestingly, note accuracy was rated especially low for ‘high-behaviour’ pianists in the audio-only condition, suggesting a possible trade-off of control for expressiveness.

Recent studies have demonstrated the strong influence of the visual component, including how a conductor’s expressivity affects quality ratings of their ensembles (Morrison et al., 2012, 2014; Price et al., 2016). Huang and Krumhansl (2011) found that audiovisual recordings in which a pianist was asked to perform with restricted motions were rated significantly lower than performances of the same works with ‘natural’ movements, with experienced musicians also preferring the latter performances in audio-only settings. Performances with exaggerated stage behaviour did not lead to any further significant increase, and in fact caused a significant drop among performances of a work by Copland (but not in Chopin or Bach) suggesting an influence of repertoire. Increased stage behaviour has also been found to increase ratings of rock guitar performances (Lehmann & Kopiez, 2013). Platz and Kopiez (2013) found that the quality of the stage entrance, tuning, and preparation up to the moment of sound production among violinists in an international competition significantly altered viewers’ wish to continue observing the performance (see Chapter 4 for further discussion of the effect of stage entrances). Tsay (2013) found that, while novices or experts could not reliably predict the winner of an international piano competition based upon audio or audio-video recordings, silent video recordings demonstrating only the musician's physical behaviour allowed the winner to be chosen at a rate greater than chance regardless of the evaluator's experience. Griffiths and Reay (2018) juxtaposed the aural and visual recordings of professional and amateur performers, finding the visual modality drove judgements, again without moderation by the evaluators' experience.

1.7.2.4 Assumed ability

Just as consumer price evaluations can be driven by comparison to higher- and lower-value products (Herr, 1989), research has demonstrated how priming a judge with false information regarding the performer's ability can affect ratings. Duerksen (1972) and Radocy (1976) found that recordings received higher rankings when listeners were told they were played by professionals rather than students. A similar effect was found in the assessment of wind band performances (Silvey, 2009). Such information relates closely to the social environment in which evaluations take place, for it implies that the conclusion of a fellow judge might be the cause of shifts in assessment.

1.7.3 The evaluative environment

Performances evaluations take place in a physical and social setting, whether listening alone to a recording, within a panel of judges, or among a live audience in an auditorium. The research examining these environmental features is discussed in more detail in Chapter 6 and briefly summarised here.

1.7.3.1 Social aspects

Despite the great deal of research in the social sciences concerning group pressures on evaluation and decision-making (Aronson et al., 2007) and the prevalence of group decisions in music performance evaluation (e.g. audiences, audition panels), little has been done to examine the social interactions inherent to the practice. Radocy (1976), in addition to demonstrating an effect of presumed performer status on evaluations, also found that assumed knowledge of how previous assessors rated a performance shifted participants' judgements. Davidson and Coimbra's 2001 study provided rare insight into the discussions of academic evaluation panels to reveal examples of implicit criteria (including appearance and behaviour) used by assessors that do not appear in written reports. This work also highlighted the role of student-teacher relationships when a professor assesses their own student, and how examiners may temper their marks in light of a wider knowledge of their students' academic standing and how the mark may affect their development. The relationships between and roles of teacher-, peer-, and self-assessment in pedagogical settings are examined further in Chapter 7.

The role of the audience member within the group has been considered, with research exploring how applause, one of the basest forms of expressing a performance evaluation, may be mathematically modelled as a form of social contagion (Mann et al., 2013). Length of applause was found to be highly variable even within the presentation of identical stimuli. One study by Springer and Schlegel (2016) has examined this effect in musical settings, finding that high-magnitude applause appended to the end of recorded performances increased evaluative ratings of a march but decreased them for a ballad.

The sport psychology literature has repeatedly demonstrated bias for the home team, with increased audience support and its influence on referees' judgements found to be stronger factors than practical advantages such as reduced travel when playing in one's home stadium (Nevill & Holder, 1999; Clarke & Norman, 1995; Garicano et al., 2005; Dohmen, 2008), although there is evidence that familiarity with the venue plays a role (Wilkinson & Pollard, 2006). The effect of biased judges must be balanced with documented cases where accusations of unfair judgements turn out to be non-existent, a form of confirmation bias in which the calls against one's team are remembered more strongly than the calls for it (Rodenberg, 2011). Research has also examined emotional contagion in audience members, hypothesising that affective reactions may be heightened in group settings (Sutherland et al., 2009), and when the reactions of other evaluators are made known (Egermann et al., 2009a), although results have been contradictory.

1.7.3.2 Practical aspects

The manner in which performance materials are presented may also alter the evaluation. Live performance carries with it variations in the physical space, and qualities of and familiarity with the venue have been shown to influence a listener's anticipation and enjoyment of a concert experience (Thompson, 2006, 2007). Video recordings, no matter their quality, change the mode of presentation and to some degree inhibit the visual characteristics of the performance (Thompson et al., 2005). Audio recordings remove what has been shown to be a vital visual component, which may result in generally lower ratings overall (Wapnick et al., 1997). Timmers (2007) found a correlation between quality rating and the recency of classical vocal recordings, which corresponded with an increase in recording quality, although causal connections could not be drawn as the change in performance practice and interpretation between recordings was not controlled.

While the present thesis focusses on the evaluation of solo performances, a great number of soloists in the western classical tradition perform with some form of accompaniment, whether in audition or concert settings and whether it be with a pianist, chamber group, or orchestra. Assuming these collaborations are performed live (i.e. not pre-recorded audio), they complicate the evaluative environment in that they introduce the complexity of a second layer interpretation, fallibility, and social influence into the situation without themselves being a focus of the evaluation. This on top of the explicit effect accompanists have on the performer's decisions and reactions in the performance and their preparation of the work with or without accompaniment, which has been found to lead to more highly-rated performances (Klee, 1999). Britten (2002) had 188 young instrumental musicians listen to performances of material with no accompaniment, piano accompaniment, or with pre-recorded accompaniment. Performances played with recorded accompaniment received the highest quality ratings, while performances with piano received the lowest. This preference for recorded accompaniments was replicated in a subsequent study with a sample in Singapore (Britten et al., 2002), which also replicated a correlation between listeners' preferences for accompaniment modality and the quality rating of the solo performance it supported.

Several studies have examined how the presence of accompaniment affects perception of expressive features of the collaborative performance. Geiringer and Madsen (1998) demonstrated that evaluators gave significantly higher expressiveness scores for use of phrasing, expression, rhythm, and dynamics when excerpts of Schubert's and Gounod's *Ave Maria* were presented with separately-recorded accompaniments versus without. Geringer and Sasanfar (2013) found that the level of the accompanist's expressivity affected ratings of such for the soloists. Springer and Silvey (2018) found comparable effects, including significant changes of ratings for both accuracy and expressivity of the soloist.

Procedural aspects of evaluation may also affect their results. Duerksen (1972) first demonstrated serial effects in evaluation in that an immediate preference for the second hearing of an audio recording manifests in laboratory conditions. A recent study (Anglada-Tort & Müllensiefen, 2017) replicated and expanded the underlying finding that listeners are often unable to recognise the same performance presented twice, dubbing the effect the *repeated recording illusion.* Three quarters of listeners in the authors' sample believed they had heard a different performance when the same was presented, whether or not confounding information was present, with those demonstrating higher neuroticism and openness to new experience being significantly more likely to commit the error. Flôres and Ginsburgh (1996) demonstrated an extension of this effect in a naturalistic setting. In a landmark study, they examined whether the final ranking of performers in the prestigious Queen Elisabeth Competition correlated with the day on which the candidate performed. The rankings of the 12 semi-finalists over 21 competitions (from 1951 to 1993; 120 violinists and 132 pianists) were aggregated. As performance order of the 12 performers (two per day over six days) was randomly chosen, the null hypothesis stated that each permutation of rankings over the 12 performance slots was equally likely. This was, however, not the case. Candidates performing later in the week were more likely to receive a higher ranking, with the peak occurring on day five of six and the lowest point on day one. The effect was more strongly pronounced for the pianists than the violinists. Suggested causes were a learning effect of the judges, both in formalising their internal rating schemes and developing familiarity with the imposed concerto (composed specifically for the competition and not yet heard by any of the jurors). A later study of the same competition (Glejser & Heyndels, 2001) supported these results. The serial effect has also been well documented in popular music domains such as the *Idol* series (Page & Page, 2010) and the Eurovision Song Context (Bruine de Bruin, 2005) and other evaluative domains including figure skating (Bruine de Bruin & Keren, 2003; Bruine de Bruin 2005, 2006), synchronised swimming (Wilson, 1977), Olympic gymnastics (Damisch et al., 2006), the grading of essay papers (Hales & Tokar, 1975), and in forming preferences for consumer products (Moore, 1999).

Time of day has also been found to influence ratings, with scheduling later in the day predicting higher ratings in a several studies of solo and small-ensemble music festivals (Bergee & McWhirter, 2005; Bergee & Westfall, 2005; Bergee, 2006). Conversely, a study by Elliott and colleagues (2000) found that students performing in a morning session of auditions for an all-state band were more likely to be chosen than those performing in an afternoon session, contradicting the standard assumptions of the serial effect. Examinations in other domains have demonstrated this effect, notability in judicial decisions where favourable parole outcomes have been found to be highest after meal breaks, steadily declining over time until the next respite (Danzinger et al., 2011).

The research discussed thus far, focussing on the repertoire, performer, environment, and expert evaluator, has focused almost exclusively on the singular product of the performance evaluation, examining a set of ratings taken at a point following the conclusion of a performance. However, as discussed in the development of the new process model (see Section 1.4) and in following the overarching theme of this thesis, the temporal process leading to these evaluative products is key to understanding the act of evaluation. Thus, the following section examines this product, first as it relates to the performance itself, then to the judgement of that performance.

1.8 MUSIC AS A TEMPORAL PROCESS

1.8.1 Process versus product

A process describes the occurrence of one or more acts over time. The growth of a sapling, the learning of a language, the fall of a rock, and the construction of a bridge all serve as examples. The result of a process is the product: the tree, the skill, the crater, the bridge. Differing processes may lead to the same product, with variations in the time taken and order of individual acts. Thus, the process of constructing a house may take a month or a year, or the order of installing the windows and doors may be reversed, and yet the same product—the house—results. This difference is reflected in art. The process of creating a painting or sculpture leads to a product, one that becomes free of the temporal aspect. It may be experienced, in theory, over any length of time, its component parts introduced, studied, and evaluated in any combination or sequence. What, then, of music? What is its product? The notated score may be the product of the compositional process, but from there it forms the starting point and guide for performance. It is a document with which an audience or evaluator may never interact. Performed interpretations of the score take place continuously over time, making music (along with other performance-based art forms such as film and dance) "an art that is based on the temporal stream" (Namba et al., 1991, p. 270). Can the performance, then, be considered the product?

Not necessarily. Music performance itself is a process. There is first the process of the performer producing the sounds; i.e. the cognitive and physical acts of the performer. The processes of performance has received considerable attention in the research literature, including the routes by which a performer develops the relevant skills (e.g. Ritchie & Williamon, 2012; Clark & Lisboa, 2013; Bonneville-Roussy & Bouffard, 2015; Hatfield et al., 2016), memorizes and recalls their repertoire (Highben & Palmer, 2004; Chaffin et al., 2010; Lisboa et al., 2014), maintains and has affected their physical and mental wellbeing (e.g. Spahn et al., 2004; Araújo et al., 2017), experiences anxiety and arousal before and during a performance (e.g. Kenny, 2011; Chanwimalueang et al., 2017), and executes thoughts and actions during the performative act itself (e.g. Mishra, 2010; Maidhof et al., 2013; Clark et al., 2014).

There is then the process by which those sounds are transmitted to the listener, involving the physical and acoustic properties of the relevant instrument, venue, recording and playback device, etc, not to mention the psychoacoustic principles by which those sounds are received and processed by the human ear. Then, the act of listening to these produced sounds results in a process, as it must too inevitably take place over time. The listener actively responds based on their experience and cross-modal perceptions, creating a continuous *perception-action* cycle (Cross, 2010) in which they create and alter their perception of the performance as it unfolds. An audio or video recording, then, may represent the product, but this is an abstract state, for the only method of accessing the captured information without changing its very nature (such as a graphic or statistical representation of its digital or acoustic qualities, for example) is by reliving it via playback. Recordings may allow for an alteration of the process, i.e. the rewinding, replaying, and skipping forward through the timeline, not to mention the varying possibilities of audio and video manipulation, but such alterations inevitably affect the nature of the performance itself. A situation where one could ask the pianist to stop and replay a section, for example, would be considered fundamentally different from a standard performance presented continuously from start to finish. It is clear, then, that music performance can be considered and studied as an ongoing process. Thus, we can now examine the degree to which the perception of those performances has been examined with regard to this temporal unfolding.

1.8.2 Studies of temporal reactions to music

The act of perceiving a performance over time has been explored primarily in listeners' emotional and affective reactions to music as it unfolds. Such examinations allow researchers to attribute correlations to specific features of the music as they are defined by their temporal location, as opposed to relying on *post hoc*, self-reported assessments provided by the listener *following* the experience of a performance.

Methods of *continuous measures* have been developed and used to examine these relations. These have included bespoke technologies and devices, such as the Continuous Response Digital Interface (CRDI; Madsen, 1990, 2011; Gregory, 1995; Geringer et al., 2004) and the Continuous Response Measurement Apparatus

(CReMA; Himonides, 2011, 2017). MIDI devices, normally used for the performance of and interaction with musical stimuli, have also been employed (e.g. Vines et al., 2006), as well as bespoke computer software (e.g. Egermann et al., 2009b). These tools have been used to examine listeners' preferences (Brittin & Sheldon, 1995), perceptions of loudness (Geringer, 1995) and phrasing (Vines et al., 2006), focus of attention (Geringer & Madsen, 1995, 1998; Madsen, 1997; Madsen & Geringer, 1999; Madsen & Coggiola, 2001; Williams, Fredrickson, & Atkinson, 2011), perception of musical intensity (Brittin & Duke, 1997), perceived tension (Madsen, 1998; Vines et al., 2006; Williams et al., 2011), perceived expressivity (Silveira & Diaz, 2014), and aesthetic (Madsen, 1997; Madsen & Coggiola, 2001; Geringer & Madsen, 2003) and emotional (Madsen, 1998; Schubert, 1999, 2004; McAdams et al., 2004; Plack, 2006; Egermann et al., 2009b) responses in relation to musical stimuli as they change over time, often comparing them to overall ratings and correlating them with physiological responses (e.g. Himonides, 2011). The temporal relation of neural activity as it relates to music perception has also been explored, using, for example, electroencephalography (EEG; e.g. Palmer et al., 2009) and functional magnetic resonance imaging (fMRI; e.g. Chapin et al., 2010). The particular use of these methodologies of *continuous measurement* are explored in more detail in Chapter 2, for they form the basis of two of the studies in this thesis (Chapters 3 and 4) examining a temporal process as yet underexplored: that of music performance quality evaluation.

1.9 MUSIC PERFORMANCE QUALITY AS A TEMPORAL PROCESS

In very few cases have continuous measures methodologies been applied to music quality evaluations. Geringer and Madsen (1998; Madsen & Geringer, 1999) approached this topic in a study of attentional focus in performance pre-judged as good or bad. Himonides (2011) conducted a pilot study examining quality ratings of sung vocal performances, including criteria such as diction, dynamics, and vibrato, and comparing their responses to physiological data (heart pulse, rate, and galvanic skin response). The data collected demonstrated the potential for continuous self-reported and physiological measures to provide insight into the complete experience of music perception and judgement.

A direct application of continuous measures methodologies in the performance quality evaluation literature was carried out by Thompson and colleagues (Thompson, 2005; Thompson et al., 2007) in which a bespoke piece of software was created to allow for continuous data to be collected using a simple mouse interface. Two pianists each audio-recorded contrasting performances (slow, natural, and fast) of two works, resulting in a total of 10 performances (one pianist's fast recordings were discounted as unrealistic). Thirty-three active musicians, teachers, and music researchers were then divided into three experimental groups which evaluated each performance's overall quality, technical proficiency, or musicality both continuously using the software and as an overall judgement using written scales following the performance.

Analyses of the resultant data provided five general results: (1) initial evaluative judgements were reached approximately 15-20 seconds into the performance, and final judgements were reached approximately 60 seconds into the two-minute excerpts; (2) judgements changed an average of 2.6 times per minute, with frequency and magnitude of changes decreasing over time; (3) written overall evaluations were closer to the final continuous judgements than the initial ones, suggesting the evolution of a final decision as opposed to an averaging of individual judgements; (4) while initial time to form a judgement and frequency of changes did not significantly differ between groups, the direction and overall pattern of changes did differ between those evaluating technical, musical, and overall quality; and (5) individual consistency in the characteristics of the judgement process between recordings was low.

Save for basic differences in the judgement criteria, no other variables within the evaluation process were examined in the study. Also, specific temporal features of the recording were not examined for correlation with judgement decisions. Rather, the authors sought to establish general patterns in the evaluation process based on recordings that were stylistically distinct. These patterns were found in the average amount of time taken to form initial and final judgements, as well as general tendencies in judgement alterations as their decisions evolved. However, the variability between the evaluations of individual judges as they assessed various recordings remains promising as an avenue for research; it implies that the temporal processes of performance quality evaluation are not uniform, but instead vary depending on the particular circumstance of the piece and perceptions of the evaluator.

While little research has been conducted examining the specific process of music performance evaluation, that which exists hints at a rich resource of untapped knowledge regarding how performances are quantified, qualified, and ultimately valued. Every factor presented in the literature review, from the performer's appearance to the evaluator's experience, is currently understood only in the context of the final rating, with nothing known regarding the temporal points at which such factors are most meaningful or how they interact over time to form the final judgement. Such knowledge is vital in understanding more fully the nature of performance itself and to better quantify the outcomes of performance-enhancing interventions. It is also centrally important to the practice of performance itself, wherein musicians not only find their performances under intense scrutiny with lasting consequence on their development and careers, but must often conduct (and learn to conduct) evaluations of others as part of their career. Thus, the present thesis examines this act of music performance quality evaluation, considering not only the products of evaluation but the processes leading to them. In this way, a performance evaluation can be treated with the same complexity as the performance under evaluation, opening new opportunities for research in line with the breadth and depth of attention given to the study of performance processes as described above. To achieve this research agenda, the specific aims of this thesis now follow.

1.10 AIMS OF THE PRESENT THESIS

Any assumptions that music performance quality evaluations are purely objective have been shown to be unfounded; like the art form they seek to capture, such evaluations are subject to all of the complexities of human perception, expression, and behaviour of the performer and assessor and the environments in which the judgements take place. Furthermore, an evaluation is a process that coincides with the experience of the performance, an experience that does not necessarily begin and end with the first and final notes of the work itself. A complete understanding of performance evaluation requires engagement with the various factors at the points in time at which they occur.

This thesis contributes to such understanding by examining the process of musical decision-making as it unfolds over time. To achieve this, the first overarching research question (RQ) begins with the existing framework set out by Thompson and colleagues (Thompson, 2005; Thompson et al., 2007):

RQ1. When are decisions made and adjusted while assessing the quality of a musical performance?

Following on this underpinning, the thesis examines a selection of variables for their effects on and relationships with the formation of performance quality evaluations. The selection of these variables is guided by the novel process model presented above (see Figure 1.3). Thus, the remaining aims are to determine:

RQ2. How is the process of music performance evaluation affected by variables relating to repertoire?

RQ3. How is this process affected by variables relating to the performer?

RQ4. How is this process affected by variables relating to the environment?

RQ5. How is this process affected by variables relating to the evaluator?

To achieve a broad view of the nature of the performance evaluation process, this thesis engages with evaluators ranging in musical experience and expertise, as well as contrasting performers (solo pianist and choral ensemble) and evaluative environments (lone laboratory experimental task and communal live professional concert). These features cut across the research questions above, not with the intention of providing an exhaustive understanding of every possible contributing variable in every setting, but rather to provide insights into as diverse a range of practices as possible.

Four empirical studies are included in this thesis, each incorporating a topic-specific literature review. Study 1 (Chapter 3) examines the role of the *repertoire* by varying and manipulating features of the works presented in audio recordings of solo pianists, using a continuous measurement methodology to track effects on the decision-making processes of a group of experienced musicians. It includes a review of error detection in performance studies. Study 2 (Chapter 4) expands the continuous methodology to variables of the *performer*, introducing the visual modality in the use of manipulated video recordings and widening the population examined to include the perceptions of those with and without musical training. It reviews the role of stage entrances and facial expressions of musical performers. Study 3 (Chapter 5) moves into a naturalistic evaluative setting to examine the affective states of the *evaluator* before and after the performance and how these relate to a final quality judgement, as well as how quality relates to aesthetic judgements. Study 4 (Chapter 6) then turns focus to the concert environment itself, both of the surrounding audience and the various extraneous factors of the physical and social space in their relation to quality judgements. As these studies are varied in their nature, their unique methodological challenges and approaches are discussed in the relevant chapters. However, they are linked by the need for a measurement tool to capture quality ratings, thus the development and use of methods to measure performance quality are discussed in Chapter 2, concerning both the final written ratings used across all four studies and the continuous measures methodologies employed in Studies 1 and 2.

Following the four empirical studies, Chapter 7 then takes a broader theoretical examination of the musical decision-making process and the complexity of the evaluative environment to consider how the act of assessment can be considered a performative skill to be tested and trained. It identifies gaps in the existing research and pedagogical literature and methodologies, then outlines the conceptualisation and development of the *Evaluation Simulator* to engage with the complexity of the evaluative environment and open new avenues for training and research. Finally, Chapter 8 summarises the findings of the thesis as they relate to each of the five research questions, and their implications for research and practice within and beyond musical decision-making.

2 THE TOOLS OF PERFORMANCE EVALUATION

2.1 INTRODUCTION

The evaluation of music performance quality is inherently reductive. Whether providing a verbal or written diagnostic summary, a numbered grade, or a placing in a musical competition, all of the complexity of a musical performance and the performers' technical and expressive ability that comprised it must be collapsed into a practical, yet meaningful representation suited to the purpose of the evaluation. In the case of this thesis, the purpose of collecting quality evaluations in the comprised studies is to examine them and determine generalisable relationships with, and effects of, existing and manipulated performance variables. This calls for a quantitative approach; performance quality reduced to numbers that can be treated to comparative statistical analyses.

Such an approach is highly common to the existing performance quality literature. Indeed, while there were several cases that employed qualitative written analysis (e.g. Davidson & Coimbra, 2001; Kokotsaki et al., 2001), virtually every piece of research discussed in Chapter 1 reduced performance quality to a numbered score and analysed it as such. However, where quality evaluations in research settings may be reductive in comparison with the musical phenomenon and performer's ability they seek to capture, they are not necessarily reductive of many of the real-world practices the research literature seeks to inform. Competitions, conservatoires, schools, exams, orchestras, and often reviews all employ numerical representations of performance quality to some degree. While their value may be debated, their functionality and use are unquestionable (Radocy, 1986). Thus, to examine the methodologies and tools of capturing performance quality in research is to simultaneously examine it in practice, and vice versa.

Whether examining the tools of research or practice, quantifying musical ability is no trivial task. Even setting aside the numerous extra-musical variables that contribute to the evaluative experience as discussed in Chapter 1, one must take the subjective perceptions and judgements of the purely 'musical' material and convert them into comparative and standardised outputs. Even items presumed uncontroversial in their contribution to a performance assessment comes with complications. Take intonation, for example. The degree to which a note is out of tune can be measured linearly, so should the tuning accuracy of, say, a vocalist not contribute in a predictable manner to their performance quality evaluation? Warren and Curtis (2015) found that the extent to which vocalists' mistuning influenced ratings of their performances depended on participants' ability to perceive the slight variations. While this effect was shown in a sample from the general population, musical experts were also inaccurate in that they overestimated the degree to which poor tuning, when it was perceived, influenced the quality ratings. Add to this complexity the influences on intonation perception resulting from harmonic context (Rakowski, 1990) and use of equal temperament or just intonation (Kopiez, 2003), and it quickly becomes clear that even the most fundamental musical constructs cannot be so easily transferred to musical ability.

Thus, this chapter outlines the development of quantitative rating systems of musical quality for musical practice and research, particularly the distinction between holistic and segmented tools, and sets out the approach used through the four empirical studies of this thesis. It also describes several tools for the collection of continuous response data in parallel with musical stimuli, which are used and adapted in Studies 1 and 2 (Chapters 3 and 4).

2.2 SCALES, RUBRICS, AND CRITERIA

2.2.1 Early work: Quantifying musical aptitude

Early attempts to standardise the process of evaluating one’s musical abilities focussed on dividing the musical performance into its component parts and testing them individually. Seashore’s (1939) classic *Measures of Musical Talent* sought to capture musical aptitude via a battery of discrimination tests, including qualities of the pitch (tone, duration, intensity), and concepts of consonance, rhythm, and aural memory. Wing (1947) expanded this to the identification of musically structural and aesthetic elements in his *Standardized Tests of Musical Intelligence*, testing the perception of concepts such as harmonic and chordal analysis, phrasing, rhythmic accents, and intensity. These and similar testing methods from the period faced three significant problems: (1) few could serve as reliable predictors of student ability in an academic or musical sense, (2) the musical community did not have the training to rigorously apply and test the measures, and (3) it was generally recognised by musicians and researchers that musical performance aptitude extended beyond sensory perception alone (Humphreys, 1998). Thus, the use of an experienced evaluator to judge a student’s music performance, and by extension their musical ability, was not replaced. However, this early work highlighted the variability between individuals in the psychological perception of auditory and musical stimuli.

Early attempts to capture solo instrumental music performance ability using more practical means were carried out by John Watkins and Stephen Farnum in the 1940s (Zdzinsky, 1991). Sixty-eight melodic exercises for coronet were created based on existing texts, then administered to students of varying ability in order to rank the exercises in terms of difficulty. Scoring could then be based upon the students’ ability to perform the exercises, with each performed exercise containing errors resulting in a deducted point. The creation of two equivalent sets allowed reliability coefficients to be established (> .90) and high (.82) correlation with teacher performance rankings was achieved. This led to the development of the *Watkins-Farnum Performance Scale* (WFPS), which included transpositions of the material for other instruments and maintained relatively strong reliability and internal consistency. Despite widespread use as a tool for evaluating performance ability in both academic and research settings, however, it remained restrictive based on its inability to distinguish between the nature and magnitude of performance errors, as well as an inability to capture the nature of the student's musicality, creativity, and tone quality (Zdzinsky, 1991).

Gutsch (1964, 1965) took a similar approach in that a set of performance examples limited in musical scope (in this case, algorithmically generated rhythmic patterns), were ranked by difficulty level and presented to students to sight-read. Reliability between two equivalent sets was high (.92) as was comparison to a re-ordered test (.95). However, as was the case in Schmalstieg's (1972) study, these tests remained an extremely limited representation of performance ability. This bottom-up approach of classifying and quantifying the individual factors, while providing promising scores of reliability and consistency, was not addressing the complexity of true performance evaluation. Research thus turned to a top-down examination of the criteria actually being used in such situations with the goal of establishing a benchmark of the qualities inherent to evaluation.

2.2.2 Establishing criteria

Evaluation criteria have traditionally been established based upon agreement between evaluators as to what qualities constitute the ideal music performance (Davidson & Coimbra, 2001). This, unfortunately, leads to a paradox in which the evaluator is guiding the terms of the criteria, which is in turn guiding the terms of the evaluator. As Johnson (1997) stated in his critique of evaluation criteria, "to invoke the examiner's subjective response as to the final arbiter in [the criteria's] validation is to remove a principal reason for having them in the first place" (p. 272).

Moore's (1972) attempt to classify the qualities of the performance being judged using a purely psychological definition of performance as communication engineering was described in Chapter 1 (see Section 1.5), as was Schmalstieg's (1972) work to define the components of a specific musical ideal, in that case vocal tone production. This was followed by a broader approach in which researchers entered the academic evaluation setting to examine what criteria jurors, judges, and teachers were actively using and considering in their evaluations. Abeles (1973) had 17 music education graduate students compose brief essays on the auditory qualities of a clarinet performance, then used content analysis to extract 54 descriptive statements. These were added to an additional 40 statements taken from descriptions of adjudication, and the resultant 94 items were converted to a rubric using both positive and negative framings of the statements (e.g. “he played too slowly”, “musical communication was effective”). These were grouped into seven categories using five-point Likert-type scales in which each evaluator agreed or disagreed with the statement. Fifty music teachers then rated recorded samples (two each) of 50 high-school clarinettists.

Varimax factor rotation was used to establish a six-factor solution in which interpretation, intonation, rhythm continuity, tempo, articulation, and tone were selected. Statements with high-loaded factors were chosen to populate each category and form the *Clarinet Performance Rating Scale* (CPRS). A second rating set using groups of 9, 11, and 12 graduate music education students rated 10 randomly selected videos to confirm the six-factor structure. Inter-judge reliability was shown to be strong (> .70 for most factors and > .90 for total scores). Abeles acknowledged the relative homogeneity of the performances (i.e. a limited range of age and skill; all clarinettists) but hoped that the method could be expanded to other instruments.

Mills (1987) followed in this research, motivated by the newly implemented General Certificates of Secondary Education (GCSEs), which sought to standardise the evaluation of various subjects, including music and music performance, in England and Wales. She also considered the work of Fiske (1975, 1977; discussed in Section 1.6) who found in his study of specialist trumpet evaluation that the criteria ‘intonation’ and ‘rhythm’ were highly correlated with each other but not ‘technique’, and that ‘overall’ and ‘interpretation’ were highly correlated with all five criteria. From this he concluded, “judges should be asked to assign only an overall grade for trumpet performances. This trait was shown to be significantly related to all other traits and, therefore, rating other traits and summing or averaging scores for other traits is a needless, time-consuming operation” (1975, p. 196). Mills found this inadequate, citing examples of assessment schemes that do not rely on marks but rather on written adjudications, often surrounding specific criteria including musicality, technique, and presentation.

Mills' first study (1987) sought to examine the nature of this 'overall' grade, first establishing a vocabulary of evaluative terms that could be applied to all instruments and understood by performing musicians with or without specialist training in the instrument being assessed or in music teaching. Six instrumental performances of students at a Grade 8 ABRSM (Associated Boards of the Royal Schools of Music) level were video-recorded and examined by eleven assessors, each of whom were asked to give an overall mark out of 30 and provide written commentary. This was followed with a semi-structured interview in which Mills questioned each evaluator on the reasons for their choices and asked them to make comparisons between performances using a form of *triangulation*: the evaluators were asked to name shared characteristics of a chosen two performances that were lacking in a third.

Twelve bi-polar statements were formed of the resultant vocabulary (e.g. "this performance was hesitant/fluent") and in a second implementation of the study protocol (using 10 novel instrumental performances and 29 assessors) the statements were presented following the completion of the first assessment. Assessors rated each statement on a four-point scale. Analysis showed that roughly 70% of the variance in the overall marks could be explained using the vocabulary collected in the first phase of the study, a relatively high proportion but leaving nearly a third unaccounted for. The result should also be considered in the light of the relatively small sample size for the number of predictor variables, and the limited degree of variance possible within the four-point scales.

2.2.3 Holistic versus segmented assessments

As a result of her work, Mills (1991) went on to define a classic distinction between two categories of music evaluation tools.

Holistic, or global assessments, which comprise a single, overall rating to encapsulate the quality of a given performance: i.e. the classic 'eight out of ten' or

78/100 or ‘Grade-A’ performance. In practical terms the advantages are clear: a single score allows for easy comparison between performances and subjects. It gives the evaluator freedom to employ their own criteria and weighting of specific points. It allows them to reflect their intuitive judgements without having to ‘show their work’. The strengths of the holistic rating in flexibility and adaptability, however, weaken its reliability. Ratings of multiple performances by a single evaluator may be comparable, but without a fixed criteria or weighting there is no way of inferring whether a second evaluator rewarded the same elements of the performance, or indeed whether one evaluator employed the same evaluative criteria over multiple judgements.

Segmented assessments, which break the ratings into specific categories, often dived into thematic groupings and totalled to give a final, pseudo-global rating. These assessments offer a greater degree of flexibility and nuance to the rating, perhaps giving greater insight into the reasoning behind the assessor’s judgement. Students (and their parents and teachers) also see them as a valuable and necessary insight into and justification for their (their child’s/their students’) assigned ratings (Conway & Jeffers, 2004). However, forcing one’s evaluation into pre-determined categories adds to the artifice of the practice. A musical performance is the result of a complex interaction of performer traits and performance idiosyncrasies—of event-specific errors colouring overall technique, creativity, and interpretation. While some evaluators find that such criteria help them focus on these criteria and pass information concerning each characteristic on to students, others find that they are inherently restrictive and interfere with their ability to provide a true holistic assessment (Stanley et al., 2002).

Mills acknowledged that the trend in musical academia showed a shift from holistic to segmented assessments but emphasised the need for careful consideration of their composition, warning that “introduction of a segmented system with arbitrary weighting does not remove the problem: it only hides it” (1991, p. 174).

Wapnick and colleagues (1993) examined the holistic/segmented dichotomy in its extreme by questioning whether access to any kind of rating scheme at all affects the consistency of their judgements. They differentiated consistency from reliability by describing the former as the stability of a judge's preferences when evaluating multiple performances, while the later refers to the stability of one's rating when evaluating the same performance twice. This was tested in a novel paradigm in which seven interpretations of the same excerpt were presented in all possible pairs over 21 trials. This was duplicated with a second excerpt. Participants simply chose their preference from each pair, and consistency was evaluated by comparing trials for inconsistencies, e.g. if A was preferred to B and B preferred to C, A should logically be preferred to C. Eighty participants (pianists ranging in experience from undergraduates to faculty members) were divided between the two groups, then further divided into categories in which one quarter were given a rating scheme, one quarter were given the musical score, and one quarter were given both the scheme and the score. The rating scheme consisted of eight scales (note accuracy, rhythmic control, tempo, phrasing, dynamics, tone quality, interpretation, overall interpretation) rated on seven-point scales (from 'good or worse' to 'superb' to maximise variance). Consistent with earlier research, the participant's musical experience did not predict their consistency, although faculty scores were slightly higher.

No direct effect of rating scales on the consistency of evaluators was found. This is, on one hand, a promising result: it implies that the use of such scales does not impinge on the evaluator's ability to make a simple statement of preference. It does, however, call into question the function of segmented schemes in the first place: while they provided more information to the researcher (and by extension, the performer) in terms of the reasons for a preference, they did nothing to improve the consistency of the evaluation itself. However, when given a rating scheme in tandem with a musical score, consistency decreased.

The utility of segmented scales was examined by Saunders and Holahan (1997) in a study of 926 high school students seeking entrance to the 1994 Connecticut All-State Band. Thirty-six experienced instrumental music specialists and teachers gave ratings on five-point scales of seven criteria of solo performance quality: tone, intonation, technique/articulation, melodic accuracy, rhythmic accuracy, tempo, and interpretation. Separate criteria existed for the performance of scales and sight-reading. Scales were both *continuous* (where one box of five sequentially more demanding criteria was checked) and *additive* (where five non-sequential criteria could be checked when displayed, each contributing to the final score). Final internal reliability was high (> .90), and when the five most-intercorrelated items (tone, technique/articulation, rhythmic accuracy, interpretation, sight-reading interpretation) were chosen from a correlation matrix and subject to multiple regression analysis, the resultant five criteria accounted for 92% of the variance among total scores. While this outstripped the 70% achieved by Mills (1987, 1991) it was established using a selection of only the five most strongly correlated criteria out of the total 15. It could be argued that aspects of musical production usually considered significant (e.g. intonation, tempo) were shown to be unimportant. On the other hand, removing them may have simply masked a greater complexity in favour of an unrealistic model. Furthermore, each evaluation was the result of a single judge's ranking (the students were divided among the 36 instructors) and thus a fair comparison cannot be made to Mills' findings.

Recognising that different criteria were being imposed upon evaluators across studies and evaluative situations with varying degrees of reliability, Thompson, Diamond, and Balkwill (1998) examined whether judge-specific criteria could be developed. They allowed five experienced music evaluators to develop their own constructs and rate six commercial performances of a Chopin *Etude* based upon their selections. These evaluations were then compared with an overall score. First, each evaluator listened to each performance and made written comments on its quality. Criteria were then chosen for each evaluator using a triangulation method, as used by Mills (1991), in which three of the Chopin performances would be presented and a construct chosen that differentiated one from the other two (e.g. one may significantly differ in expressiveness). The construct was then made bi-polar by having the evaluator define the extremes (e.g. too expressive versus no expression). Finally, the six performances were each rated by placing them along a visual representation of the scale (displayed on a screen). This was repeated for all five constructs. A final trial allowed the evaluator to rank the performances in terms of overall quality, with the instruction that the ranking on this item would determine the ranking in a hypothetical competition.

Inter-judge reliability of performance ranking was shown to be high, with a median correlation coefficient value of .68 between adjudicators. Fourteen distinct constructs were used by the five judges, ranging from the typical dynamics and articulation to repertoire-specific “expression in bars 27-30”, the most common being pedalling (used by four of the five judges). Overall preference was highly correlated with the criteria, but results of both statistical analyses indicated that evaluators deviated to some degree from their own criteria, and post-test interviews indicated that this was often intentional, with one judge choosing to rank a performance where they thought it deserved to be placed despite the low stylistic rankings given on their own constructs.

Thompson and Williamon (2003) examined the utility of a measurement scheme using 13 criteria over three general categories (perceived instrumental competence, musicality, communication) plus an overall quality mark, each assessed on a scale of 1 to 10. Three expert evaluators assessed 61 video recorded performances of varying instruments. While analyses showed that the three general categories were able to predict a high degree of variance in the final mark (approximately 90%), correlations between the three general categories and the dependant variable of the overall quality rating were also extremely high (> .80).

This result brings to light a troubling aspect in accounting for the results of a holistic ranking in terms of segmented criteria. When inter-item correlations are low, it implies that evaluators are able to distinguish the categories as distinct musical features and evaluate them independently. However, such results will often lower the criteria’s ability to predict global score variance. Alternatively, high variability percentages, as those found by Saunders & Holahan (1997) and Thompson & Williamon (2003) have been achieved only when the criteria are highly intercorrelated. This again calls into question the utility of a segmented system, as it implies that evaluators either cannot distinguish between the musical aspects described or that such concepts are so closely linked in their relation to performance that their ratings will always remain close, providing little useful information for those wishing to understand the intricacies of the performance. Furthermore, inter-judge reliability when using only a holistic rating has been shown to be high (Smith, 2004), supporting Fiske's (1975) early dismissal of the segmented system.

Attempts to create novel rating schemes and categorise performance criteria for a variety of instrument- and task-specific outcomes continue (e.g. Madura, 1995; McPherson, 1995; Zdzinsky & Barnes, 2002; Wrigley, 2005; Ciorba, 2009; Geringer et al., 2009; Wrigley & Emmerson, 2013; Russell, 2015; Bergee, 2015; Wesolowski, 2016, 2017; Wesolowski et al., 2017). In the 2017 study by Kopiez and colleagues examining the effects of apparent memorisation on quality ratings (see Section 1.7.2.2), the authors assembled 13 items from existing sources including two non-musical (Ambady & Rosenthal, 1993; Berlo, 1969), which comprised the terms *concentrated*, *committed*, *relaxed*, *stressed*, *authentic*, *certain/confident*, *expressive*, *empathetic*, *rousing/enthusiastic*, *precise*, *sonorous/resonant*, *persuasive*, and *professional*. In doing so, they note the continued lack of standardised music evaluation rubrics that are robust to the standards of current testing procedures, and their own addition to the plethora of available scales remains one more step in the search for a universal system (Gingras, 2017). Qualitative research has found some preference for holistic approaches within higher education settings, with examiners citing lack of agreement with their colleagues on the detail of nature of individual items, and the distraction of attending to multiple criteria as factors weighing against the use of segmented approaches (Gynnild, 2016). Radocy (1989) asserted that evaluators will inevitably differ on the degree of attention and weighting they apply to the various components of multifaceted systems, often turning back to their own holistic assessments.

2.2.4 Approach for the present thesis

Based on this literature, it is apparent that no standardised segmented tool for measuring performance quality across contexts exists. Furthermore, based on the literature presented in Chapter 1, it can be presumed that any of the segmented systems presented above are incomplete as they do not consider the myriad extra-musical performance features already shown to influence performance evaluation though not consciously perceived by the evaluators. It is beyond the aim and scope of the present thesis to posit and establish a novel segmented rating scheme that could be considered superior to those already available. Finally, the purpose of the final written evaluations was to make direct comparison with the continuous ratings given throughout the performance. With current methods to collect continuous responses to music performance, researchers are limited to two simultaneous contracts beyond which the cognitive load of the participant becomes a significant issue (Schubert, 2001). All of the segmented criteria listed above involved three or more items, and, as discussed, the reduction of a segmented set to a single holistic score remains inconsistent. Thus, this thesis employs a *holistic* rating of performance quality assessment, captured as a single number on a Likert-type scale, across each of the four empirical studies. This strategy has three advantages.

1. The single generated score is ideally suited to serve as a single dependant variable in the various inferential statistical models to be used throughout the thesis.
2. A holistic score allows the participant to form a quick, intuitive rating following the processes and criteria they themselves deem appropriate and available, thus facilitating the influence of unconscious cognitive biases resulting from the extraneous performance features examined in each study.
3. A single holistic overall score allows for direct comparison with a single continuous score, potentially allowing participants to use the same internal criteria to generate each and allowing for direct comparison between the two constructs in the first two experimental studies of this thesis.

While a holistic score is reductive in nature and does not provide insight into the nature and weighting of the component criteria, it facilitated the research as described above and the resulting data are interpreted with these limitations in mind.

With the underlying method for measuring music performance quality evaluation for this thesis established, this chapter now focusses on the methodology of continuous measures employed in Studies 1 and 2. After defining the terms to be used throughout the thesis, a summary is provided of continuous measures methodologies used in music research including two bespoke devices developed for the measurement of continuous musical stimuli. Software developed at the Royal College of Music is then presented as an example of bespoke software created for the purpose of continuous measurement in musical contexts. This software is presented as is represents the only method used specifically for the study of performance quality evaluation over time, and is used as the measurement tool in the first empirical study (Chapter 3) of this thesis. Finally, some general considerations in handling the data collected by such devices are discussed.

2.3 CONTINUOUS MEASURES IN MUSIC PSYCHOLOGY

The use of the term *continuous* may be misleading in the discussion of the decision-making literature both within and beyond music research, for the majority has not followed its strict definition (Schubert, 2001). A continuous phenomenon implies an event unfolding over time, without gap or pause. An unbroken set of data documenting the event would also be described as continuous. An example of such a dataset would include that from an analogue kymographic device; i.e. one which translates physical information to a line on a moving sheet or drum of paper via a stylus. However, such devices are limited to a simple physical input that corresponds directly to the subject of measurement, such as geological movement (e.g. the early seismometer), physiological response (e.g. early indicators of blood pressure, skin conductance, heart rate, etc.), or a participant's continuous self-report of a subjective measure (e.g. a human drawing a continuous line that raises or lowers reflective of their response). By comparison, digital devices collect data in discrete, time-specific packages; even a medical-grade ECG device providing an apparently continuous output might be taking 5000 individual measurements per second (i.e. 5 kHz). As these data points are technically separated by (albeit tiny) stretches of time, they are by the strictest definition not *continuous*. Rather, they are *continual*, in that they comprise an event reoccurring at regular, time-separated intervals. The same principle applies in collecting "continuous" measurement of musical judgement. The approaches described below either collected discrete measurements and mapped them to a musical stimulus, or used a digital device that collected one or more discrete measurements per second. In these cases, while the stimulus itself (i.e. the musical performance) could be described as *continuous*, the data collected were, technically speaking, *continual*. Despite this, the term *continuous* has been consistently used within the literature, both within music and beyond, with reference to *time* as the continuous factor (Schubert, 2001).

For consistency and clarity, the present thesis will maintain the convention of referring to the methods and measures as *continuous*, while acknowledging that the data collected in the present and existing studies as described are, by definition, *continual.* With these definitions in mind, the section will now examine the methods and tools used in music research.

2.3.1 Continuous approaches in music research

A 'method of continuous judgements by category' has been used in which participants reported changing perceptions of a single aural stimulus over time by selecting categorical representations (e.g. very loud, very soft, etc.) along a spectrum, captured digitally. This has been used to examine the perceived loudness of aircraft (Namba & Kuwano, 1980) and of traffic noise (Kuwano & Namba, 1985). Continuous studies of response to music have been aided by the nature of the stimulus; as a music performance (and especially a recorded one) is predictable in terms of its general content and the sequence in which it unfolds, temporal correlations can be made to the output itself. Early research used the music's structure as its foundation. Hevner (1936) had participants assign affective terms (e.g. spiritual, dreamy, exciting, etc.) to distinct musical sections of classical works by writing them on the score, thus was able examine affective response both as it changed throughout a performance and as it correlated with musical features unique to each section. Sloboda (1991) used a similar method to track affective responses.

Some research has employed physical indications, such as Goldstein (1980) who had participants raise one, two, or three fingers as they listened to recordings to indicate the temporal location and intensity of the experience of musical 'thrills'.

Clynes (1989) developed and utilised his *sentograms,* in which displacement of a force-sensitive finger pressure sensor was used to measure emotional reactivity to continuous musical stimuli.

Another approach has been the ‘selected description’ methodology, in which evaluators chose adjectives they believed captured their impression of the performance (e.g. graceful, strong, tragic) at the moment they felt it appropriate and entered it into a computer (Namba & Kuwano, 1990; Namba et al., 1991). The frequency of temporal use of each judgement correlated with the adjectives chosen to describe the overall impression of the work, with the authors hypothesising that the overall impression is based on a weighted average of temporal impressions. Napoles and Madsen (2009) developed a paper-based line-drawing method in which participants drew a continuous line on a grid comprising a horizontal, minute-by-minute representation of a musical work and a vertical axis indicating experiential intensity.

2.3.2 Bespoke devices/software

Numerous devices have been employed to improve the reliability and consistency of continuous measurements in music research. Some cases use the continuous sliders, knobs, and tools available on Musical Instrument Digital Interface (MIDI) equipment (Vines et al., 2006) or custom software (Nagel et al., 2007; Grewe et al., 2009; Ferguson & Schubert, 2011). Custom devices have also been developed. Examples of three bespoke solutions are highlighted here as demonstrations of ways in which the various methodological challenges have been addressed and which approach best suits the present study.

2.3.2.1 CRDI

The *Continuous Response Digital Interface* (CRDI) was developed at Florida State University’s Center for Music Research in the late 1980s (Gregory, 1989; Madsen, 1990). The device takes two forms; a large dial that rotates 256 degrees (see Figure 2.1) and a box with a sliding lever, both of which transmit continuous responses to bespoke software that catalogues the temporal data and can present it for analysis, and both of which can be overlaid with a variety of rating scales. While simple, the

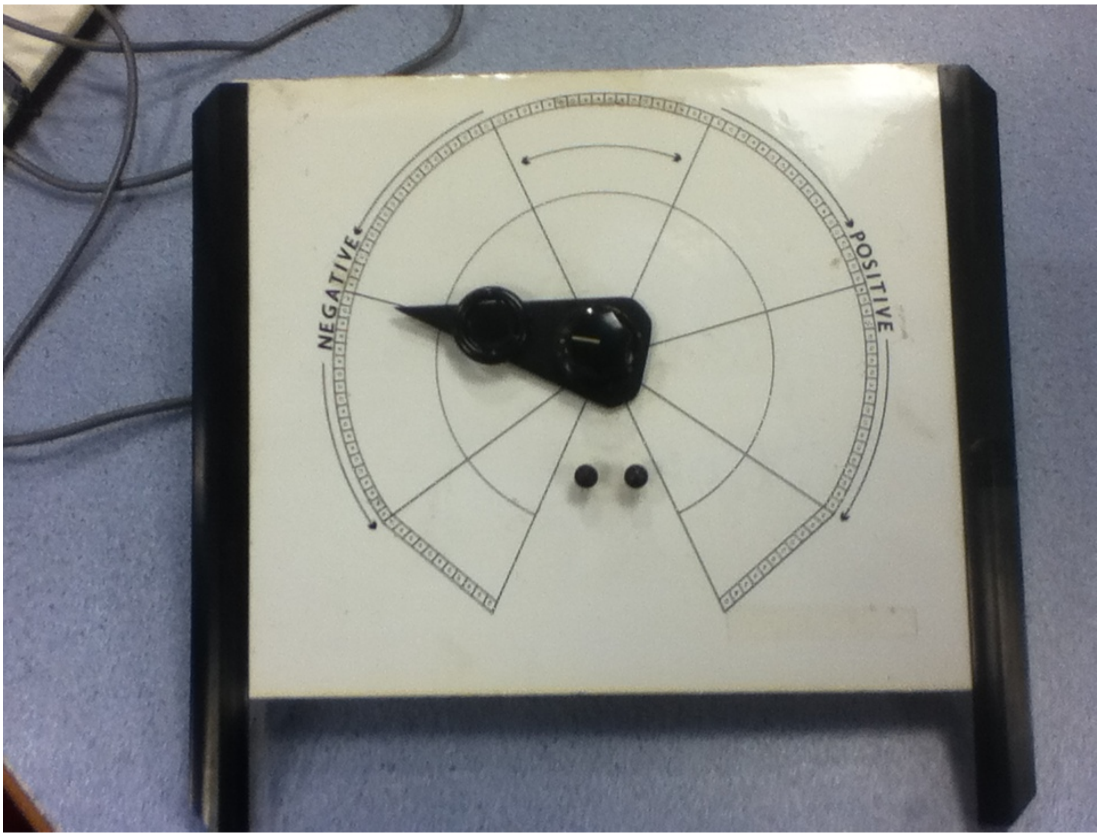


Figure 2.1. The dial version of the Continuous Response Digital Interface (CRDI).

devices were constructed with several goals in mind. First, new advancements in computing technology called for a device that could provide a direct link between physical movement and digital storage and analysis. Secondly, the creators acknowledged that asking evaluators to provide continuous ratings of performances could cause cognitive overload and redirect attention from the stimulus being measured to the measurement itself, especially in children and those without musical training (Madsen, 1990). Thus, the interfaces were designed to be as simple as possible. Finally, the CRDI is flexible in the nature of data it may collect. While the temporal occurrence of decisions remains the primary focus, those decisions may be made within two general types of scale:

- *Continuous*, in which ratings are given between extremes of a single dimension (e.g. loudness, tension, affective response) along a continuous scale. In practice, the scale will be broken into individual sections for measurement (similar to the sampling rate of the temporal data). In the case of the CRDI, divisions are made along the 256 degrees of rotation (Gregory, 1995)
- *Categorical*, in which the dial is moved to correspond to predetermined zones, each representing a specific category (e.g. instrument as focus of attention, term to describe affective response, Likert-type multiple choice scale, etc.) The position within each zone is irrelevant; rather, the frequency and timing of category selection is measured (Gregory, 1995).

Multiple devices may be used at one time, allowing both for a group of participants to be simultaneously tested and for multi-dimensional studies of a single listener (e.g. simultaneously measuring emotional valence and affective arousal). However, the cognitive load of participants in such cases and their ability to accurately report multiple responses simultaneously must be taken into consideration, and usually limits the number of simultaneous continuous contracts to two (Schubert, 2001).

The CRDI has seen widespread use in the music psychology community. As of 2004 the device's creators documented over 70 studies and 20 dissertations in which it was employed, primarily in studies of affective response and focus of attention (Geringer et al., 2004). Its use has continued over the past decade in an expanding variety of topics, including rubato tendencies in horn performers (Johnson et al., 2012), the effects of subtitles in perceptions of expressivity in opera (Silveira & Diaz, 2014), and how the performance of the accompanying pianist affects expressivity ratings of violin/piano and vocalist/piano duos (Geringer & Sansafar, 2013).

2.3.2.2 CReMA

The *Continuous Response Measurement Apparatus* (CReMA) was introduced by Himonides and Welch (2005; Himonides, 2011) and built to expand upon the capabilities of the CRDI. Primarily, the creators noted that a challenge in interpreting the CRDI lies in the nature of the dial (or sliding rod); if the dial is left stationary there is no way of indicating whether the evaluator is consciously holding their judgement at a particular value, or whether their attention has been brought away from the assessment process. Furthermore, if the assessor wishes to make an instantaneous decision change between points or categories within a scale, they must pass through every intervening point. While the researcher may infer that a leap in judgement was intended based on the speed of such a movement (and the change may have taken place at a pace too quick to be registered at the sample rate employed) there is again room for error in inferring what was intended. The CReMA resolves this issue by allowing for a ‘no-input’ situation. As ratings are measured by the physical position of the evaluator’s finger across a physical strip (see Figure 2.2) they may simply remove their finger to indicate that no judgement is being made at that point. They may also indicate the nature of a judgement change; sliding the finger from one point to the next indicates a gradual or continuous change, while lifting the finger from one point on the device and returning it elsewhere may indicate an instantaneous change.

The CReMA device is also designed to consider physiological aspects of the listening task. In addition to location data, the sensor measures the physical pressure exerted, which the creator hypothesised may be exerted subconsciously as a correlate of emotional response (Himonides, 2011). Furthermore, the data output was designed to synchronise and allow comparisons with physiological data (e.g. heart rate, body temperature, galvanic skin response; see Figure 2.3) via its high sampling rate (up to 200,000 Hz) and compatibility with software used for physiological monitoring.

Early use of CReMA has explored the relationship between affective response, physiological response, and quality ratings when assessing vocalists (Himonides, 2011). General correlations were found between galvanic skin response and quality ratings of diction, dynamics, and overall quality. The conceptual framework of the

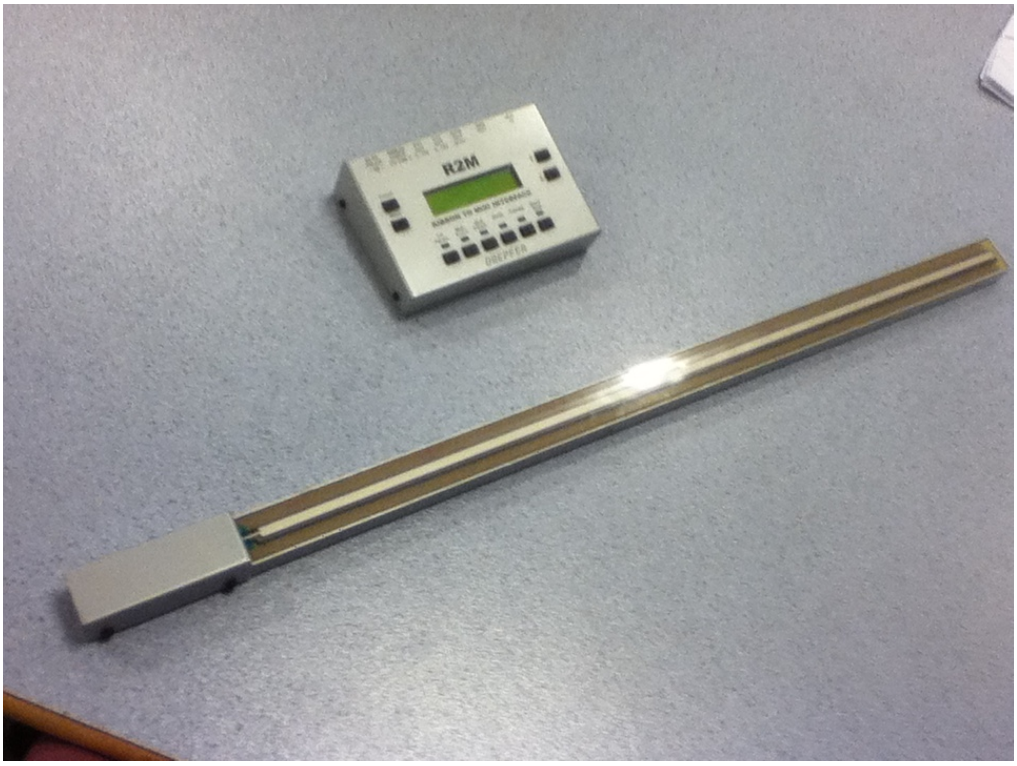


Figure 2.2. The Continuous Measurement Response Apparatus (CReMA).

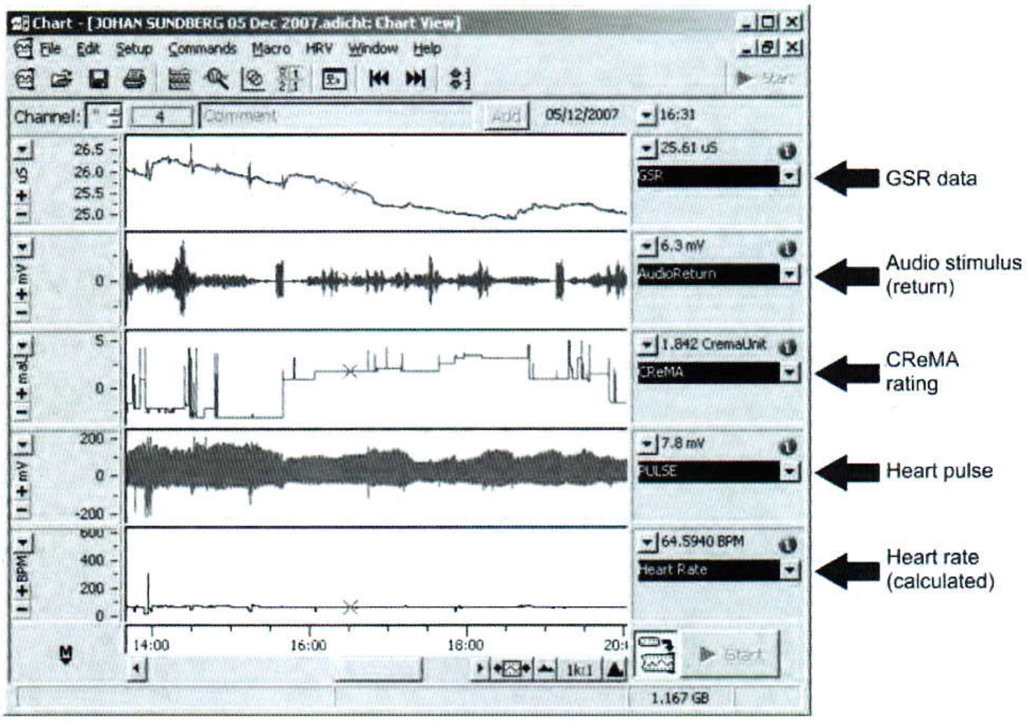


Figure 2.3. Affective responses from CReMA synchronised with the physiological responses of the listener and the audio stimulus (from Himonides, 2011, p. 16).

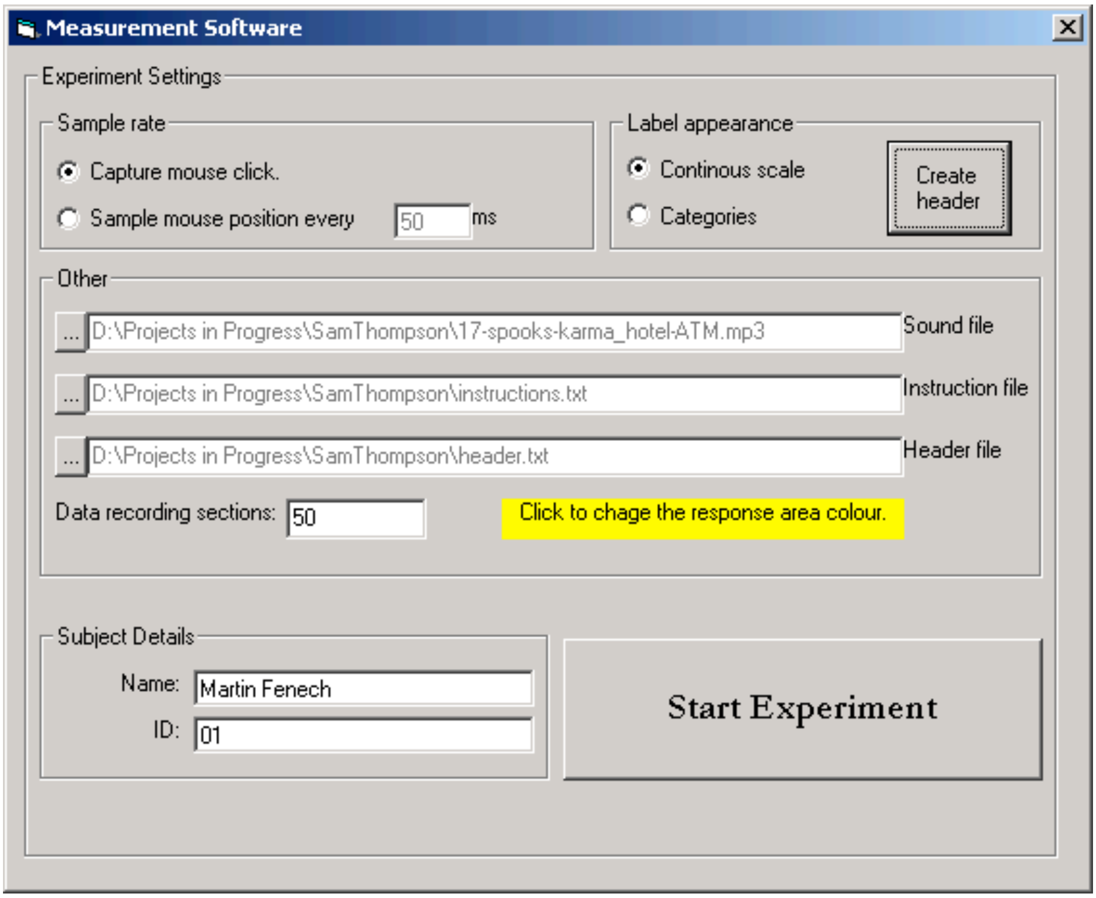
CReMA device has since been expanded to a software tool that allows for the capture of continuous data via commercial MIDI instruments and their varied sliders, dials, keys, and other control interfaces (CReMA MIDI; Himonides, 2017).

2.3.2.3 RCM software

Devices like CRDI and CReMA are, at their core, fulfilling a simple principle: capturing self-reported decision-making via conscious motion over time into a computer. While such bespoke devices offer a greater ability to make comparisons across studies due to identical input methods, any modern computer has the ability to collect categorical data via the keyboard or continuous (scalar) data via mouse or trackpad movement. Thus, software that uses inherent computer hardware as the input source has been developed for continuous measures studies in music. This has three distinct advantages: (1) flexibility in the on-screen display on which the mouse pointer is moved or key choices described, (2) flexibility in the nature of the input device (mouse, trackpad, joystick, customised keyboard, etc.), and (3) familiarity of the participant with the use of a mouse or computer keyboard.

While software has been developed for examinations of perceived emotions in music (e.g. EMuJoy: Nagel et al*.,* 2007; Grewe et al*.,* 2009) the software developed and employed by Thompson and colleagues (2007) in the Royal College of Music's (RCM) Centre for Performance Science is here examined. As described in Chapter 1 (see Section 1.9), this research represents one of the only uses of continuous measures methodologies to specifically address performance quality evaluation, and it thus forms the foundation of the present thesis' methodology so that cross-study comparisons may be made.

The RCM software uses mouse movement along a rating area to collect continuous rating data. The main screen (see Figure 2.4) allows for the alteration of several variables (Thompson, 2005):



Measurement Software

Experiment Settings

Sample rate

● Capture mouse click.

○ Sample mouse position every 50 ms

Label appearance

● Continous scale

○ Categories

Other

D:\Projects in Progress\SamThompson\17-spooks-karma_hotel-ATM.mp3 Sound file

D:\Projects in Progress\SamThompson\instructions.txt Instruction file

D:\Projects in Progress\SamThompson\header.txt Header file

Data recording sections: 50

Click to chage the response area colour.

Subject Details

Name: Martin Fenech

ID: 01

Figure 2.4. Main screen of the RCM continuous measurement software.

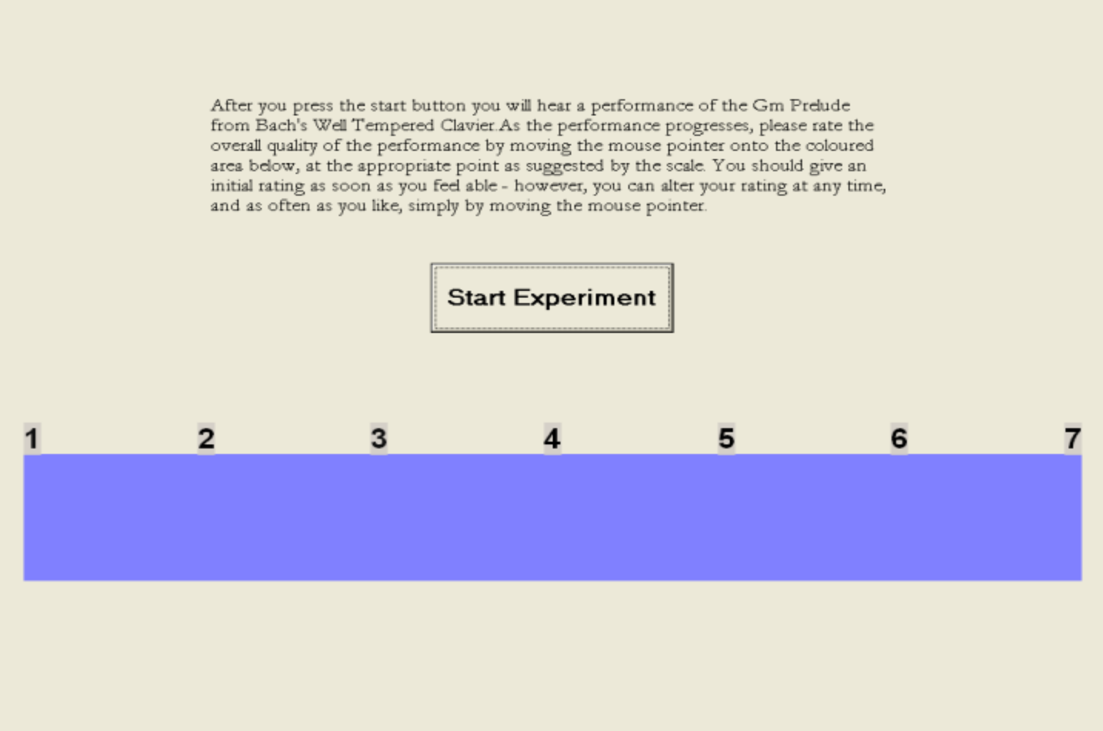
- The sample *rate* (see Section 2.3.3.1) of mouse position along the rating area (see Figure 2.5), as well as the option to capture mouse clicks within the area instead of regularly timed positions. This indirectly addresses the concern that Himonides and Welch (2005) had with the CRDI device, in that the researcher may wish to record the points at which instantaneous decisions were made, and not track movement from one point to the next.

- The number of *data recording sections* along the rating area. This allows for both a practically continuous scale of up to 100 discreet sections where most mouse movements would be detected while allowing for simplified data analysis, and a forced categorical scale in which a series of large, evenly spaced areas may be applied where movement within each block is not recorded. The researcher may then alter the *label appearance* to visibly indicate the categories with vertical lines and label them with a custom header, or alternatively they may use a continuous scale with a header providing a basic guide (e.g. a Likert-style series to correspond with a standard written evaluation) and no lines. For easy data handling, Thompson (2005) applied a 7-point guide over a continuous scale of 70 areas (as well as written reports using 7-point scales) so that the continuous data could be easily mapped to and compared with the overall judgements. The colour of the rating area may also be adjusted.

- The *sound file* to be evaluated by the participant. Timings of mouse position are measured against the point at which the audio track is begun, therefore the researcher must manually insert a period of silence into the track if desired and take the length of that silence into account when analysing the data.
- The *instruction file*, containing text instructions to be presented to the participant before beginning the experiment.
- The *subject details*, which are included in a spreadsheet of two columns: the timing data (in milliseconds from the start of the audio playback) and the mouse position data (in terms of the number of data recording sections specified).

Participants are shown the rating screen in Figure 2.5. After reading the supplied instructions they press the “Start Experiment” button to begin playback (following any silence the researcher may have added to the track). Data recording starts at the point the mouse enters the coloured area, thus allowing a first judgement to be measured.

For the first experimental study of the thesis, the RCM software served as the initial tool for data collection. This allowed for comparisons to be made with the work of Thompson and colleagues (2007) by replicating the basic methodology while varying the nature of the repertoire being studied. In comparison to tools such as the CRDI and CReMA, its principal disadvantage is the lack of bespoke hardware with



After you press the start button you will hear a performance of the Gm Prelude from Bach's Well Tempered Clavier.As the performance progresses, please rate the overall quality of the performance by moving the mouse pointer onto the coloured area below, at the appropriate point as suggested by the scale. You should give an initial rating as soon as you feel able - however, you can alter your rating at any time, and as often as you like, simply by moving the mouse pointer.

Figure 2.5. The participant display of the RCM software, showing a 7-point heading over a continuous rating area on which mouse position is tracked.

which physical input is collected. However, this was remedied with the consistent use of a mouse between participants. In Study 2, the addition of the visual modality required a system that could present video, thus a custom software solution was developed (see Section 4.2.3).

The remainder of this chapter will now discuss general considerations that must be given to data collected from such devices.

2.3.3 Handling continuous data

The particular nature of continuous measurements methodologies requires special consideration in the collection and handling of the resultant data. Some decisions, such as the sample rate, must be made before collection, while others involve processing of the collected material before analysis can take place. While each specific device and methodology carry with it specific requirements, some general considerations are outlined here.

2.3.3.1 Sample rate

The sample rate of a continuous measurement is simply the number of regularly spaced measurements taken per second, measured in hertz (Hz). A system measuring one response per second, for example, would be recording at 1 Hz. Setting an appropriate sample rate is key: too low and pertinent information may be missed; too high and the sheer amount of data to be processed may complicate the analyses and lead to false precision. Traditionally, the sample rate of continuous signals is set via the *Nyquist frequency*, defined as a rate twice that of the highest frequency to be measured (Upham, 2011). In terms of music performance assessment, that frequency is determined by the rate at which listeners are able to form judgements of newly presented material. Studies examining the latency of loudness judgements in music performance (e.g. 2.5 seconds by Kuwano & Namba, 1985; one second by Geringer, 1995) have thus far been used to set the current standard of 2 Hz (one measurement every 500ms) in music studies (Schubert, 2001). While this standard was not set upon any research concerning latency in music performance quality judgements, recent work by Plazak and Huron (2011) has demonstrated that listeners, while able to identify descriptive features such as instrumentation in as little as 100 milliseconds of musical stimulus, required approximately one second of music before emotional responses were reported and two seconds before offering judgements of performer skill. Should these periods represent minimum response times to performance quality they would easily be captured in a 2 Hz measurement.

2.3.3.2 Other considerations

Due to the large amounts of data that may be collected in a continuous measures methodology, special consideration must be given to preparing the data so that the inclusion of redundant and irrelevant information might be minimised. Such processes include filtering, removing outliers, and normalisation (Upham, 2011).

Filtering involves the removal of redundant information, often an artefact of the data collection tool. For example, in his examination of performance quality measurement, Thompson (2005) asked participants to move the mouse to the rating area only when they felt they were ready to make a first decision. However, he found that, in practice, this first movement did not always correspond to a single score, but rather showed an initial cluster of rapid changes as a result of the mouse not being moved in a perfectly even fashion. As a result, the first position that was held for at least two seconds was considered the value of the first judgement, and all preceding position data were altered to that value.

Removing outliers (data sets that vary excessively from the results, often as a result of methodological error and/or a participant not understanding the task) is a fundamental part of any behavioural study. In terms of continuous measures methodologies, Upham (2011) describes the case of the *flat-liner*, wherein participants fail to provide or change judgements for unusually long periods of time. Such cases may be grounds for removal if they are determined to be the result of a participant forgetting to provide judgement information or a malfunction of the equipment.

Continuous measures data allow for an examination of the contour of response changes independent of the overall or moment-to-moment values. In these cases, *normalisation* can be used to simplify such value differences between cases and highlight the temporal differences (Upham, 2011). Two forms can be of particular value: *normalisation of range*, which involves the stretching and shifting of each response so that its maximum = 1 and minimum = 0, and *z-normalisation*, in which the distribution of each response is set to have a mean of 0 and a variance of 1 standard deviation.

2.4 SUMMARY

This chapter outlined the development of scales for capturing music performance quality ratings. Holistic scales comprising a single overall score and segmented scales comprising a selection of criteria were defined and compared. While a large and growing literature has endeavoured to establish a standardised segmented rating scheme, no one system has yet been established. Thus, a holistic approach was employed in the present thesis to avoid suppression of extra-musical performance influences and allow direct comparison to continuous measures data. Several tools for the collection of continuous data were described, and the custom software employed by Thompson and colleagues (2007) chosen to be used in the first empirical study to allow for direct comparison of the data, particularly in addressing the first research question of the present thesis.

3 STUDY 1: THE REPERTOIRE

3.1 INTRODUCTION

This chapter represents the first empirical study of the present thesis, designed to address the questions raised the overarching research aims 1 and 2:

RQ1. When are decisions made and adjusted while assessing the quality of a musical performance?

RQ2. How is the process of music performance evaluation affected by variables relating to repertoire?

This study comprises two principal components. First, a partial methodological replication of the work by Thompson and colleagues (2007; see Section 1.9) to determine whether the benchmarks for time to initial decision could be replicated. Second, an expansion of focus to include the experimental manipulation of variables relating to the repertoire to determine their effect on the evaluative process and products. As outlined in the new process model of music performance evaluation introduced in Chapter 1 (see Section 1.7.1), *repertoire* in this context refers to qualities of the work itself, both of its inherent nature and its relation to the evaluator. This is considered separate to qualities of the performance. An examination of every known variable is not intended here. Rather, representative variables have been chosen, and specifically those that have been addressed previously in the literature so that hypotheses may be drawn. Three principal variables were therefore examined: repertoire length, repertoire familiarity, and repertoire likeability.

In addition to these features, the use of audio-based stimuli and the examination of continuous measures data provided an opportunity to examine the nature of a key component of performance and of evaluation; the commitment of performance errors. Within the context of the process model posited by this thesis, performance errors fall most comfortably into the domain of the *performer*, reflecting upon their technical skill, control, preparation, and reaction to performance circumstance. However, such errors also bear a strong relationship with the *repertoire* in that they represent deviation from the expected notated standards, and they take place at points relative to the work's temporal structure. Thus, performance errors comprised the fourth variable examined in this study.

3.1.1 Features of the repertoire: Length

Length is a fundamental feature of any composition and provides an easily quantifiable differentiator—a work may be said to be two, five, or one hundred times longer than another—while mode, expressivity, or difficulty, for example, are not so easily quantified. Of course, the tempo of an interpretation may alter the length of a performance, but the repertoire itself determines the baseline. The length of an excerpt used in research settings is often not varied or questioned, although studies by Wapnick and colleagues found that excerpts of differing lengths were rated differently. In a first study (2005), two groups of participants rated recordings of 19 classical music excerpts 20 or 60 seconds long. Length condition was randomised within each group and counterbalanced between groups. Participants were not informed in advance of the length of each recording. The results showed that the longer excerpts received significantly higher and more consistent ratings, measured as differences in group standard deviations. In a subsequent 2009 study, participants were given 25-, 55-, or 115-second excerpts of performances, again rating the longer two excerpts significantly higher than the 25-second excerpt. The researchers varied certain visual characteristics, finding that attractiveness, sex, dress, and stage behaviour produced conflicting effects for different lengths of excerpt, such as dress affecting men's ratings of the 25- and 115-second but not the 55-second excerpts. Overall, these studies highlight variation in the evaluation process depending on the length of the task, although they did not examine these effects with full-length compositions or in situations where participants knew in advance the length of time they had to make their decisions.

Outside the musical domain, research has suggested that the total time to determine an applicant's suitability in interviews is mediated by the predetermined length of the interview (Buckley & Eder, 1988). One such study demonstrated that participants viewing video-recorded interviews of approximately 15 minutes took longer to reach a final decision if they were first informed that the session would take 30 minutes (Tullar et al., 1979). Crucial to the method was the participant's knowledge (though faulty) of the interview length prior to its beginning. Thus, in a musical context, the length of the excerpt would need to be explicitly stated before its presentation for an accurate comparison; in a range of settings, from listening live in concerts and examinations to listening to recordings, it is not uncommon for timing information to be available to the listener.

3.1.2 Features of the repertoire: Familiarity

Familiarity with the work takes into consideration the knowledge of the evaluator. Indeed, the very definition of a musical expert in evaluation settings usually includes knowledge of common repertoire or experiences of engaging with new repertoire. Such a connection makes sense: evaluators who are aware of the framework on which the interpretation is to be made are, in theory, primed for the information that is to be presented to them and have a standard to which they can compare variances in individual interpretations. In line with this, Kinney (2009) found that evaluators' familiarity with a work improved their internal consistency when forming quality judgements of performances of that work. In terms of the temporal aspects of decision making, one could hypothesise that familiarity with a work would decrease the time to the first and final judgement, as less effort would be needed to understand and process the nature of the work being presented and thus the attention could be shifted to the quality of the performance itself. However, another advantage of familiarity could be increased awareness of the structure of the work, including perhaps the points at which the most technically challenging and musically defining moments will take place. One could then hypothesise that familiarly would *increase* the time to a final judgement, as evaluators would delay their decision until the expected points of interest arrived. This would be specific to the work and the points which the particular evaluator considered of interest. Such hypotheses have not yet been investigated, but a continuous measures methodology would allow for the relationship to be examined directly.

3.1.3 Features of the repertoire: Likeability

Related to familiarity is the concept of the *likeability* of a work—that is, does the evaluator have an inherent preference for the composition itself? While every listener (and evaluator) is entitled to such preferences, it would be problematic if they were to interfere with the evaluative process if it were taking place in educational or competition contexts. Research specifically examining the relationship between performance quality rating and preference for the work is lacking. Several studies have found a relationship in the perceived quality and liking of popular and classical music recordings (Hargreaves et al., 1980; North & Hargreaves, 1998), although these framed quality as the overall value judgement of the work, not the specific performance or interpretation of it. It is generally assumed that one's preference for a work is tied closely to one's familiarity with it, although Thompson (2007) found that the two concepts could be separated to some degree in that likeability, but not familiarity, of a work was predictive of enjoyment. The same study also found that performance quality could be separated from affective response, suggesting that the evaluative process may be unchanged despite differences in preference for a work, but such assumptions have not been experimentally tested. In a study of music in popular charts, Russell (1987) found that listeners' familiarity with a piece increased with repetition, but likeability did not.

3.1.4 Performance errors

Regarding *performance errors*, previous studies have examined the ability of musicians of varying experience to detect manipulated 'mistakes' in recordings. Byo (1993) asked participants to detect errors in recorded excerpts of polyphonic wind band repertoire, manipulated to contain performance errors. Analyses found that listeners were better able to identify rhythmic than pitch errors and improved in identifying both when the instrument timbres were similar across voices. A later study (Byo, 1997) supported these findings, also demonstrating that experience and rating monophonic (versus polyphonic) textures increased error detection rates. Repp (1996) found that listeners detected only 38% of pianists' pitch-based errors, including missing or unnecessarily repeated notes. The nature of performers' errors has also been examined, with research demonstrating that errors are more likely to be made in the middle of phrases away from structural boundaries (Mishra, 2010); that the majority of pitch errors in a corpus of Chopin recordings were note omissions, with a significant proportion of errors systematically repeated (Flossmann & Widmer, 2011); that performers can detect that they are about to perform an error immediately before the motion is executed via electroencephalographic (EEG) event-related potentials (Ruiz et al., 2009; Maidhof et al., 2009, 2013); and that EEG negative potentials immediately following the perception of an error are more pronounced when performing than when listening (Maidhof et al., 2010).

No studies to date have examined the effects of errors on temporal quality ratings, and in particular, issues relating to their location. Thus, the question remains as to whether a mistake at the beginning of a piece is more harmful to one's evaluation than one at the end. Research in interpersonal impression formation would suggest so, as negative first impressions have been found harder to alter than positive ones (Ybarra, 2001), yet this is still to be examined in the context of music performance evaluation.

3.1.5 Aims of the present study

The present study aimed to examine the effects of repertoire length, familiarity, and likeability, as well as the location of performance errors, on the temporal process of forming quality evaluations. Previous studies have demonstrated that participants form their initial quality judgements within an average of 15 seconds when rating audio (Thompson et al., 2007) recordings of standard repertoire, with no correlation found between time to first decision and overall quality rating. Thus, hypothesised increases or decreases in decision time resulting from differences in features of the works themselves were posited based on the existing literature:

1. Works of lesser familiarity and likeability would result in an altered time to first and final decision. This effect would be increased in the case of a work of unfamiliar tonal framework and composer. The direction of the effect was not hypothesised.
2. A work of shorter length (when work length is known beforehand) would result in a shorter time to first decision.

Regarding the performance errors, two hypotheses were:

3. A performance error inserted at the beginning of a composition would result in a shorter time to first decision.
4. A performance error inserted at the beginning of a composition would result in a lower final rating than the same error inserted part way through the performance.

To test these hypotheses, works of varying length and familiarity were chosen. In addition, a difference in genre (i.e. Romantic versus twentieth-century) and popularity of composer (famous versus relatively unknown) was used to emphasise the familiarity contrast in one of the five chosen works. Performance errors were added digitally to several of the performances, with every effort made to create the impression of live, undoctored recordings.

3.2 METHOD

3.2.1 Participants

Forty-two musicians were recruited via email and in person from the Royal College of Music (RCM) and Imperial College London. The cohort comprised 24 women and 18 men with a mean age of 27.2 years (SD ± 9.9, range = 18 - 55). Musical experience among the group varied, ranging from undergraduate to doctoral students and including 4 professional musicians, with a mean 19.9 years of musical experience (SD ± 9.6, range = 5 - 51). Fifteen participants reported the piano as their primary instrument, and of the remaining 27 (12 strings, 8 winds, 4 voice, 1 brass, 1 organ, 1 harp), 20 reported the piano as a second study instrument. Informed written consent was obtained from all participants following the ethical guidelines of the British Psychological Society and with internal RCM approval on behalf of the Conservatoires UK Research Ethics Committee. No payment was given in exchange for participation.

3.2.2 Stimuli

Repertoire was chosen to vary in length, familiarity, and genre. The piano works of Frédéric Chopin were selected as they provided a wide range of compositions with a distinct, overarching style by a well-known composer and including compositions of less than one minute in length. Four of Chopin's works were chosen: the 'Black Key' Etude in G-flat Major, Op. 10 No. 5; the 'Minute' Waltz in D-flat Major, Op. 62, No. 1; the Prelude in D Major, Op. 28, No. 5; and the Tarantelle in A-flat Major, Op. 43. These were selected to match in mode (major key), tempo (fast: 100-150 beats per minute), and texture, with a scalar and arpeggiated right hand over accompanying figures in the left. Of these, the Etude, Waltz, and Tarantelle were chosen as longer pieces (> 100 seconds) and the Prelude as a short piece (< 30 seconds). They were also chosen to vary in familiarity, ranging from very popular with the Etude and Waltz to relatively unknown (as much as is possible with a work of Chopin) with the Tarantelle. To create a stark familiarity contrast, the Caprice No. 6 'Klavierstuck' by twentieth-century composer Sophie Carmen Eckhardt-Gramatté was chosen. The work bears technical similarities to the selected Chopin works in its use of melodic material in the right hand over accompaniment in the left but employs an expanded, less familiar tonal framework. As performances of shorter complete works of this nature were not available, an excerpt taken from the beginning to a point that could be perceived as a functional finale was used to match the length of the Tarantelle, the most unfamiliar Chopin work. The selected compositions, their lengths (in terms of the performance used), and their approximate tempi are shown in Table 3.1. Piloting was undertaken via informal discussions with undergraduate- and

Table 3.1. Works used as stimuli for Study 1.

Composer	*Title*	*Length (s)*	*Tempo (bpm)*
Chopin	Etude in G-flat Major, Op. 10, No. 5 'Black Key' Etude	108	~110
Chopin	Waltz in D-flat Major, Op. 62, No. 1 'Minute' Waltz	117	~100
Chopin	Prelude in D Major, Op. 28, No. 5	27	~100
Chopin	Tarantelle in A-flat Major, Op. 43	156	~150
Eckhardt-Gramatté	Caprice No. 6 'Klavierstuck' (excerpt)	152	~130

graduate-level pianists to confirm that assumptions made concerning familiarity and the choice of end-point of the twentieth-century piece were valid.

MIDI (Musical Instrument Digital Interface) recordings of the Chopin works were used to allow for the controlled insertion of performance errors at strategic points. These recordings were taken from *The Classical MIDI Resource*, an online repository of openly submitted MIDI recordings that are editor-reviewed for quality and accuracy before being posted for free download. The Eckhardt-Gramatté Capricewas recorded acoustically by a graduate-level pianist, itself requiring no manipulation as it was not a part of the error examination due to its lack of recapitulating material. To ensure that the artificially inserted errors would be both believable and easily perceived, dissonant errors of pitch in a single voice were chosen as they have been shown to be both the most common in piano performance (Flossmann & Widmer, 2011) and the second-most-easily perceived, after rhythmic errors (Byo 1993, 1997). To test the effect of error location, the two familiar Chopin works of the same length (the Etude and Waltz) had also been selected due to the recapitulation of their opening thematic material. Thus, a pitch error in the opening seconds of the performance could be recreated mid-way through, differing only in temporal location and structural context. To match error type as closely as possible, Logic Pro 9 was used to transpose an arpeggiated figure in the right hand of approximately one bar in length up one semitone in each work, simulating a pianist that had played a brief passage with the hand in the wrong chord position. Three tracks were then created for each of the two works: one with an error at the beginning (error-start), one with an error at the recapitulation (error-recap), and one control condition without an error (no error). An error of the same nature was added to the beginning of the Prelude to test the interaction of opening error and work length. A summary of the variables associated with each work is provided in Figure 3.1.

Although MIDI files of piano recordings have been successfully employed in previous studies of music performance evaluation (e.g. Winter, 1993; Sloboda & Lehman, 2001; Thompson et al., 2007; Kinney, 2009), digital enhancement was undertaken to add further realism to the files and to match the recording quality of the contemporary excerpt. Specifically, Logic Pro 9 was used both to realise the MIDI

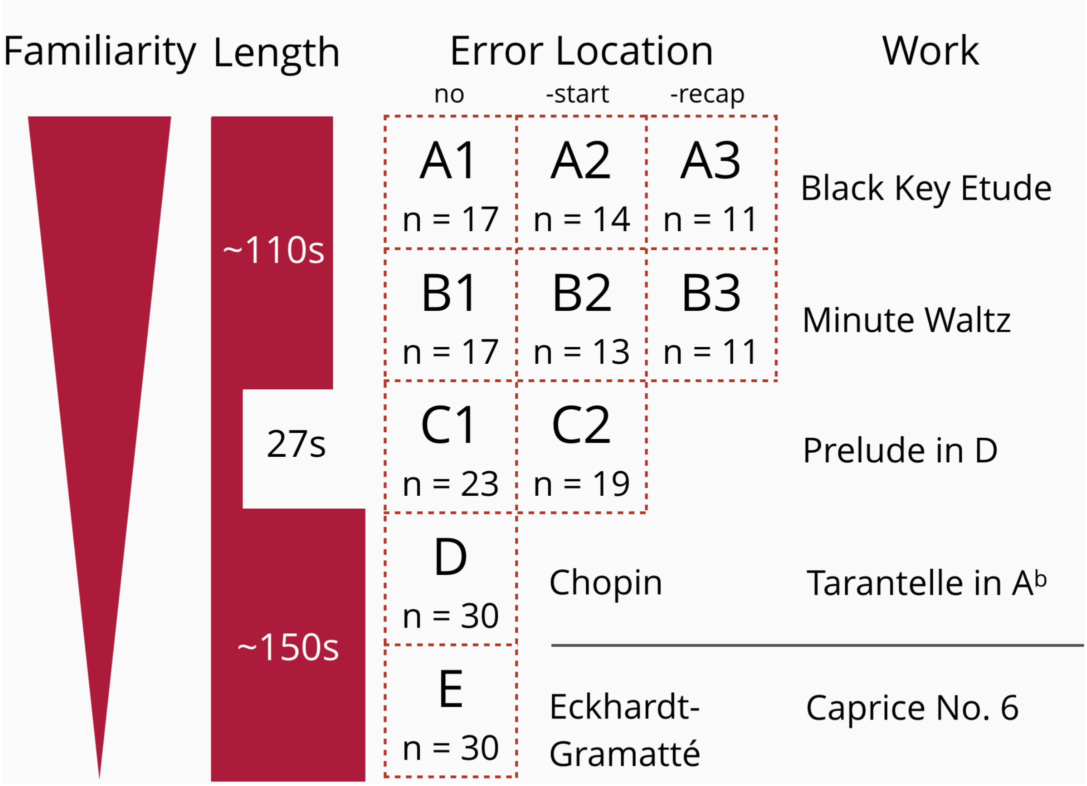


Figure 3.1. The Study 1 research design, showing the variables of repertoire length, relative familiarity, and error placement in the five works. The number of participants assigned to each condition are shown (total N = 42; see Section 3.2.5 for a description of participant distribution). Audio recordings of the 10 experimental stimuli are available to download via the reference in Appendix 1.

data into audio formats and to add three effects: (1) reverb, to emulate the acoustic of a performance space in a live recording; (2) stereo split, to break the mono output of the MIDI realisation into slightly varying signals as one would experience in a true stereo recording; and (3) distortion, applied sparingly to approximate the signal-loss inherent to audio recording and dull the overly bright and harsh quality often associated with MIDI recordings. Manipulations of the MIDI data also allowed for the removal of overt performance eccentricities (e.g. occasional over-accented, jarring notes or the addition of slight tempo fluctuation in overly-metronomic passages, a common characteristic of MIDI recordings). The tracks were then converted to .wav format, and 4 seconds of silence were added to the beginning of each to allow for the listeners to prepare themselves after commencing each trial. Informal piloting with graduate-level pianists confirmed that the performances could pass for genuine acoustic recordings. Audio recordings of the 10 experimental stimuli are available to download at the link in Appendix 1.

3.2.3 Continuous measures

For the present study, the continuous measures data were collected using the software developed at the Royal College of Music described in Chapter 2 (see Section 2.3.3.3) and used by Thompson and colleagues (2007). To summarise, the software comprised a horizontal blue bar onto which the participant moved their mouse pointer when they were ready to register their first judgement and then along which they could move the pointer to increase or decrease continuously their rating as appropriate. For this study, the horizontal area was divided into 70 discrete sections, not visible to the participant, while a 7-point scale (from 1 “poor” to 7 “excellent”) was overlaid above the rating area for easy transfer to the written evaluations (see below). Data points were sampled at 2 Hz. The software was presented to each participant on the same Windows-based laptop with USB mouse and Sennheiser HD 380pro headphones.

3.2.4 Written evaluations

Two bespoke questionnaires were used in the study. The first was completed immediately following each trial and assessed the participants’ relation to the work and overall evaluation of the performance on 7-point Likert-type scales along several categories: overall quality of the performance (1 “poor” to 7 “excellent”), familiarity with the work (1 “never heard it” to 7 “extremely familiar”), and degree to which they like the composition (1 “not at all” to 7 “very much”). The typicality of the performance in relation to others they have heard (if applicable) and the perceived difficulty of the work to perform was also measured on 7-point scales. Participants were encouraged to provide comments concerning each performance. The second questionnaire, completed at the end of the study, elicited background information on musical training and listening preference by musical genre: Baroque, Classical, Romantic, and twentieth-century, each measured on a 7-point scale. The questionnaires can be found in Appendix 2.

3.2.5 Procedure

Participants met the researcher in a quiet room at the Royal College of Music or Imperial College London and were presented with an information sheet and consent form. They were then introduced to the continuous measures software and encouraged to make and record their decisions as instinctively and intuitively as possible, emphasising that their decisions should be made not on the basis of their enjoyment of the performance but rather on the objective quality of the performance “as though they were a competition judge”. A brief (< 20-second) excerpt of a Beethoven piano sonata was used as a test piece, which the participants were allowed to repeat as many times as they wished until they felt comfortable with the input method. Following this, participants were told that they were about to hear several live performances by different undergraduate pianists—as opposed to studio, professional recordings so that the obvious performance errors would not seem implausible—and to rate the performance quality. For each trial, the name of the composer, the name of the work, and the length of the recording was presented orally to the participant. They were then able to start the first recording in their own time, and when it finished, they completed the first questionnaire. This procedure was repeated for each work in a randomised order with a questionnaire following each continuous measurement. Concerning the performance errors, participants randomly heard either the no error, error-start, or error-recap condition of the Etude and Waltz and either the no error or error-start condition of the Prelude; separate randomisation procedures were used for each work. The randomisation was established to favour conditions with no error to maximise opportunities to compare performances without such manipulations across the five works. Following the final trial, the second questionnaire was presented and participants were invited to give comments concerning the procedure as a whole. Each session lasted 30 - 40 minutes.

Due to time constraints, 12 of the 42 participants were presented only the three works containing variations in errors (i.e. the Etude, Waltz, and Prelude) following the same randomisation procedures described previously. These pieces were emphasised to maximise opportunity for between-groups examination of error placement, as the other 30 participants had rated the Tarantelle and Caprice but only 10 would have rated the no error, error-start, or error-recap versions of the other three works. The final n values for each condition are shown in Figure 3.1. These shorter sessions lasted approximately 20 minutes.

3.2.6 Data treatment and analyses

Data were treated to several operations, primarily following Thompson et al. (2007) in which three discrete variables were extracted from the full continuous data, along with the quality rating provided in the written comments:

1. Time to first decision, $\mathbf{T_1}$: As a brief amount of time was necessary to move the mouse to the desired first rating point, the moment the cursor entered the horizontal bar and data collection began was noted as the initial decision time, T_1. The continuous measurement ratings were measured from the moment the trial was started, yet the first note was not played until 4 seconds; therefore, 4 seconds were subtracted from each score, giving initial ratings made prior to the first note a negative time value.

2. First rating, $\mathbf{R_1}$: The first point at which the participant maintained a stable rating of at least 2 seconds was taken as the first rating.

3. Final rating, $\mathbf{R_2}$: The final continuous score reported formed the final rating.

4. Overall rating, **R_3**: The overall written score provided in the questionnaire on a scale of 1 - 7. When comparisons were made directly with continuous ratings, R_1 and R_2 were converted from 70-point to 7-point figures as per Thompson et al. (2007).

Three general approaches were taken to the analyses, requiring careful selection of subgroups and tests necessitated by the complex nature of the experimental setup. For analyses of scores that would not be affected by the presence of errors in the performance (i.e. familiarity and likeability), 5x2 factorial repeated-measures ANOVAs were conducted among the 30 participants who had rated a version of all 5 trials. For analyses of scores affected by the presence of an error (e.g. T_1, R_1, R_2, R_3) comparisons could only be made between participants who had heard an error-free version or, in measuring time to (T_1) or rating at (R_1) the first decision, between participants who had heard the error-free version or the error-recap where the error took place after first decisions had been recorded. Between-groups analyses of the error conditions were conducted using factorial ANOVAs. Planned repeated contrasts and t-tests were used to examine the four hypotheses as appropriate. Where Mauchly's W indicated a violation of sphericity ($p < .05$), Greenhouse-Geisser corrections are reported.

3.3 RESULTS

The first section of analyses examines familiarity and likeability levels of each of the works to validate assumptions of familiarity made in work selection and to define groupings for between-groups comparisons. This is followed by repeated-measures examinations of the five works to determine effects of familiarity, likeability, and composition length on time to first decision (T_1), final continuous rating (R_2), and overall written rating (R_3). Between-groups analyses are then used to determine effects of the error placement within the Etude, Waltz, and Prelude on the rating profile, and examine differences in the rating profile between the relatively unfamiliar Tarantelle and completely unfamiliar Caprice. The final section examines the influence of participants' perception of the difficulty of the works, musical experience, and listening preferences on the rating process.

3.3.1 Preliminary analyses: Establishing familiarity and likeability

The first stage of analyses involved defining reported familiarity and likeability levels of each of the works. As participants rated familiarity and likeability regardless of assigned error condition (and as they were asked to rate opinions of the composition itself, not of how it was performed), analyses could be conducted between all 30 participants who rated the five works. Table 3.2 shows descriptive values for the two dimensions, including correlations between likeability and familiarity for each piece (using Kendall's tau due to the smaller sample size and large number of tied ranks). While a matching overall trend from high to low familiarity and likeability can be seen across the compositions (see Figure 3.1), correlations between each pair varied across the pieces, with only the Etude and Prelude showing significant medium correlations between the two items. For the remaining three works, familiarity with the work was not necessarily indicative of the degree to which participants liked the piece. The low mean familiarity score and standard deviation (where a response of 1 indicated that the participant had never heard the work) for the Caprice resulted from the fact that 28 of 30 participants had indicated that the work was entirely unknown to them.

To examine overall trends, a 5x2 factorial repeated-measures ANOVA was conducted with work (Etude, Waltz, etc.) and rating construct (familiarity and likeability) as within-subjects variables. The ANOVA was followed by planned repeated contrasts in which the mean of each score was compared with that of the next (i.e. A versus B, B versus C, C versus D, D versus E). This was done as the works

Table 3.2. Familiarity and likeability ratings and correlations for each of the five works.

Work	*Familiarity* *Mean (SD)*	*Likeability* *Mean (SD)*	*Correlation* *τ (p)*
Etude	5.10 (1.90)	5.67 (1.21)	.46 (< .005)
Waltz	5.48 (1.86)	5.63 (1.40)	.09 (ns)
Prelude	2.70 (1.99)	4.73 (1.36)	.46 (< .01)
Tarantelle	2.28 (1.48)	4.72 (1.35)	.01 (ns)
Caprice	1.06 (0.25)	4.45 (1.66)	.06 (ns)

were chosen with a hypothesised pattern towards descending familiarity from the Etude to the Caprice and as likeability was predicted to follow a similar trend across the five works, both of which were confirmed by the descriptive values. Mauchly's W indicated a violation of sphericity ($p < .05$), thus Greenhouse-Geisser corrections were used. A significant main effect of work was found ($F_{(3.06,88.61)} = 34.53$, $p < .001$, $\eta_2 = .29$), resulting from the descending familiarity and likeability scores moving from the Etude and Waltz to the Caprice (see Table 3.2). The repeated contrasts showed that the descent was not uniform, however, with significant differences of familiarity and likeability (when combined) between the Waltz and Prelude ($F_{(1,29)} = 35.42$, $p < .001$, $r = .74$) and between the Tarantelle and Caprice ($F_{(1,29)} = 10.99$, $p < .005$, $r = .52$), but not between the Etude and Waltz or between the Prelude and Tarantelle (see Figure 3.2). A significant main effect of rating construct was found ($F_{(1,29)} = 55.24$, $p < .001$, $\eta_2 = .18$) where likeability scores were generally higher than those of the familiarity scores. These differences between constructs were not uniform, highlighted by the significant interaction between piece and construct ($F_{(3.11,90.16)} = 22.00$, $p < .01$, $\eta_2 = .08$). Once again, the planned repeated contrasts demonstrated that these interactions were only significant between the Waltz and Prelude ($F_{(1,29)} = 24.14$, $p < .001$, $r = .67$) and between the Tarantelle and Caprice ($F_{(1,29)} = 4.38$, $p < .05$, $r = .36$).

Together, these two sets of contrasts demonstrated three distinct groupings between familiarity and likeability scores in which the works were rated similarly: the Etude-Waltz pair, the Prelude-Tarantelle pair, and the Caprice (see Figure 3.2). The Etude-Waltz pair showed significantly higher scores overall (as demonstrated above) with no significant differences between familiarity and likeability, tested with multivariate simple effects tests using the estimated marginal means. The Prelude-Tarantelle pair showed lower overall familiarity and likeability, although both showed significantly higher familiarity scores than likeability scores with nearly identical effect sizes ($F_{(1,29)} = 44.43$, $p < .001$, $r = .78$; $F_{(1,29)} = 44.85$, $p < .001$, $r = .78$). Finally, the Caprice showed the lowest familiarity, with a significantly higher likeability score than its familiarity score ($F_{(1,29)} = 124.73$, $p < .001$, $r = .90$).

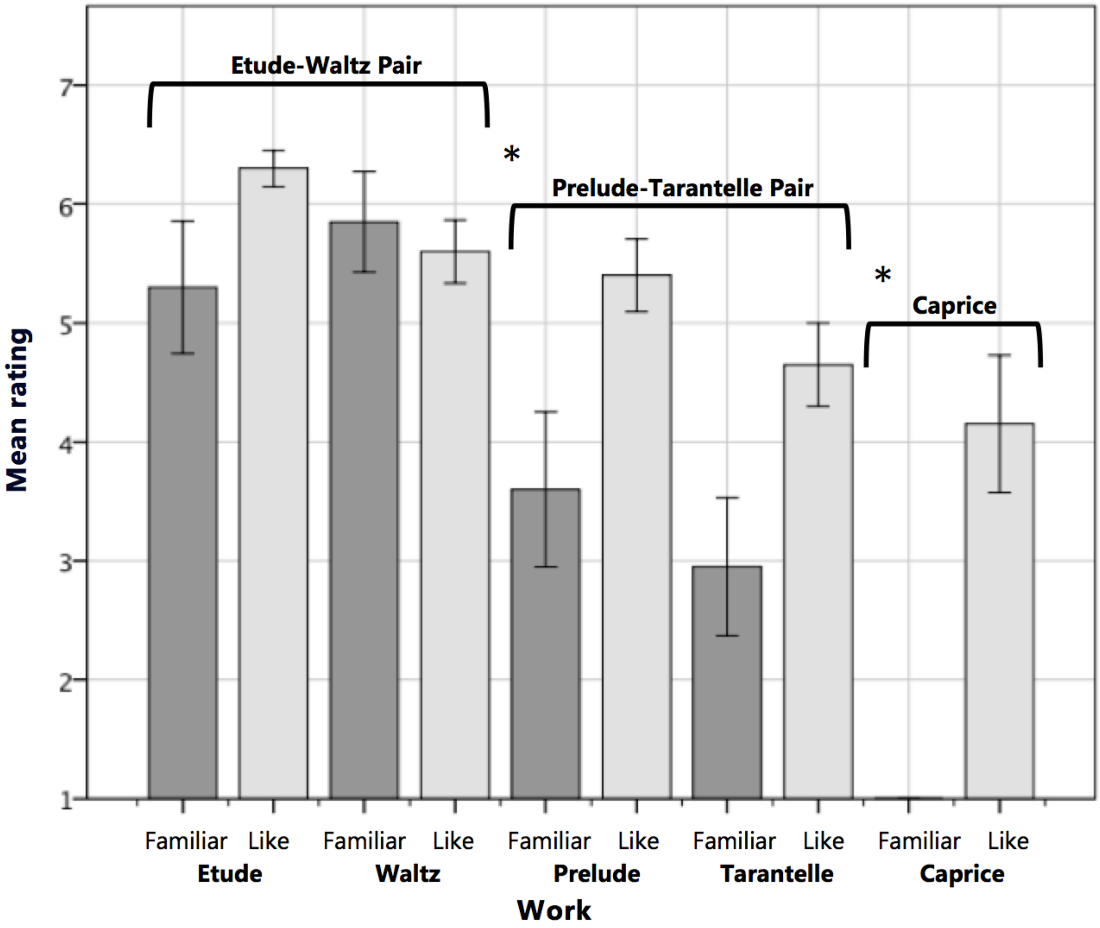


Figure 3.2. Mean familiarity and likeability scores for the five works. Three distinct groupings appeared: the Etude-Waltz pair showed high familiarity with no significant difference in likeability scores; the Prelude-Tarantelle pair showed significantly lower overall scores with significantly lower familiarity ratings than likeability ratings; the Caprice showed the lowest familiarity (approaching the minimum possible) with a significantly higher likeability score. Error bars show +/- 1 SE. * = p < .005, as tested using planned repeated contrasts in which the mean of each combined familiarity/likeability score was compared with that of the next.

With the Etude-Waltz, Prelude-Tarantelle, and Caprice familiarity/likeability groupings established, these were used for the basis of repeated-measures comparisons to test the relationship between familiarity/likeability and the time to first decision (T_1) as posited in hypothesis 1. Furthermore, the Prelude-Tarantelle grouping provided an opportunity to compare works of differing lengths while maintaining a consistent familiarity/likeability profile. This allowed for a direct examination of hypothesis 2, which predicted a decrease in time to first rating (T_1) for a work of shorter length.

Correlations (Kendall's tau) between time to first decision (T_1) and the first (R_1) and overall (R_3) ratings were conducted for each of the five works to test the assumption that any significant differences in time to first ratings were due to the nature of the works and not simply a result of differences in the perceived quality of the individual performances. Correlations remained very low ($\tau < .2$) and nonsignificant across the 10 tests, supporting this assumption.

3.3.2 Hypotheses 1 and 2: Repeated-measures effects of familiarity, likeability, and length on time to first decision (T_1)

To examine the effect of condition on the time to first decision (T_1) a repeated-measures ANOVA was calculated between the five works among the 11 participants who had rated all five performances without an error at the beginning. Despite the small sample size, a significant main effect of condition was found ($F_{(2.16,21.66)} = 5.20$, $p < .05$ $\eta_2 = .52$). Again, a planned reverse contrast was used to compare the differences between each condition and the previous condition, as was employed in the likeability/familiarity comparisons. The only significant difference was between the Tarantelle and Caprice where a mean 34.50 seconds (SD ± 24.93) was taken to first decision versus 15.50 seconds (SD ± 8.35; $F_{(1,10)} = 6.78$, $p < .05$, $r = .64$; see Figure 3.3). No significant difference was found between the other levels, although medium effect sizes were seen between the Etude and Prelude ($r = .29$) and between the Prelude and Tarantelle ($r = .38$; for reference, the Etude versus Waltz comparison showed $r = .04$) suggesting the descriptively shorter time to first decision for the Prelude (M = 12.90, SD ± 8.56 seconds) versus the Waltz (M = 16.27, SD ± 7.83 seconds) and Tarantelle (M = 15.50, SD ± 8.35 seconds) could represent a significant effect in an analysis with greater power.

While the small sample size afforded by the five-group (n = 11) test was able to reveal the relatively large difference between the Caprice and the remaining works, with participants taking on average twice as long to register their first judgement, the nature of the experimental setup allowed for larger sample sizes in focussed comparisons. Hypothesis 2 suggested that the shorter Prelude would result in shorter

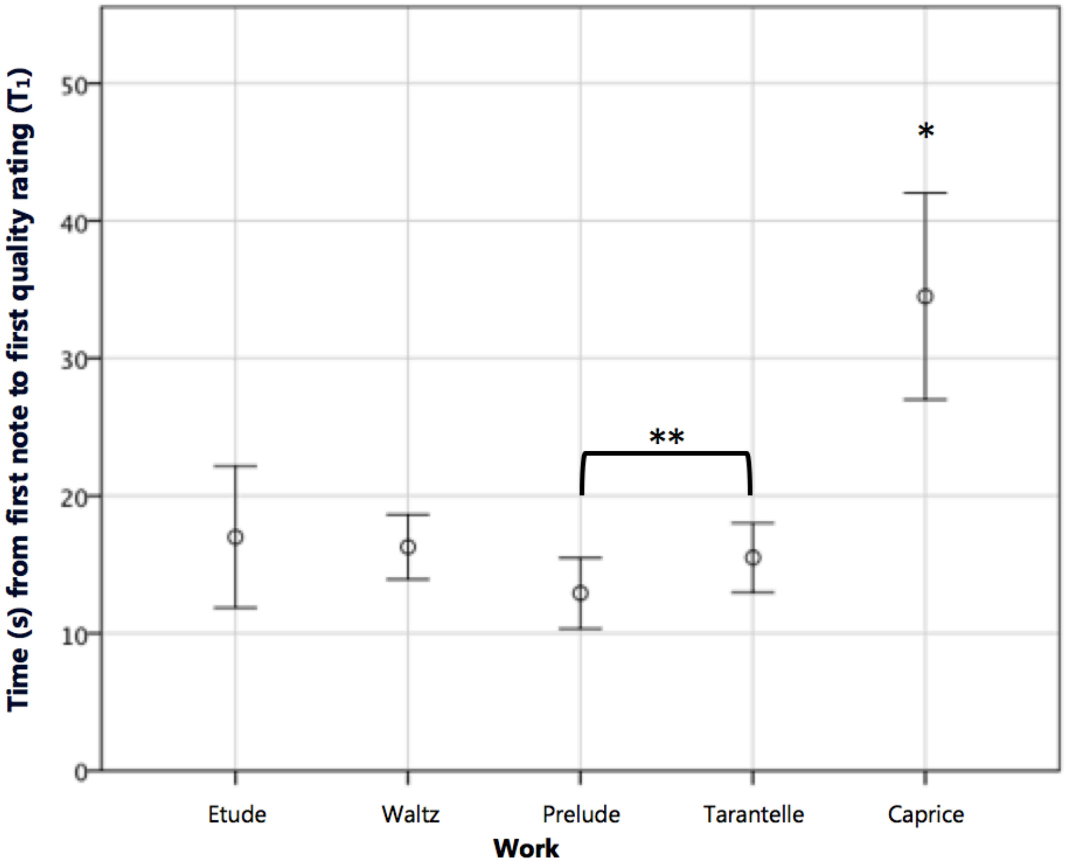


Figure 3.3. Mean time in seconds from the first note to first decision (T_1) for the five works. The Caprice resulted in a significantly longer time to first decision than the four stylistically similar works of Chopin in a repeated-measures comparison of 11 participants. * = p < .05 as tested using planned repeated contrasts in which each time was compared with that of the next. A further test between works of equal familiarity but differing length (the Prelude and Tarantelle) with n = 16 found a significantly lower time to first decision for the shorter Prelude (27 seconds in length) versus the Tarantelle (156 seconds in length). ** < .05 as tested using a Wilcoxon signed-rank test. Error bars show +/- 1 SE.

time to first decision than a work of equal familiarity, which was above demonstrated to be the Tarantelle and could be tested with a higher degree of power as 16 participants rated both the Tarantelle and the error-free version of the Prelude. This hypothesis was confirmed with a one-way related-samples Wilcoxon signed-rank test ($Z_{(16)} = 1.66$, $p < .05$, $r = .28$) with first decisions for the Prelude taking a mean 10.83 seconds (median = 7.75, SD ± 7.70) and for the Tarantelle a mean 13.38 seconds (median 10.25, SD ± 7.74). For comparison, a similar test run between groups of similar familiarity (the Etude-Waltz pair) showed no significant difference, despite an even greater availability of matched pairs (n = 20) and the corresponding increase in power.

Correlations between each of the familiarity scores for the Etude, Waltz, Prelude, and Tarantelle and their respective times to first decision (T_1) were tested using Kendall's tau; the Caprice could not be tested as only 2 of the 30 participants indicated they had ever heard the work. T values were low (< .10) and none approached significance, further suggesting a lack of relationship between familiarity and time to first decision among the stylistically familiar Chopin works. Examination of the likeability scores (which included the Caprice) also showed no significant correlations between how much one liked the work and the speed with which a first rating was made.

Overall, these analyses revealed a significant effect of work on the time taken to form a first decision. Participants rating the Caprice took significantly longer to form their first judgements, due perhaps to the unfamiliarity of the piece and its composer. This relationship between familiarity and decision time was not reflected among the stylistically similar Chopin works, although a significantly faster time to first decision was demonstrated within the shorter Prelude.

3.3.3 Comparisons of the final ratings (R_2 and R_3)

Direct comparisons of the final ratings in this study are complicated by the experimental setup, in which very few (n = 3) participants heard no error (i.e. uncontaminated) versions of all five works. While such comparisons were not the primary focus of the study due to its focus instead on the decision-making process, two of interest could be made: (1) whether final continuous scores (R_2) were representative of the final written ratings (R_3) and (2) individual correlations between familiarity, likeability, and the final scores within each work.

For the first comparison, the R_2 scores were converted to a 7-point scale as described in Section 3.2.6 allowing for direct comparison with R_3. A 5x2 factorial repeated-measures ANOVA was then calculated with work and rating condition (converted R_2 versus R_3) as within-subjects factors. A significant main effect of work was found ($F_{(4,112)} = 6.18$, $p < .001$, $\eta_2 = .16$), where final scores increased from the Etude as the lowest to the Caprice as the highest (see Table 3.3), unsurprising as these ratings included versions of the Etude, Waltz, and Prelude that contained performance errors. Crucially, no significant main effect of rating condition was found, or any significant interaction between work and rating condition. This suggests that the final continuous ratings (R_2) were reflected in the overall written scores (R_3) across all works, supporting the use of continuous ratings as an adjunct for standard written rating procedures and for using R_2 scores to examine the effects of error placement on final scores.

Correlations were tested between each of the familiarity and likeability scores for each of the works (again, correlations could not be checked with familiarity for the Caprice) and their respective final continuous ratings (R_2) using Kendall's tau. The strongest correlation, and the only one to reach significance following a Bonferroni correction for multiple comparisons, was a medium correlation between likeability and final continuous score for the Caprice ($\tau = .46$, $p < .01$). A linear regression between the two variables produced a significant model ($F_{(1,27)} = 9.76$, $p < .005$, $R^2 = .27$, $b = 3.10$) wherein an increase of one point on the 7-point likeability scale predicted a 3.10-point increase on the 70-point final continuous rating (see Figure 3.4).

Table 3.3. Mean final continuous scores (R_2), final scores converted to a 7-point scale, and overall written scores (R_3) for the five works.

Work	*R_2 (SD)*	*Converted R_2 (SD)*	*R_3 (SD)*
Etude	38.31 (15.52)	4.31 (1.58)	4.31 (1.55)
Waltz	39.93 (14.98)	4.48 (1.45)	4.41 (1.23)
Prelude	46.31 (9.51)	5.10 (1.01)	4.93 (0.80)
Tarantelle	44.66 (13.67)	4.93 (1.31)	4.84 (1.25)
Caprice	51.72 (10.03)	5.62 (0.98)	5.41 (0.81)

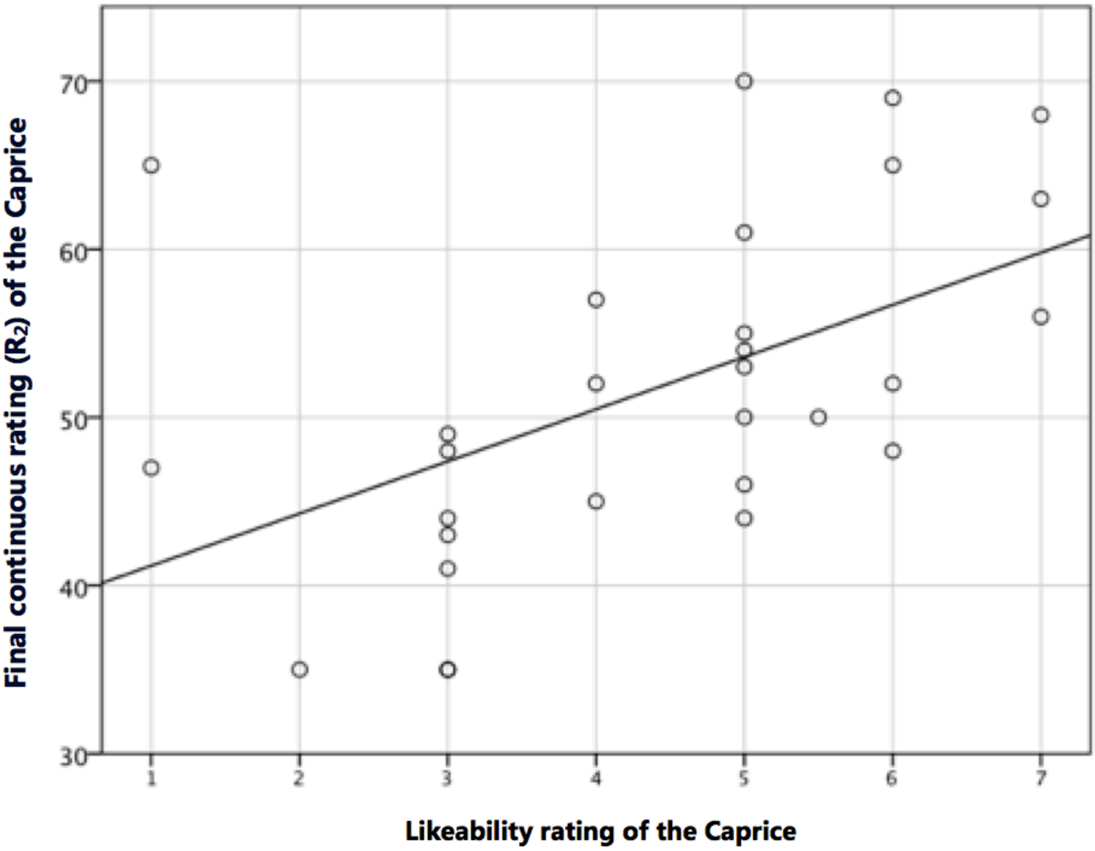


Figure 3.4. Scatter plot showing likeability score and final continuous rating (R_2) for the Caprice, wherein greater liking of the composition predicted a higher quality rating for the performance ($R^2 = .27$).

3.3.4 Hypothesis 3: Between-groups effects of the error on time to first decision (T_1)

In the cases of the Etude, Waltz, and Prelude, listeners were randomly assigned to a condition with no performance error (no error), a performance error in the opening seconds (error-start), or in the case of the Etude and Waltz, that same performance error at the recapitulation of the opening material (error-recap). This randomisation was not consistent for each work; a participant hearing a no error version of the Etude, for example, may have heard a start-error version of the Waltz. Thus, direct repeated-measures comparisons were not possible. Instead, the data offered the opportunity for an effective replication of the test with the same sample but a new stimulus and different randomisation.

Hypothesis 3 predicted that the time to first decision (T_1) for a performance would be lower in conditions with an error at the beginning when compared with those without. Thus, one-way ANOVAs were conducted for the Etude and Waltz with error condition (no error, error-start, and error-recap) as a between-subjects factor and T_1 as the dependent variable. These were followed by planned simple contrasts where each error condition was compared with the no error control. For the Etude, while no main effect of error condition was found, the contrast showed that the mean 6.36 seconds (SD ± 3.43) to the first decision in the error-start condition was significantly shorter ($t_{(39)} = -8.85$, $p < .05$, $d = 0.72$) than the mean 15.21 seconds (SD ± 17.11) in the no error control condition (see Figure 3.5A). This finding was replicated in examining the Waltz; the main effect of error condition was non-significant, but the contrasts again showed the mean 7.69 seconds (SD ± 4.94) to an error-start first decision was significantly shorter ($t_{(39)} = -13.51$, $p < .05$, $d = 0.63$) than the 21.21 seconds (SD ± 29.78) in the no error control condition (see Figure 3.5B). No significant differences were found between the error-recap conditions and the no error control in either work. As the Prelude was the shorter work, only two conditions (no error and error-start) existed and required testing. However, to maintain consistency in alpha inflation, ANOVA was also used to examine differences between the conditions. No significant main effect was found (see Figure 3.5C), influenced perhaps by the fact that the shorter length of the work already reduced times to first decision in the Prelude condition. Overall, these results support hypothesis 3; participants made their first decisions more quickly when an error was present in the opening seconds.

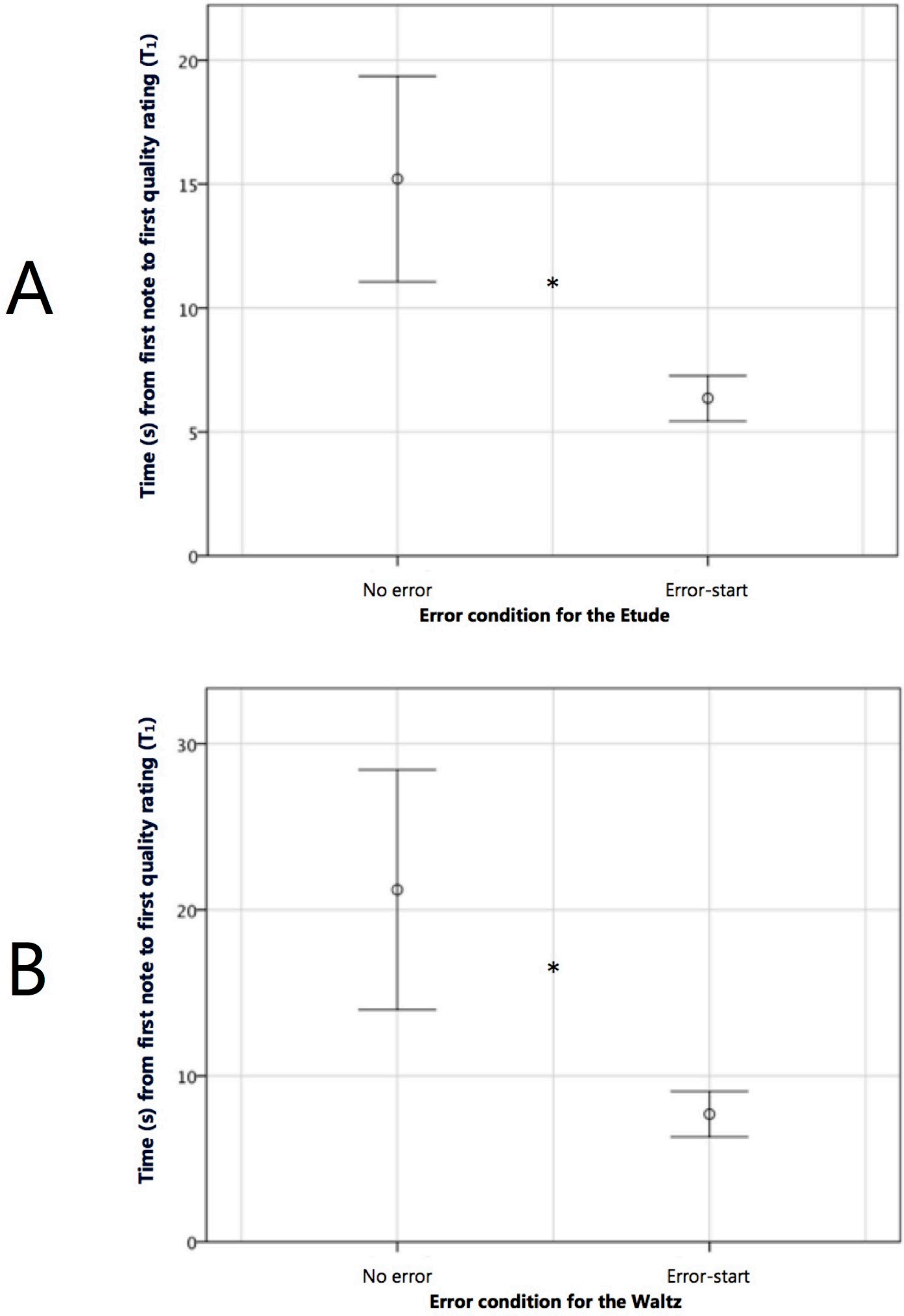


Figure 3.5. Continued on next page.

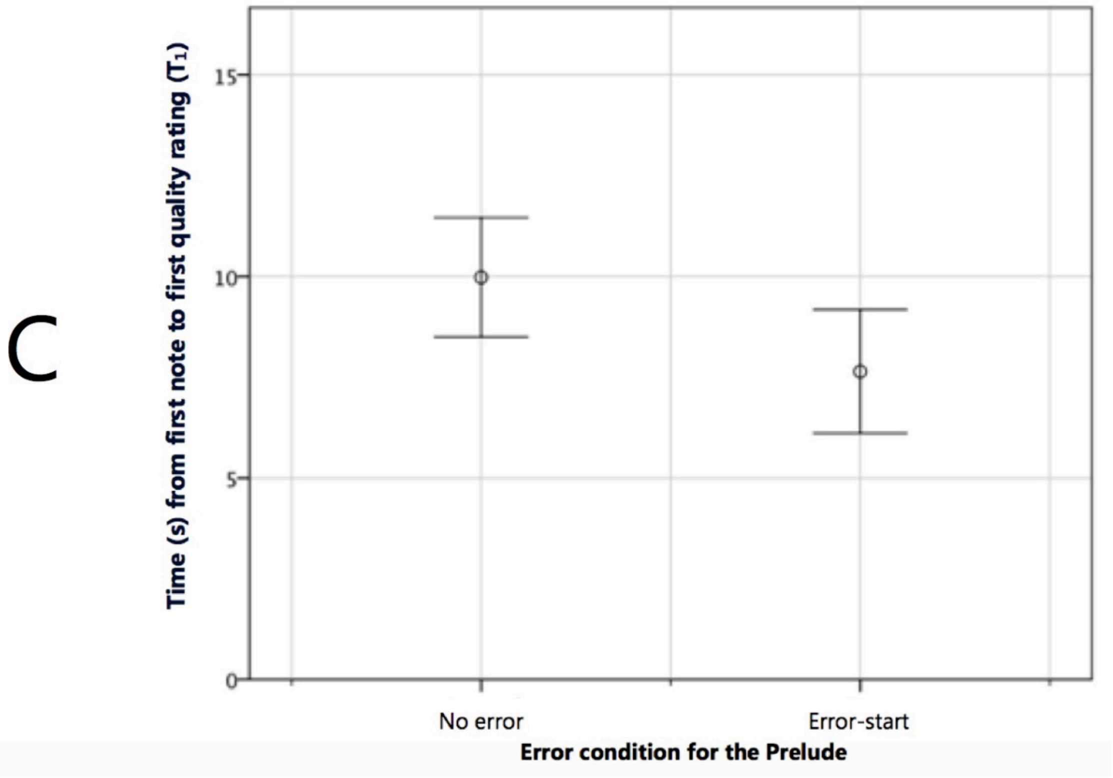


Figure 3.5. (Continued from previous page.) Mean time to first decision between the no error control and error-start conditions for the (A) Etude, (B) Waltz, and (C) Prelude. The Etude and Waltz showed a significantly shorter time to first decision when an error was inserted into the opening seconds of the performance; * p < .05 as tested with planned simple contrasts where each error condition was compared with the no error control. The Prelude did not show a significant difference. Error bars show +/- 1 SE.

3.3.5 Hypothesis 4: The effects of the errors on first and final ratings and continuous rating profile

The same approach as above could be taken for analyses of the rating profile, treating the Etude and Waltz as replications of the same study with different randomisation procedures. In this case, tests examined differences between first (R_1) and final (R_2) ratings–as the analyses above demonstrated that R_2 scores were representative of the final R_3 written scores–and how they were affected by the presence of errors. For the Etude and Waltz, differences in the overall rating profile were tested with a mixed 2x3 ANOVA in which the first and final ratings (R_1 and R_2) served as the within-subjects variable and the 3 error conditions (no error, error-start, and error-recap) the between-groups. A planned simple contrast was used to determine group differences between the error conditions in which the error-start and error-recap conditions were compared with the no error control.

In the case of the Etude, no significant repeated-measures effect was found, although a significant main between-groups effect of error condition was demonstrated ($F_{(2,39)} = 4.78$, $p < .05$, $\eta_2 = .20$) as was a significant interaction between rating and error condition ($F_{(2,39)} = 3.38$, $p < .05$, $\eta_2 = .14$). As can be seen in Figure 3.6A, this was due to the general downward trend of the no error and error-recap conditions and the upward trend of the error-start condition. The simple contrast confirmed that, while the error-recap performance did not differ significantly from the no error performance in terms of first and final ratings, the performance with an error at the beginning did ($t_{(39)} = -11.83$, $p < .05$, $r = .88$), prompting first ratings (M = 28.00, SD ± 13.81) well below those of the standard performance (M = 42.71, SD ± 11.56) and concluding with a narrower but still significant gap (M = 36.57, SD ± 16.40 versus M = 45.53, SD ± 13.98). Thus, when the error was placed at the beginning of the work, the evaluators penalised the performer with a significantly lower rating that did not recover to no error levels by the end of the performance. In the case of the performance with an error part way through, the continuous measures data revealed a sharp drop in ratings immediately following the missed notes, but interestingly, this deficit was ‘forgiven’ by the end of the work (see Figure 3.6A), with no significant difference in the final score. An observer seeing only the final ratings would have no indication that an error had been made.

An analysis of the Waltz replicated the overall finding but did so under different circumstances. While the between-group analyses again showed a significant difference of error condition ($F_{(2,39)} = 4.35$, $p < .05$, $\eta_2 = .18$), there was in this case an additional main repeated-measures effect of first-to-final rating ($F_{(1,39)} = 10.45$, p <

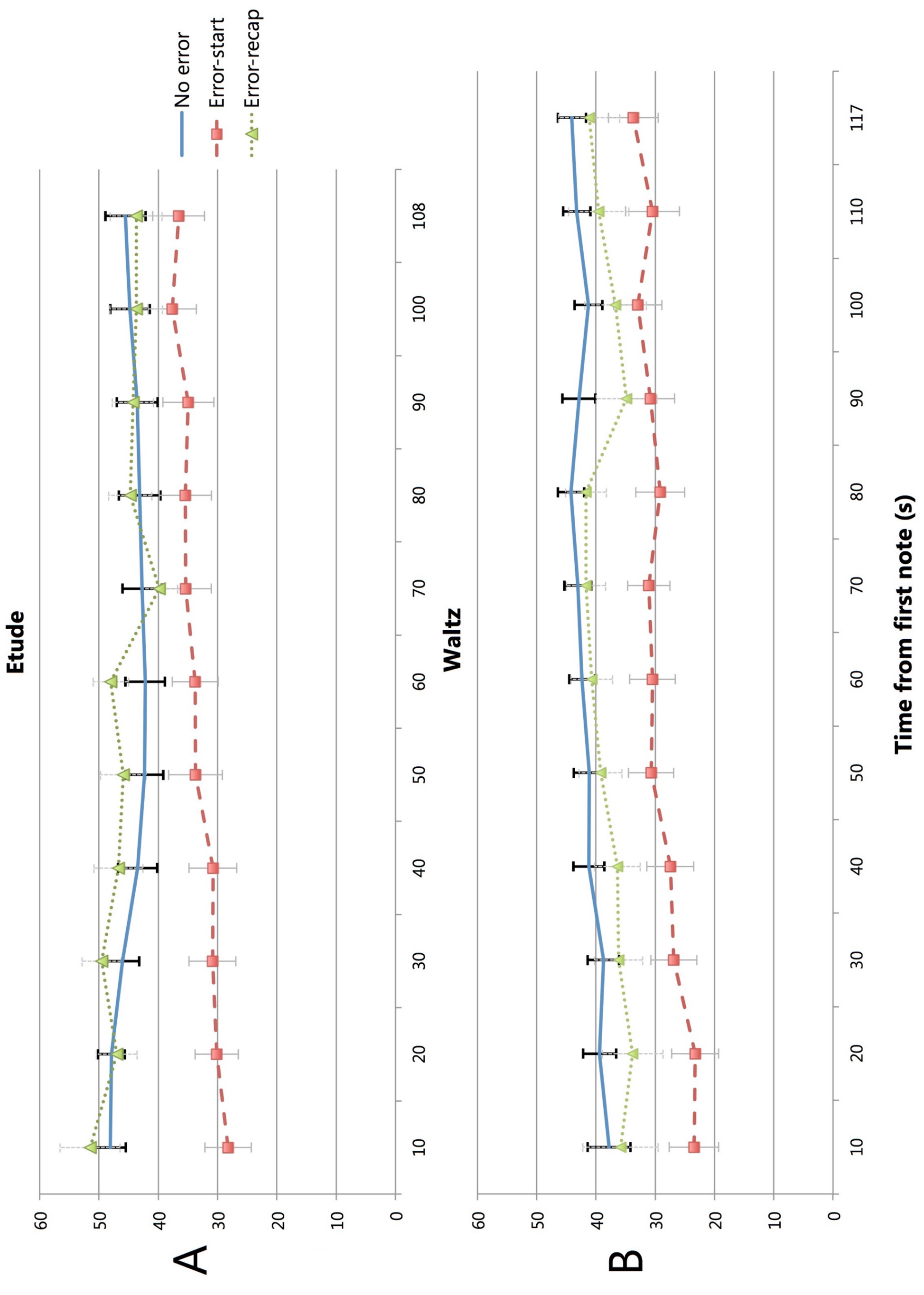


Figure 3.6. Continuous rating profiles of the no error, error-start, and error-recap conditions for the Etude and the Waltz, showing mean ratings at 10-second intervals. In both cases, the error-start condition resulted in a significantly lower first (R_1) and final (R_2) rating than the no error control. The error-recap condition resulted in a noticeable drop at the point of the error – between 60 and 70 seconds in the Prelude and 80 and 90 seconds in the Waltz – that recovered by the end of the performance, resulting in a final score (R_2) not significantly different from the no error control. Error bars show +/- 1 SE.

.005, $\eta_2 = .20$) and no significant interaction. The reason for this reverse of significant main and interaction effects can be seen in Figure 3.6B where all three conditions show a similar upward trend for the Waltz in contrast to the converging lines of the Etude (see Figure 3.6A) and thus show a significant overall increase in rating across the performances of the Waltz. However, the error-start condition once again lay significantly lower than the standard performance, confirmed by a significant difference from the standard condition shown by the simple contrast ($t_{(39)} = -12.10$, $p < .01$, $r = .89$) with lower first ratings ($M = 24.92$, SD ± 12.10 versus $M = 38.77$, SD ± 8.67) and final ratings ($M = 33.69$, SD ± 15.25 versus $M = 44.06$, SD ± 9.87). As with the Etude, the version of the Waltz with an error mid-way through, despite again causing an immediate drop in rating at the point of the mistake, did not differ significantly from the standard performance in terms of first or final ratings (see Figure 3.6B).

Regarding the Prelude, no significant main effects of rating or condition, or interactions between them, were found as a result of the error at the start. This mirrors the previous section, where the error also failed to affect time to first decision in the Prelude despite a significant effect within the Etude and Waltz. This suggests that the error itself may not have been dramatic enough to cause a reaction in the Prelude. For the Etude and Waltz the results are clear: an error in the opening material caused a shorter time to first decision and a lower initial rating that never fully recovered, where an error mid-way through caused a temporary drop that was not significantly reflected in the final ratings.

3.3.6 Hypothesis 1 revisited: The effects of familiarity on continuous comparisons of the Tarantelle and Caprice rating profiles

As 30 participants provided continuous ratings of both the Tarantelle and Caprice, and as analyses of time to first decision (T_1) demonstrated a different rating process between the two works in the greater amount of time taken to form a first decision, similar continuous analyses could be conducted to further test hypothesis 1, which predicted that familiarity would affect the time to form a final decision. To determine the point at which the cohort reached a final consensus on the two works, scores at 10-second intervals from the beginning of the performance were extracted and analysed to determine the point at which raters' responses did not differ significantly from their final scores. Repeated-measures ANOVAs were calculated for each work followed by reverse simple contrasts comparing each 10-second mean score with the final, beginning with the interval at which at least 50% of the participants had first reported (thus providing full datasets for analysis): this was the 10-second mark for the Tarantelle (with 15 respondents) and the 20-second mark for the Caprice (with 20 respondents). For the Tarantelle, the overall effect was not significant, although the contrasts showed a significant difference between the final score (M = 49.07, SD ± 10.54) and both the 10-second point (M = 43.93; SD ± 10.56; $F_{1,14}$ = 10.33, p < .001, r = .65) and 20-second point (M = 43.87, SD ± 11.11; $F_{(1,14)}$ = 6.43, p < .05, r = .56), with no significant difference from the end from the 30-second point onward. In the case of the Caprice, a significant main effect of the ANOVA was found ($F_{(3.51,66.78)}$ = 9.48, p < .001, η_2 = .33) and the contrast revealed significant differences between the 20-80-second points and the final score, with no significant results following. As can be seen in Table 3.4, effect sizes at the cut-off are still moderately strong, but using the significance value as a conservative cut-off, these results suggest a time to final group decision at least 3 times longer in the Caprice than the Tarantelle (see Figure 3.7).

Table 3.4. Mean performance ratings for the Caprice at 10-second increments from the beginning of the recording, with results from a repeated-measures ANOVA comparing each score with the final continuous rating.

Time (s)	*Mean*	*SD*	*F*	*p*	*r*
20	43.95	8.94	20.97	.000	.72
30	45.65	10.04	16.75	.001	.68
40	46.20	10.56	14.90	.001	.66
50	47.70	10.19	9.00	.007	.57
60	48.20	10.56	11.52	.003	.61
70	48.80	10.46	8.76	.008	.56
80	49.55	11.10	5.81	.026	.48
90	50.00	11.26	3.71	.069	.40
…					
Final (152)	52.50	10.10			

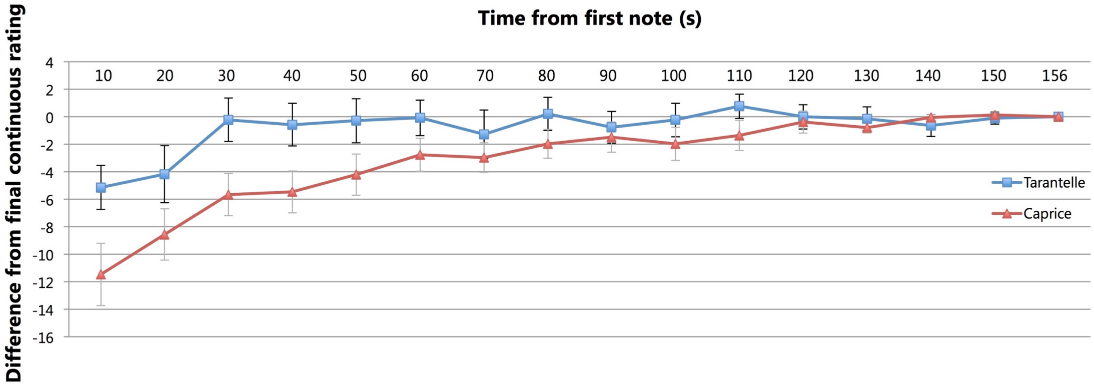


Figure 3.7. Continuous rating profiles of the Tarantelle and the Caprice. Data are normalised to show mean difference from the final score at 10-second intervals. Using a reverse simple contrast, the Tarantelle showed no significant difference from the final score from 30 seconds onward, whereas the Caprice showed no significant difference from 90 seconds onward. Final time for the Caprice was at 152 seconds. Error bars show +/- 1 SE.

3.3.7 Correlations with experience, difficulty, and listening preferences

Further tests were conducted to determine whether years of musical experience, perceived difficulty of the work, typicality of the performance, and listening preference (Romantic when examined against the Chopin works, twentieth-century when examined against the Caprice) correlated with time to first decision (T_1) or final continuous ratings (R_2). The only significant correlations (after correcting for multiple comparisons across the five works) were between perceived difficulty of performance and final continuous score (R_2) for the Etude ($\tau = .38$, $p < .005$) and the Caprice ($\tau = .40$, $p < .01$), where higher difficulty scores correlated with higher performance ratings. This relationship showed small but non-significant correlations across the other three works.

3.4 DISCUSSION

The purpose of most music performance quality assessments, whether conducted as part of an audition, recital, competition, or examination, is to determine the quality of the performance and performer. They are not intended to be an assessment of the quality of the work being performed, at least not in most Western classical contexts where the composer and performer of the work are separate entities. Otherwise, music competitions intended to identify a top performer would become repeating debates over the relative merits of Mozart and Haydn or of Beethoven's Op. 110 and 111. This study questioned this assumption, examining how qualities related to the repertoire – such as its length, familiarity, and likeability – affected the process by which assessments are formed. It also examined the nature of performance errors, and whether an error placed at the beginning of the performance had the same effect as the same error placed mid-way through the piece. To achieve this, trained musicians evaluated recordings of five works, selected to vary in familiarity and length, using a continuous measures methodology and standard written questionnaires. Furthermore, three of the works were manipulated to create conditions with performance errors at the beginning of the performance, and two of those manipulated again to have errors mid-way through the performance. The continuous measures approach revealed effects of these variables that could not have been seen in the standard written evaluations which followed, allowing for direct examination of each of four hypotheses set out in Section 3.1.5.

3.4.1 Hypotheses 1 and 2: Familiarity and length affect first decision time

The first hypothesis predicted that works of lesser familiarity would result in an altered time to first decision and that this would be exaggerated for a work of unfamiliar tonal structure and composer. This hypothesis was partially confirmed: within performances by a familiar composer (Chopin), relative familiarity and likeability had no effect on or correlation with time to first decision (T_1). However, for the unknown work by the unfamiliar composer, the first decision took significantly longer. Furthermore, the rating profile for the Caprice showed that the group took three times longer to settle on their final decisions than they did for Chopin's Tarantelle of equal length and that the likeability of the Caprice showed a medium correlation with the final continuous score (R_2). The second hypothesis predicted that a work of shorter length would also result in a shorter time to first decision. This was confirmed, wherein the 27-second long Prelude resulted in a significantly shorter T_1 score compared with the 156-second Tarantelle, which matched in familiarity and likeability ratings.

Overall, these results support previous findings that time to first decision takes place within an average window of 15 seconds when rating audio recordings of performances of high musical quality (Thompson et al., 2007). The present results go on to demonstrate that the time to first decision can vary. Here, the unfamiliar nature of the work and composer led to a twofold increase. This could suggest that the listeners needed more time to orient themselves to the work and determine their criteria for assigning performance quality. Alternatively, the unfamiliar nature of the work could have taken attentional focus away from the task at hand. Moreover, a shorter work resulted in a decrease in time to first decision. This supports the findings of Tullar et al. (1979), who found that the decision-making process took longer when assessors were informed that job interviews would be longer. This suggests that assessors accelerate the decision-making process when they are aware that they will have less time to conduct it. Anecdotally, the participants in this study often expressed visible and/or verbal surprise when informed that the work they were about to assess was less than 30 seconds in length; many seemed aware that this was a relatively rare situation in rating full performances of standard repertoire and perhaps prepared themselves accordingly.

The positive correlation between likeability and final quality ratings in the unfamiliar Caprice raises interesting questions about reactions to a completely unfamiliar performance, as the finding was not replicated in the other works where familiarity scores were higher and rating processes (represented by the time to first decision) were unchanged. As this is a correlational finding, the direction of causality can only be speculated upon, although the fact that the finding was not replicated among the more familiar works suggests that it was not the case of participants being unable to separate the constructs of likeability and performance quality or having a third variable (e.g. tendency to provide generally higher responses on the rating scales) influencing both. It may be that, when orienting oneself to an unfamiliar work in an unfamiliar style, one’s enjoyment of the work itself influences the interpretation of performance quality. Alternatively, those that felt the work was performed better may have developed a stronger liking for the composition itself. Further work is required with other unfamiliar compositions to determine whether this is a generalisable effect, as well as the direction of causality.

3.4.2 Hypotheses 3 and 4: Performance error locations

Hypotheses 3 and 4 concerned the placement of performance errors, predicting that an error in the opening seconds of a performance would both reduce the time to first decision and result in a significantly lower final quality rating when compared with a performance with no error or an error in the middle. The continuous ratings confirmed both. Time to first decision was shorter for both the Etude and Waltz when the initial error was present, and while a decreasing trend was seen for the Prelude, it was not significant. For both the Etude and the Waltz, the error-start condition caused a significantly lower first and final rating than the no error control, and the continuous measurement profile demonstrated that the error-recap condition, while not differing from the control in terms of first or final decision, caused an immediate negative reaction to the error that recovered by the end of the performance. No effect of the error on ratings was seen for the short Prelude. Thus, participants were more temporally reactive to negative than positive (or at least neutral) information in the opening moments of the performance. That this effect was not replicated within the Prelude could be explained by the corresponding lack of significant effects on the first and final quality ratings; it could be that the error itself was not as easily perceived or considered as serious as the error in the other two works.

The effects of the errors at the start and middle of the Etude and Waltz were dramatic, demonstrating that the temporal location of an otherwise identical error matters. This provides strong support of Ybarra's (2001) findings that it is difficult to reverse judges' negative first impressions. In this study, the significantly lower first ratings did recover over time, but never reached the height of the final score in the no error conditions. There are at least two possible explanations for these findings. It may be that the low quality of the opening seconds caused an anchoring effect in the listener, whereby the remainder of the performance was perceived as being of lower quality and was rated as such, with the perceptual effect of the error gradually fading. Alternatively, the listeners may have perceived the quality of the rest of the performance as high as those rating the no error condition, but their moment-by-moment continuous rating represented an overall decision reflecting both the current material and that which has come before it. The lack of a significant difference between final continuous (R_2) and overall written (R_3) scores in this study supports the latter explanation, as it suggests that an extract of a moment-by-moment continuous rating emulates the same performance-averaging result as provided when a judge is asked to give an overall quality score. This continual comparison is also supported by research examining evaluations of affective experience that show global evaluations can be best predicted by an averaging of extreme peaks in rating and the material in recent memory (e.g. Fredrickson & Kahneman, 1993; Varey & Kahneman, 1992). Retrospective ratings of pain, for example, have been found to correlate most strongly with the point of highest pain intensity and the intensity during the final stage of the treatment, not reflecting the duration of treatment or accumulated pain ratings (Redelmeier & Kahneman, 1996). The question remains as to whether, given enough time, ratings of performances with an initial error could eventually recover regardless of their severity. Future studies could examine the effect in pieces of significantly longer length; the classical repertoire offers examples of works that are hours long. They could also examine how the presence of errors in one performance affects the ratings of subsequent performances by the same performer, as participants in this study were informed that they were rating different pianists. The role of musical structure may also be important. Perhaps those hearing the mistake at the recapitulation were more forgiving because they had already heard an example of the performer navigating that exact passage correctly at the beginning of the piece. On the other hand, those hearing the mistake in the introduction did not obviously reward the performer for avoiding the error later on.

The examination of the errors in the present study focussed only on the works of higher familiarity in a recognisable tonal style, in which an error could be easily perceived as a harsh dissonance. This raises questions about the nature and perception of performance errors within a contemporary work that lacks the familiar tonal frameworks of standard Romantic repertoire. It is interesting to note the similarities between the continuous data for the Caprice and those for the Etude and Waltz when the error was inserted at the beginning of the performance. It stands to reason that participants took significantly longer to come to their first decision when rating the Caprice; as discussed above they had to acclimatise to an unfamiliar style. However, when that first decision was eventually made it was established, on average, at a point significantly lower than the final rating, gradually increasing across the length of the performance to reach the highest mean final rating of the five works (see Table 3.3 and Figure 3.7). This was mirrored in the error-start conditions. While the comparison is cursory in this case, one could hypothesise that participants initially could not determine whether errors were being made and, hearing the repeated extra-tonal dissonances, rated it as though they were. It could also be that the performance was of genuinely lower quality at the beginning, although this would contradict the performer's reported intention and perception of a polished performance throughout and one that was true to the notated score. Interpretation is limited by the fact that only one such composition was investigated in this study. As discussed above, future work should examine whether this effect is generalisable. It should also compare ratings of unknown works with a specialised cohort familiar with its structure, language, and style, either through prior experience or an experimental intervention. A growing body of research has demonstrated the ability of listeners to remember and perceive errors in non-tonal contexts when first given an accurate reference (e.g. Dienes & Longuet-Higgins, 2004; Samplaski, 2004; Ockelford & Sergeant, 2013; Kuusi, 2015).

3.4.3 Directions for future research

While the present study was conducted in laboratory settings with digitally manipulated stimuli, every effort was made to replicate the experience of rating audio recordings of genuine performances as a juror might be asked to do in an audition or competition setting. Nonetheless, care must be taken in determining the degree to which these results are generalisable to live performance and evaluative settings when the environment is under less experimental control. Furthermore, the sample in question represented a relatively homogenous group of musical expertise. A wider sampling, especially one allowing for between-groups comparison, would be able to determine whether the present results apply to a wider population. In particular, whether non-musicians' differing knowledge, and thus expectations, of the piece might alter the degree to which they perceive and respond to performance errors. Limitations of the use of a single 'familiar' composer and a single wholly unfamiliar work have been discussed above, but it bears repeating that expansion of this research to a wider repertoire base will be required to determine whether the effects demonstrated are replicable.

With these caveats in mind, there remain several points of which musicians can take note. The nature of their repertoire, whether its length or its familiarity, can affect the process by which their performances are judged. In particular, unfamiliar works may cause their audiences to take longer to orient themselves to the performance and be more critical in their initial judgements of quality. In addition, the adage that 'first impressions count' appears to hold true. Performers are well advised to ensure that, if nothing else, the opening seconds of their performances are as prepared and polished as possible. Otherwise, a few misplaced notes could tarnish judgements of the thousands that follow.

3.5 SUMMARY

This chapter examined the effects of composition length, familiarity, and likeability – as well as the location of performance errors – on the process of forming performance quality ratings. Forty-two musicians provided continuous and final quality ratings of five works varying in length and familiarity, several of which were manipulated to contain performance errors. The use of continuous measures to examine the performance evaluation process revealed findings that could not have presented if examining the traditional final ratings alone. Familiarity with the repertoire had no effect within works of a well-known composer, but times to first and final decision were significantly extended for an unfamiliar work of an unfamiliar composer. A shorter piece led to a shorter time to first decision. An error at the beginning of a performance caused a shorter time to first decision and lower initial and final ratings, where the same error at the recapitulation did not have a significant effect on the final judgement, despite causing a temporary negative drop.

4 STUDY 2: THE PERFORMER

4.1 INTRODUCTION

Central to any performance, and any evaluation thereof, is the performer themselves. It is their ability that the assessor seeks to quantify, and it is their technical, expressive, and communicative skill and interpretative intent that sets one performance apart from another. Nonetheless, there are numerous qualities of the performer that may be considered extraneous to the performance. As discussed in Chapter 1 (see Sections 1.6 and 1.6.2), many of these properties are communicated visually, such as race (Elliott, 1995; Davidson & Edgar, 2003; VanWeelden, 2004), dress (Griffiths, 2008, 2010, 2011), attractiveness (Wapnick et al., 1997, 1998, 2000; Ryan & Costa-Giomi, 2004; Ryan et al., 2006), and sex (Davidson & Edgar, 2003). The performer's physical behaviour is also expressed through this modality, both those movements necessitated by activating the instrument and those facilitating the communication of emotional and expressive intent (Thompson et al., 2005; Dahl & Friberg, 2007). That such visual information relating to the performer influences performance quality ratings is, at this point, unquestionable. Platz and Kopiez' (2012) meta-analysis demonstrated a global effect (*d* = 0.51 SDs) of visual variables across performance quality, expressiveness, and appreciation ratings. Recent work continues to strengthen the case, finding that when audio and video material of professional and amateur performers was juxtaposed incongruently, the resulting quality evaluations more strongly reflected the visually-presented ability regardless of the experience of the evaluator (Griffiths & Reay, 2018).

Where Study 1 of this thesis employed continuous measurement methodologies to examine ratings of audio-only recordings as they unfolded over time, the present study returns the performer to the centre of the evaluative process by including the visual variable in this process. Several existing studies have focussed on temporal aspects of visual information. Tsay (2013) gave participants 6-second clips of the three finalists in international piano competitions and asked them to identify the jury's top performer in each case. When provided with either audiovisual or audio-only information, the participants did no better than chance at selecting the winner, irrespective of musical training. However, those who were provided silent video clips identified the winner at a rate significantly higher than chance, a finding that was replicated with a second study using orchestral performances (Tsay, 2014). A key feature of Tsay's research was the use of very brief excerpts, forcing participants to form snap judgements of the recorded performances. The question remains as to whether the immediate influence of these visual features will persist over the course of an entire performance. This has been examined with the use of excerpts of varying lengths, although not with full performances and with conflicting findings. In supplementary studies, Tsay (2013) replicated her primary results using excerpts ranging from 1 to 60 seconds in length, suggesting that the effects may not be time-dependent. Research by Wapnick and colleagues (2009), however, found that the effects on ratings of some extra-musical visual attributes (attractiveness, dress, and stage behaviour) varied as a function of excerpt duration (25, 55, and 115 seconds), although results were inconsistent between attributes and performers' sex. For example, high attractiveness significantly increased ratings for women only and only in the 25-second excerpts, while dress affected ratings for men only in the 25- and 115-second (but not the 55-second) excerpts.

Combining the temporal nature of music with the visual modality provides numerous opportunities for the study of the evaluative process as it unfolds, for "at a basic level, visual information often signals the timing of musical events, focusing listeners' attention to (or away from) critical acoustic information at specific moments in time" (Thompson et al., 2005, pp. 203 - 204). As such, to determine what of countless features of the performer could be examined, this study focussed on two variables intrinsically linked to visually-communicated performer behaviour and tied to particular temporal points in the performance. First, the performer's stage entrance. Second, and following on the results of Study 1, whether the reaction to a major performance error is affected by the facial reaction of the performer who committed the mistake.

4.1.1 Performers' stage entrances

One method of examining the long-term effect of visual information is by examining cues that are specific to one point in the performance, thus allowing for a residual effect to be studied after the cue is presented. The stage entrance provides such an opportunity, marking the time from when the performer emerges into the audience's field of view to the production of the first note, often incorporating a bow, acknowledgement of applause, and a brief preparation of the instrument (e.g. tuning, adjusting the seat). No music is being produced, thus any effect on evaluation of the subsequent musical material emanating from the stage entrance can be linked entirely to visual features. Platz and Kopiez (2013) compiled an inventory of 141 stage-entrance features drawn from previous studies, interviews with a small concert audience, and transcriptions of an acting tutor's commentary on select entrance videos. As stimuli, 27 videos of stage entrances were extracted from an international violin competition and manipulated to ensure consistent ambient audience noise (including applause) across conditions. Through appropriateness ratings of each video's entrance behaviour on a 5-point scale by 435 participants across two preliminary studies, the corpus of 141 features was reduced to 56 and then to 10 salient behaviours via probabilistic test theory and item response theory models. In the final study, 1002 participants rated the appropriateness of these 10 items while viewing 12 of the videos of entrance behaviour and then indicated whether they would like to continue watching the ensuing performance. Of the 10 behaviours, six were found to be the most salient to judging the appropriateness of a stage entrance: nodding, direction of gaze, touching oneself, stance width, step size, and making a resolute impression. While participants were not asked what specific nature each of these items should take, one can infer from the initial item set and attributes of the performance videos that participants favoured some nodding of acknowledgement directed at the audience, not too many (nor to few) shifts of eye direction, minimal touching of one's own body, a stance and step size of moderate proportions, and a high degree of resolute confidence. High-scoring entrances correlated positively with the viewer's motivation to continue watching. This suggests that the process of performance evaluation had already begun with the stage entrance and may have influenced perception of the musical content itself, although as the videos were stopped before the first note sounded, the effect on musical perception was not explicitly examined.

4.1.2 Performers' facial reactions

Performers' facial reactions to specific performance events can also provide dramatic visual markers. The role of facial expression in music performance has been given greatest attention among singers, where studies have found their expressions to aid in lyric comprehension (Jesse & Massaro, 2010), to alter pitch perception (Thompson et al., 2005; Thompson et al., 2010), to indicate musical phrasing (Ceaser et al., 2009) and to enhance emotional expression (Quinto et al., 2014b; Livingstone et al., 2015; Thompson et al., 2008). However, facial expression has been experimentally examined far less in instrumentalists. Thompson et al. (2005) demonstrated that body and facial movements by blues guitarist B.B. King increased ratings of perceived aural dissonance by participants. In the context of the musical genre (the blues), this dissonance is expected, if not desired, thus the expression enhanced its effect. How then might facial expression influence the perception of inappropriate and unintended aural dissonance, such as when an explicit performance error has been made? Errors of pitch and timing are not considered trivial in the classical music tradition, although Repp (1996) found that only a relatively small percentage of errors in pianists' performances were noticed, even among highly trained listeners. This is to the performer's advantage; the goal should be to avoid drawing attention to a misplaced note or, if it has been detected, not to emphasise its importance. Even when errors are noticed, the results of Study 1 of this thesis found that a musically-trained audience may forgive a mistake committed part way through a performance, leading to a final rating no different than that had the mistake not occurred. How might this change if attention is drawn to the mistake by a negative facial expression? The role that the performer's face might play in this process has not been systematically investigated.

4.1.3 Aims of the present study

The aim of the present study was to examine the influence of visual cues on participants' quality ratings of music performances. It first examined stage entrance behaviour, following the work of Platz and Kopiez (2013), to determine whether 'appropriate' and 'inappropriate' entrances indeed affect the perception of the musical content that immediately follows and whether such an effect lasts throughout the performance. It then examined the presence of facial reactions to a severe performance error. To determine the degree to which reactions to these visual variables were mediated by pre-existing expertise and expectations, performances were evaluated by representative samples of musicians and non-musicians, differentiated by their level of musical experience. From the existing literature, the following study-specific hypotheses were posited:

1. The presence of an 'inappropriate' stage entrance would cause a lower initial rating when compared with the same performance with an 'appropriate' entrance. This first rating would also be made sooner, as a result of the performers' deviation from expected stage entrance behaviour. No hypotheses were drawn concerning the degree to which any initial effect of the stage entrance would persist through the performance and affect the final ratings.
2. As musically trained evaluators would have a stronger heuristic for 'appropriate' stage entrances based on their extensive experience, they would show a shorter time to first decision and lower initial rating than the non-musician group.
3. The addition of a severe performance error would cause an immediate decrease in performance ratings, measured from pre-determined points before and after the inserted error and the effect on the final rating when compared with a

control performance. A corresponding negative facial reaction would intensify this response.

4. As with hypothesis 2, musicians' reactions to the error would be more severe than non-musicians' as a result of stronger expectations.

Testing these hypotheses required the measurement of participants' reactions to the performances as they unfolded, thus participants provided responses in real-time in addition to completing overall, *post hoc* quality ratings as was done in Study 1. To achieve this, a bespoke rating tool was developed that allowed for the collection of continuous data in tandem with presented video. In order to maximise ecological validity, full performances were used that, despite experimental manipulations, gave the impression of live, undoctored performances.

4.2 METHODS

4.2.1 Participants

Participants (N = 105) with and without musical training were recruited via email and in person from conservatoires, universities, and public music and science festivals held in southeast England. Musicians (n = 53: 28 men, 25 women, mean age = 27.38, SD ± 12.16 years) were defined as participants currently undertaking undergraduate music training (n = 27), those completing or holding postgraduate music training (n = 23), and/or practicing professional musicians (n = 18). Participants not meeting these criteria were classified as non-musicians (n = 52: 31 men, 21 women, mean age = 30.82, SD ± 16.23 years), which included amateurs without specialist training (n = 30), participants who had undertaken some undergraduate training in music but did not currently practise (n = 6), and those who did not play an instrument or sing (n = 16), thus representing a variety of musical engagement. Primary instrument families represented across groups were piano (n = 30), string (n = 16), guitar (n = 11), woodwind (n = 11), voice (n = 7), brass (n = 6), and other (n = 6). The musician group had greater exposure to visually presented (live or recorded) classical performances, with 81% viewing at least monthly, in contrast to just 31% of non-musicians (13% of non-musicians reported never seeing performances). This study was conducted according to the ethical guidelines of the British Psychological Society following internal Royal College of Music (RCM) approval on behalf of the Conservatoires UK Research Ethics Committee. Informed consent was obtained from all participants, and no payment was given in exchange for participation.

4.2.2 Stimuli

To maximise the ecological validly of the stimuli, recordings were created that would give the impression of a genuine live performance. Chopin's *Aeolian Harp* Etude (Op. 25, No. 1) was chosen as the work to be performed due to its short length (approximately 3 minutes), its familiarity to Western classical audiences, and its homogenous structure: the composition features a perpetual-motion texture that is maintained throughout. Therefore, a brief break and resumption of that texture would be easily perceived by non-musicians as a severe unintentional error, similar in effect to a layperson with no knowledge of figure skating technique recognising the severity of rare but occasional cases of professional skaters falling to the ice. A postgraduate pianist at the RCM performed the work in the RCM's Concert Hall on a grand piano. The lighting, staging, and performer's dress reflected a live concert experience. Audio was recorded via two Schoeps MK41 microphones hung above the stage, and video was recorded through two remotely controlled Panasonic AW-HE50 cameras.

Musicians have been shown to be highly sensitive to audiovisual asynchronies when viewing recordings of musicians with their hands in frame, particularly of their own instrument type (Bishop & Goebl, 2014). Therefore, footage of genuinely synchronised aural/visual information with the hands in view was cut with views wherein the hands were occluded during asynchronous moments. Camera 1 was positioned at the back of the hall and captured a lateral view showing the entire pianist and instrument including a clear view of the hands on the keyboard. Camera 2 was positioned at stage left, looking across the body of the piano with a clear frontal and tightly framed view of the performer's face and upper body, obscuring the hands. Behne and Wöllner (2011) demonstrated that such manipulations can give the impression of undoctored performances even among participants with high levels of musical training and knowledge of audiovisual and experimental manipulation techniques.

The pianist was instructed to perform the complete work from memory at a high, but not necessarily 'perfect', standard, achieved by recording the work shortly before the performer considered it to be concert-ready. This resulted in several minor inconsistencies in the performance (e.g. a wrong note at ~128 seconds) maintained throughout each condition to increase the validly of such a performance containing a catastrophic error in the relevant conditions. Following the performance, the pianist bowed and walked off stage. The pianist was also recorded making two stage entrances; one appropriate and one inappropriate. These were based on the criteria outlined by Platz and Kopiez (2013), in which the appropriate entrance displayed a confident stride, repeated eye contact with the audience, a deep bow, and nods of appreciation for the applause, while the inappropriate entrance featured a narrow gate, limited eye contact, hands in pockets, and an abbreviated bow. Additionally, a performance error as described above was recorded in which the pianist was instructed to begin playing approximately two-thirds of the way into the piece (bar 27), and then make a critical error in which the performance stops for several seconds, he struggles momentarily to find his place, then continues onward. He was also given the explicit instruction to convey intense frustration at having committed the error through his facial expression. Finally, a wide shot was filmed displaying the set stage without the pianist present with the first several rows of audience seats visible. Previously recorded pre-concert activity in the same venue was then superimposed over the bottom section of the screen, along with corresponding audio, giving the impression of a live audience present for the performance. Audience applause (taken from existing footage from the venue to ensure acoustic validity) was added to the stage entrances and to the final bow. With the resulting footage, five conditions were constructed using Final Cut Pro 7, each exactly 3 minutes in length plus an additional 4 seconds in the two videos (3 and 4) containing an aural performance error (see Table 4.1 below and Figure 4.2 in Section 4.2.5 for summaries and Videos 1-5 in Appendix 3).

Table 4.1. Properties of the five videos used in the study. Each video was formed of manipulations of the same recording of Chopin's *Aeolian Harp* Etude. Videos 1-5 can be found in Appendix 3.

Condition	Description	Stage entrance	Length (s)	Error at 100 s
Video 1	Standard	Appropriate	180	None
Video 2	Inappropriate stage entrance	Inappropriate	180	None
Video 3	Aural error with facial reaction	Appropriate	184	Aural/facial
Video 4	Aural error only	Appropriate	184	Aural only
Video 5	Facial reaction only	Appropriate	180	Facial only

4.2.3 Continuous measures: Development of a new tool

As the RCM continuous measurement tool used in Study 1 (and Thompson et al., 2007; see Section 2.3.2.3 for a full description) only allowed for the presentation of aural stimuli, a new capture system was required for the purposes of this study. Thus, a bespoke tool was created within the software package *Presentation* (Neurobehavioral Software, v. 17.2) in which custom experimental trials can be developed and executed using an adapted Python programming language. Several core features of the RCM software were maintained: the use of the trackpad to slide the scale from left (low) to right (high), the synchronisation of the response data to the stimulus via time-stamp indicating time since the trial and stimuli were initiated by the participant, and the ability to indicate when the first rating had been given. However, two substantial changes were made to adapt the system to a visual stimulus.

First, the RCM software had users move the mouse cursor left to right within a blue rectangle indicated on the screen. As maintaining this cursor position required some visual attention, there was concern that this might distract participants from focussing on the visual stimuli. Thus, the present tool tracked only horizontal movement along a laptop trackpad once the trial was initiated and did not display the cursor on the screen. A corresponding brightly-coloured visual scale was placed below the video to give a quick indication of the current rating position while minimising distraction from the stimulus. As this necessitated the scale having a fixed start-point, it was decided to fix this at the scale mid-point to avoid biasing participants towards the top or bottom of the rating scale.

Second, participants using the RCM software indicated the time at which they had made their first judgement by moving the mouse vertically into the blue rectangle. As the lack of visual cursor or tracking of vertical movement in the current system prevented this method, time of first decision was instead measured via the point at which participants first moved the cursor. To account for cases where the user's first impression corresponded with the scale's default, mid-scale starting position, the system recorded the time of a mouse click to allow participants (following instruction) to indicate the time and location of their first decision. A further advantage of this approach was that it allowed for the use of a visual reminder (i.e. the scale appearing as red and a line of explanatory text at the bottom of the screen) that participants had not yet began recording their continuous rating, ensuring that they did not forget to begin engaging with the system.

In using the system, participants were first shown an initial screen with instructions to "rate the quality of the following performance from 'Poor' to 'Excellent'". Upon the participant starting the trial, the software presented the video across the top of the screen. Underneath, a horizontal grey bar was presented alongside a rating scale ranging from 1 (poor) to 7 (excellent), following the scale used by Thompson et al. (2007). Horizontal movement on the laptop trackpad corresponded with a red bar moving across the grey space, which also recorded the position from 1 to 70 at 2 Hz in a separate file for analysis. The red bar began at the midpoint (35 out of 70), and clicking the trackpad recorded a timestamp and turned the red bar to blue to confirm a first decision had been entered. Figure 4.1 displays a screenshot of the continuous measurement interface, and the full code as developed and written by the author and used to execute the tool can be found in Appendix 4.

4.2.4 Procedure

After providing informed consent, participants were told that they would be evaluating a recording of a classical pianist. They were instructed to base their ratings "not on how much you enjoy the performance, but by how 'good' you feel the performance is, as if you were a competition judge". This differentiation was emphasised because the constructs of performance enjoyment and quality ratings,

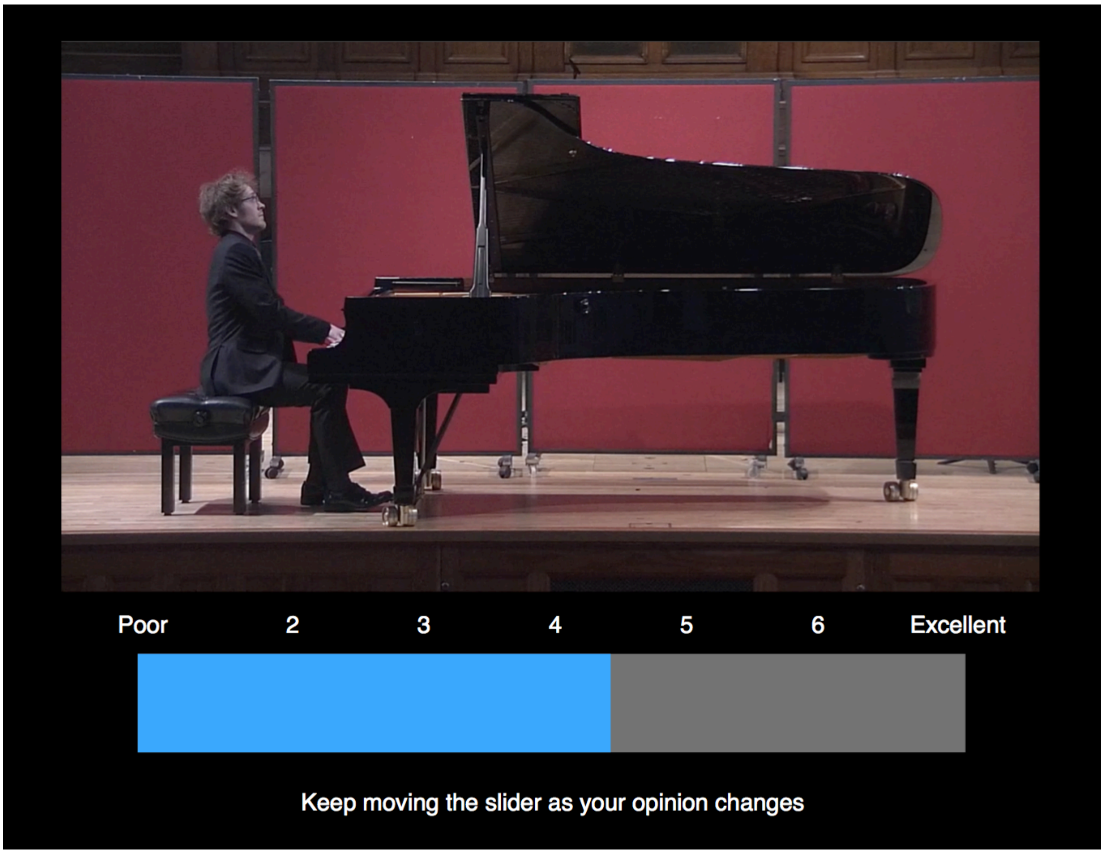


Figure 4.1. Screenshot of the custom continuous measurement interface. As the video plays the user can move the slider across the screen via the trackpad. Here, the user has already clicked to register the first judgement, turning the slider blue.

while correlated (Thompson, 2007), are assumed to be mutually exclusive in the act of professional performance evaluation (Thompson & Williamon, 2003). They were then able to try the continuous measurement software using a brief recording of a violinist playing unaccompanied Bach, with the instructions that:

- as soon as they had an opinion of the quality of the performance they should move the slider to the appropriate point and click (the click served to mark a first decision in the few cases where the slider's midpoint already indicated the participant's first rating), and

- they should feel free to move the slider at any point (without needing to click) if their opinion changed over the course of the performance.

They then initiated, watched, and rated one of the five videos (randomly assigned). Following the video, they completed a questionnaire on which they rated the performance's quality and typicality, their familiarity with the work, their enjoyment of the performance, and the appropriateness of the performer's on-stage behaviour on 7-point Likert-type scales. They were also free to provide open comments on the performance. The questionnaire also collected basic background information including age and musical training.

4.2.5 Data treatment and analyses

Data were first treated to several operations, primarily following Thompson et al. (2007) and as employed in Study 1, resulting in five general indicators of time to and score of first and final ratings. As a preliminary check, a visual examination of the data revealed one obvious erroneous spike in one participant's data caused by an accidental touch of the trackpad (i.e. a quick movement to an extreme score followed by an immediate return to the original score); this was removed and replaced with the score indicated immediately before and after the spike. Following this, five discreet variables were extracted from the full continuous data (see Figure 4.2):

- Time to first decision, **T_1**: As a brief amount of time was necessary to move the slider to the desired first rating point, the time of first movement (or the first click in the 3 of 105 cases where there was no initial first movement) was noted as the initial decision time, T_1. The continuous measurement ratings were taken from the beginning of the video, yet the first note was not played until 25 seconds in; therefore, 25 seconds were subtracted from each score, giving initial ratings made prior to the first note a negative time value. Two outliers wherein a first decision was not registered until after two-thirds of the performance had elapsed were removed, based on an admission from one participant that she had forgotten to indicate any judgement until late into the trial.

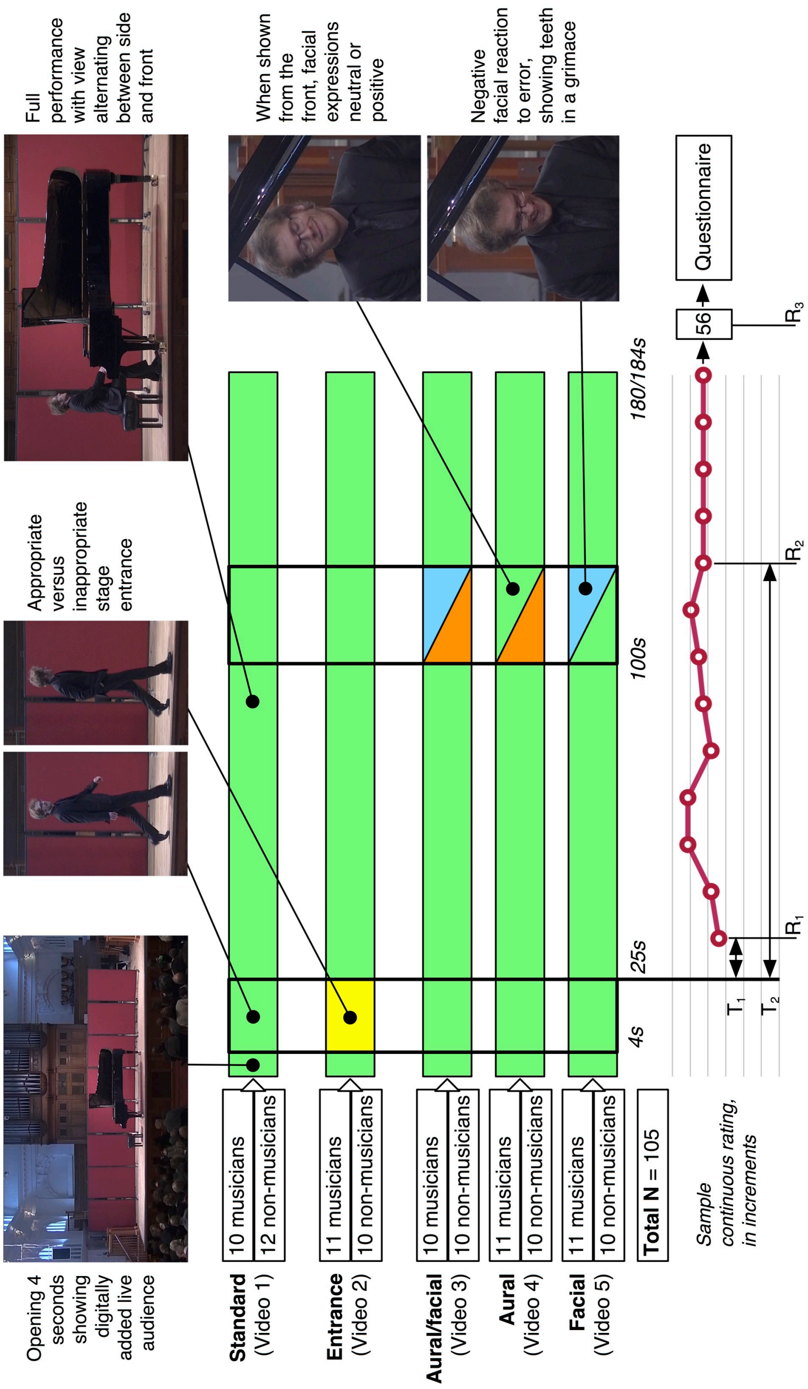

Opening 4 seconds showing digitally added live audience
Appropriate versus inappropriate stage entrance
Full performance with view alternating between side and front
When shown from the front, facial expressions neutral or positive
Negative facial reaction to error, showing teeth in a grimace
Standard (Video 1)
10 musicians
12 non-musicians
Entrance (Video 2)
11 musicians
10 non-musicians
Aural/facial (Video 3)
10 musicians
10 non-musicians
Aural (Video 4)
11 musicians
10 non-musicians
Facial (Video 5)
11 musicians
10 non-musicians
Total N = 105
Sample continuous rating, in increments
4s
25s
100s
180/184s
T1
T2
R1
R2
R3
56
Questionnaire

Figure 4.2. The Study 2 research design. 105 participants were each randomly assigned one video condition to view. From their continuous data, the time to first decision (T_1), time to final rating (T_2), first rating (R_1), and final rating (R_2) were calculated. The overall rating (R_3) came from a written score completed after the video and continuous measurement were finished, followed by a questionnaire. Shading: yellow = inappropriate stage entrance, orange = aural performance error, blue = negative facial reaction.

- First rating, **R_1**: The first point at which the participant maintained a stable rating of at least 2 seconds was taken as the first rating.
- Final rating, **R_2**: The final score reported in the continuous data.
- Time to final rating, **T_2**: Participants' continuous data tended toward brief, direct movements between stable plateaus. Thus, the time of final rating, T_2, was recorded as the point at which the movement leading to the final rating (R_2) was started. As with T_1, 25 seconds were subtracted from each score to account for the stage entrance.
- Overall rating, **R_3**: The overall written score provided in the questionnaire on a scale of 1-7. For a direct comparison with the final continuous rating, R_2 was also converted from 70-point to 7-point values following Thompson et al. (2007).

Preliminary analyses using a series of t-tests showed no significant differences between men and women on the group T and R scores; subsequently, sex was discounted as a between-groups variable. Differences in R and T scores between conditions (Videos 1-5) and experience groups (musicians versus non-musicians) were analysed using 5x2 factorial ANOVA models. Planned contrasts were run specifically for the hypotheses being tested. In examining the effect of the stage entrance on T_1 and R_1 (i.e. hypothesis 1), only Video 2 with the 'inappropriate' entrance differed in opening material that could affect these measurements. Therefore, a Helmert contrast was employed as this allows a condition to be compared with the sum mean of the following conditions (i.e. Video 2 versus 1, 3, 4, & 5; Video 1 versus 3, 4, & 5; Video 3 versus 4 & 5; Video 4 versus 5). Simple contrasts, in which each video was compared with the *standard* control, were used for the remaining tests (i.e. hypothesis 3). T-tests were used for direct comparisons of experience level in hypotheses 2 and 4. As R_1 and R_2 were commensurable, they were tested using a mixed 2x5x2 ANOVA to examine changes between first and final ratings. To analyse moment-by-moment changes within each group resulting from the stage entrance behaviour, performance errors, and facial reactions, repeated-measures ANOVAs were calculated using mean scores at 10-second increments from the beginning of the video. This followed the method reported by Thompson et al. (2007), who used 15-second increments; the value was reduced to 10 seconds to provide greater precision around the performance error.

4.3 RESULTS

Analyses in the first two sections below examine between-group differences (i.e. conditions 1-5 and musicians versus non-musicians) and within-group comparisons of time to first decision (T_1), time to final decision (T_2), and first (R_1), final (R_2), and overall written (R_3) ratings. A complete set of means and SDs are provided in Appendix 5. The following two sections focus on repeated-measures analyses of the continuous effects of the stage entrance and aural/facial errors. The final section examines relationships between general features of the participants' attitude toward the work, such as familiarity with and likeability of the piece.

4.3.1 Hypotheses 1 and 2: Effects of stage entrance on time to first decision (T_1) and first rating (R_1)

Four of the five conditions used the same opening material: that of the appropriate, confident stage entrance by the performer. Only the condition featuring the inappropriate stage entrance (Video 2) varied from the others in its opening material, thus we investigated whether participants responded differently to the altered stage entrance in both the time to and result of their first ratings: T_1 and R_1 (hypothesis 1). To test this, ANOVAs comparing condition (x5) and musical experience (x2) with T_1 and R_1 as dependent variables were each followed by a planned Helmert contrast.

For T_1, while the ANOVA showed no overall differences between conditions, experience groups, or any interaction, the Helmert contrast showed a significantly lower time to first decision ($t_{(93)} = -10.42$, $p < .05$, $r = .73$) while watching the inappropriate stage entrance (M = 8.00, SD ± 17.00 seconds) versus the combined effect of the remaining four (M = 18.52, SD ± 20.64 seconds; see Figure 4.3). Level 2 of the contrast, in which the *standard* condition was compared with the remaining three, showed no significant difference, demonstrating consistent decision times across groups viewing videos with identical opening material. Furthermore, 6 of the

21 *entrance* raters (29%) recorded a first decision before the performer had played his first note, compared with 6 of the remaining 84 participants (14%) that viewed one of the other four conditions.

For R_1, the ANOVA showed a significant overall effect of condition ($F_{(4,95)} = 4.94$, $p < .005$, $\eta^2 = .16$), with no overall effect of or interaction with experience group. The Helmert contrast mirrored that of T_1, showing a significantly lower score reported ($t_{(95)} = -7.78$, $p < .005$, $r = .62$; see Figure 4.4) by those watching the inappropriate

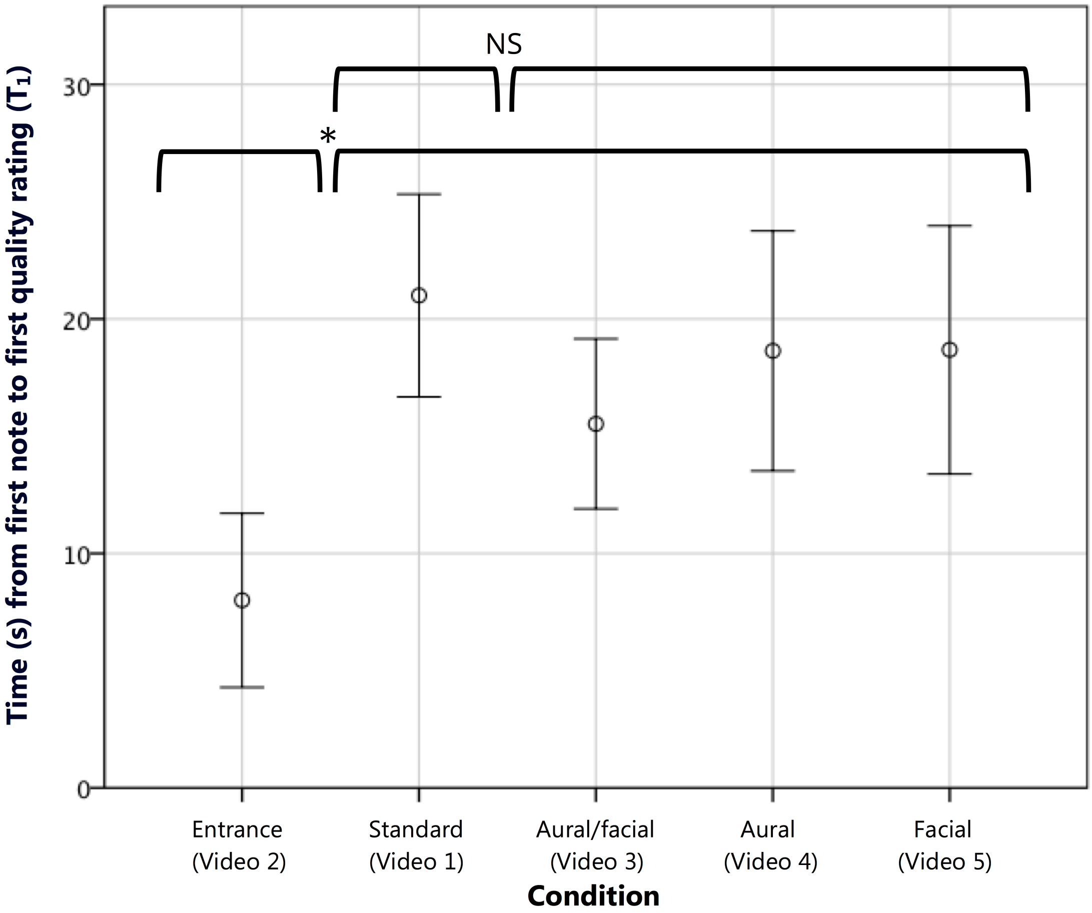


Figure 4.3. The combined mean time to first judgement (T_1) in seconds measured from the first note played. The inappropriate *entrance* condition resulted in a significantly lower time to first decision compared with the other four conditions. Error bars show 95% CI. * = $p < .05$, as tested using a Helmert contrast in which the *entrance* condition was compared with the mean of all subsequent conditions.

stage entrance versus the remaining conditions. Also, as with T_1, no significant effect was seen at the second contrast level comparing the *standard* and remaining videos. The hypothesis that musicians would more harshly penalise an inappropriate stage entrance (hypothesis 2) was confirmed with a comparison ($t_{(19)} = -2.00$, $p < .05$, $r = .42$; one-tailed) wherein musicians gave an average initial rating of 34.91 (SD ± 17.18)

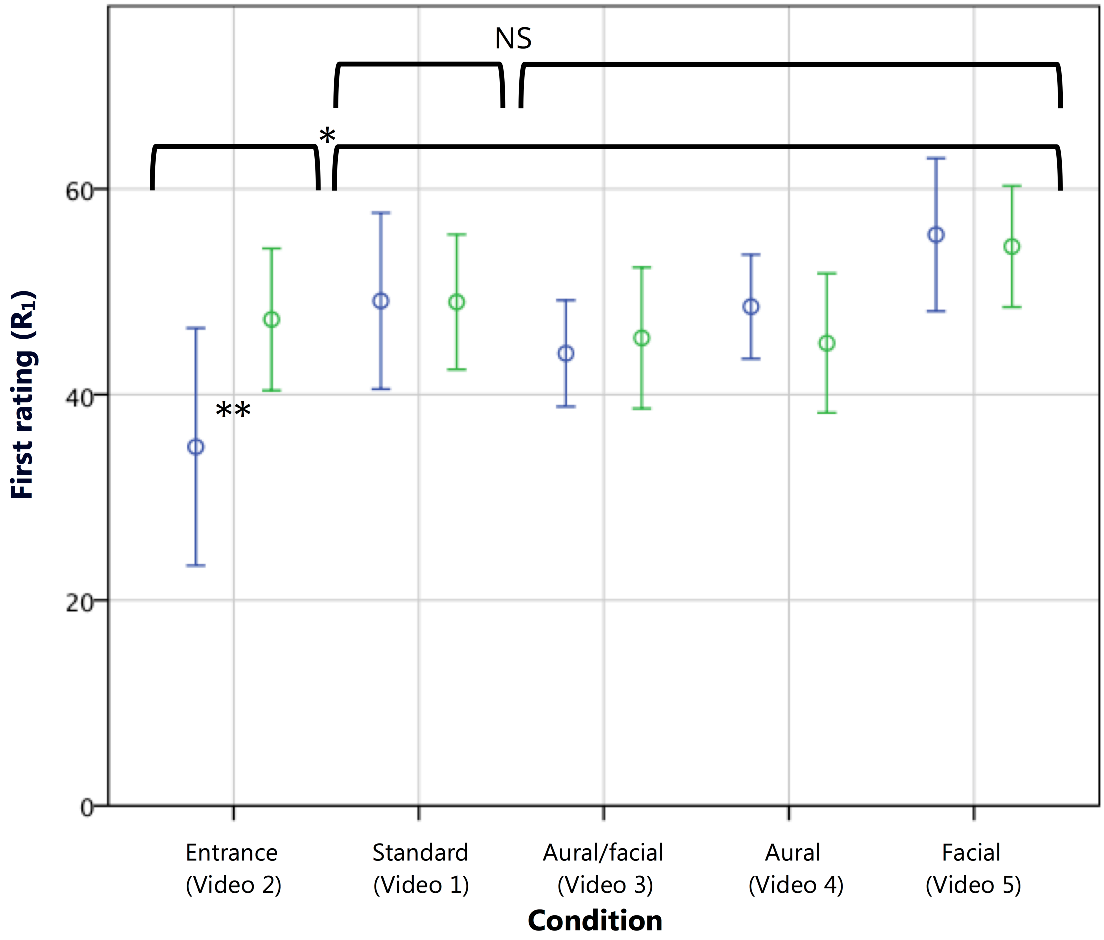


Figure 4.4. First continuous ratings (R_1) of musicians (blue) and non-musicians (green) on a scale from 1 - 70. The inappropriate *entrance* condition resulted in a significantly lower first rating compared with the other four conditions. A direct comparison revealed that this difference was due to a significantly lower first rating among musicians as compared with non-musicians. Error bars show 95% CI. * = $p < .005$, as tested using a Helmert contrast in which the *entrance* condition was compared with the mean of all subsequent conditions. ** = $p < .05$ in a comparison between musicians and non-musicians within the *entrance* condition.

and non-musicians a rating of 47.30 (SD ± 9.66), on par with first ratings across the other conditions. No significant difference in time to first decision (T_1) was found between musicians and non-musicians in a similar comparison. Thus, the manipulated stage entrance was indeed found to have an effect on continuous quality evaluations. Musicians gave significantly lower initial ratings when viewing the inappropriate stage entrance, and both musicians and non-musicians delivered their first ratings of this condition in a significantly shorter length of time.

4.3.2 Hypotheses 3 and 4: Effects of condition on final decision (T_2) and final rating (R_2 and R_3)

The mean time to a final, stable rating (T_2) across conditions was 128.31 seconds (SD ± 24.51) of the total 180 seconds of the entire performance (or 184 seconds for Videos 3 and 4, in which the aural error incorporated an extra 4 seconds of musical material). The ANOVA revealed no significant difference in final decision times based on condition or experience, and a Helmert contrast with the *entrance* condition in the first position and *standard* in the second showed no effect of condition at any level (see Figure 4.5). Thus, while the inappropriate stage entrance caused raters to make their first judgements more quickly (hypothesis 1), it showed no significant effect on how long they took to come to a final decision about the performance.

The mixed 2x5x2 ANOVA comparing the first (R_1) and final (R_2) continuous scores showed that, overall, the groups' initial mean ratings did not differ significantly from their final ratings. However, a significant interaction of rating and condition was shown ($F_{(4,95)} = 5.56$, $p < .001$, $\eta^2 = .18$), and a planned simple contrast comparing each condition to the *standard* showed that the *aural/facial* condition followed a different overall profile ($t_{(95)} = -7.55$, $p < .01$, $r = .61$). As the ANOVA examining R_1 showed no significant difference in the first score for this condition, it followed that a significantly lower final score would instead be the cause of the significant interaction effect. A 5x2 ANOVA examining R_2 confirmed this with a significant effect of condition ($F_{(4,95)} = 5.56$, $p < .001$, $\eta^2 = .19$) with no effect of or interaction with experience. Again, a planned simple contrast was conducted comparing each condition with the *standard*. Only the *aural/facial* condition (M = 36.00, SD ± 13.37) was found

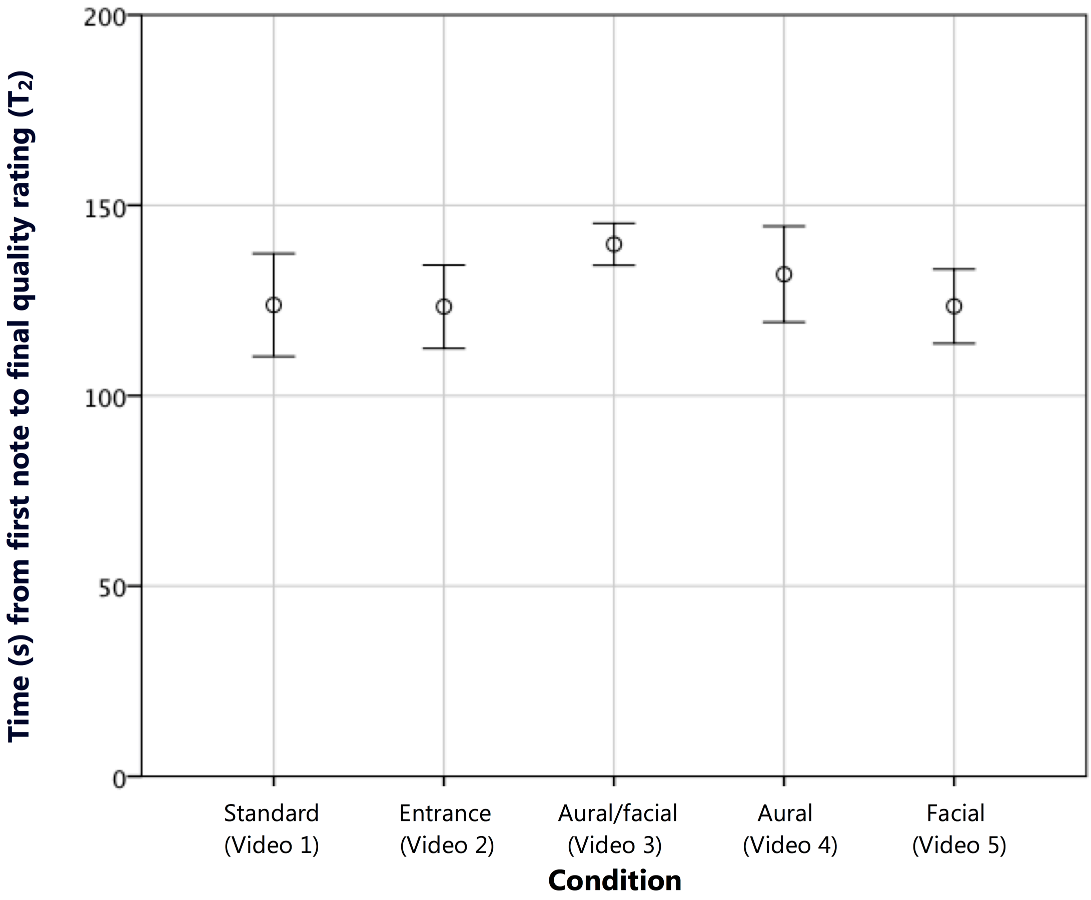


Figure 4.5. The combined mean time to final judgement (T_2) in seconds measured from the first note played. No significant difference was found between conditions. Error bars show 95% CI.

to have received a final continuous rating significantly lower than the *standard* (M = 46.82, SD ± 11.55; $t_{(95)} = -10.80$, p < .005, r = .74; hypothesis 3; see Figure 4.6). An analysis of the final written scores (R_3) showed similar findings, with a main effect of condition ($F_{(4,95)} = 4.87$, p < .005, $\eta^2 = .17$) and contrasts revealing that only the *aural/facial* score (M = 3.90; SD ± 0.97) was significantly lower than the *standard* on the 7-point scale (M = 4.86, SD ± 1.32; $t_{(95)} = -0.96$, p < 0.005, r = .10; see Figure 4.7). A direct overall comparison of R_2 and R_3 with a repeated-measures ANOVA (following a conversion of R_2 from a 70-point to a comparable 7- point scale, as described in Section 4.2.5 above) with experience and condition as between-subjects

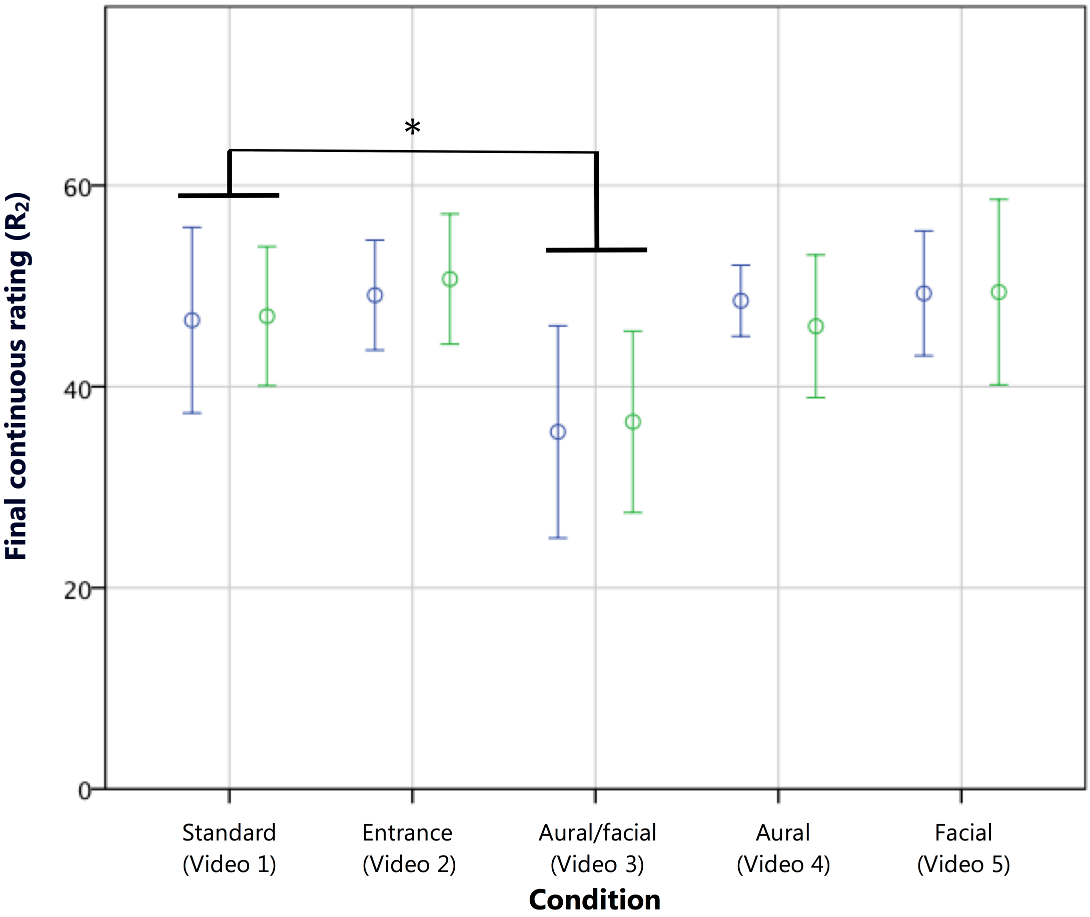


Figure 4.6. Final continuous ratings (R_2) of musicians (blue) and non-musicians (green) on a scale from 1-70. The *aural/facial* condition, comprising a performance error with corresponding negative facial reaction, resulted in the only significantly lower performance rating. Error bars show 95% CI. * = p < .005, wherein a simple contrast compared each condition with the *standard*, with no interaction with experience group.

variables also showed no main effect of rating type on the reported scores. R_2 and R_3 also showed a strong correlation (τ = .70, p < .001). This suggested that the final continuous ratings accurately reflected the opinions given by the more routinely used written scores, thus confirming the validity of continuous rating as a proxy for evaluation scores given in standard summative procedures (Thompson et al., 2007). R2 and R3 both showed small correlations with R1 (τ = .23, p < .005 and τ = .23, p < .001, respectively).

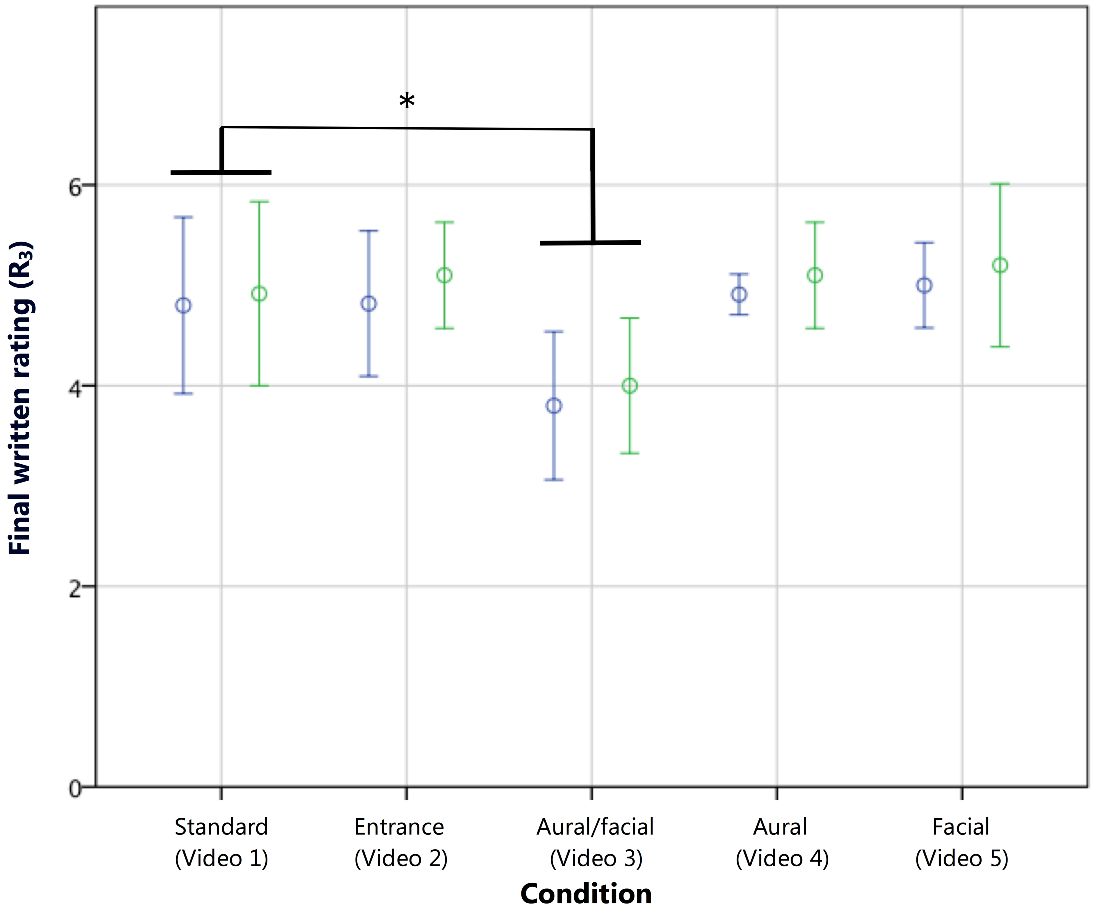


Figure 4.7. Final written ratings (R_3) of musicians (blue) and non-musicians (green) on a scale from 1-7. As with R_2, the aural/facial condition, comprising a performance error with corresponding negative facial reaction, resulted in the only significantly lower performance rating. Error bars show 95% CI. * = $p < .005$, wherein a simple contrast compared each condition with the standard, with no interaction with experience group.

These analyses found that the inappropriate stage entrance did not have a lasting effect on the final ratings (R_2 and R_3) given by either musicians or non-musicians. As this contrasted with the lower initial ratings (R_1) given by musicians as reported in the previous section, the following section examines the point at which this difference in rating converged with the *standard* condition. Regarding the performance errors, only the *aural/facial* condition had a significant effect, lowering the final ratings (R_2 and R_3) of both experience groups. No overall effects of the *facial* or *aural* errors alone were found on the final ratings. Again, repeated-measures analyses of the continuous measures data were then employed to examine the effect of the errors at the point of occurrence, as reported below.

4.3.3 Continuous effects of the stage entrance

As the above analyses of the final and overall ratings (R_2 and R_3) showed that those viewing the inappropriate stage entrance condition did not yield significantly lower scores than those in the *standard* condition, the lower R_1 scores reported by the musicians seemed to have rebounded by the end of the performance. To identify how soon after the initial stage entrance this was accomplished, average ratings at 10-second intervals from the beginning of the video were extracted and analysed using a repeated-measures ANOVA with planned contrasts of each interval to the final score. When conducted from the 50-second mark (25 seconds from the first note played), where 8 of the 10 musicians in this subsection were already reporting a mean score of 50.13 (SD ± 7.08), no significant difference from the final score was found in the remaining 12 levels. Thus, any negative impression caused by the inappropriate entrance, reflected in the quicker first rating among both experience groups and lower initial rating by musicians, was not reflected in the rating after 25 seconds of musical performance. Direct repeated-measures analyses prior to the 25-second point were not possible using this method due to the number of missing pairwise data sets resulting from participants who had not yet recorded their first rating. These results should be considered in light of the non-significant difference between the *entrance* and *standard* conditions in their change of R_1 to R_2, as shown by the 2x5x2 mixed ANOVA contrasts described above, where the difference in this subgroup did not emerge as significant when examined in conjunction with the other four conditions. Thus, any effect of the stage entrance on initial ratings among musicians did not persist when the pianist began playing, despite having formed their initial, more negative impressions significantly earlier.

4.3.4 Hypotheses 3 and 4: Continuous effects of the performance errors

Three conditions related to performance errors: *aural/facial* (Video 3), in which a performance error with corresponding negative facial reaction was spliced into the *standard* recording (Video 1); *aural* (Video 4), in which audio from the same performance error was superimposed with the visual recording of the *standard* condition; and *facial* (Video 5), in which the visual reaction to the mistake was superimposed over the correct playing. As reported above, only the *aural/facial* condition triggered a significantly lower overall rating than the *standard*, reported by both musicians and non-musicians. Visual examination of the data revealed that this stemmed from a dramatic, immediate drop in continuous ratings immediately following the error by respondents when compared with the *standard* (see Figure 4.8; a visualisation of the raw continuous data for both the aural/facial and standard conditions, emphasising the immediate and consistent nature of this reaction with comparison to the overall variability in the continuous data, can be seen in Appendix 6).

To determine the individual and combined effects of the aural and visual (i.e. *facial*) components on musicians and non-musicians, average continuous ratings at 10-second intervals were again extracted and plotted. To determine when the final *aural/facial* score was finalised, a mixed ANOVA was conducted with 12 time intervals from the 70-second mark as a repeated-measures factor (30 seconds prior to the error, where 19 of the 20 participants across both experience groups had begun registering their continuous responses) and experience as a between-group variable. Planned contrasts comparing each point with the final score were used to isolate when the final decision was reached. A significant effect of rating over time was found ($F_{(11,187)}$ = 20.20, $p < .001$, $\eta^2 = .53$) with no main effect of or interaction with experience, and contrasts were significant ($p < .05$, $r = .32 - .62$) until the 120-second point (20 seconds following the error) which followed a slight increase from the 110-second point following the error-invoked drop. To examine musicians' and non-musicians' specific reaction to the error, difference scores were calculated between ratings immediately before (100 seconds) and after (110 seconds) its presentation for the *standard*, *aural/facial*, *aural*, and *facial* conditions. The ANOVA revealed a significant effect of condition ($F_{(3,80)}$ = 14.85, $p < .001$, $\eta^2 = .28$) with contrasts

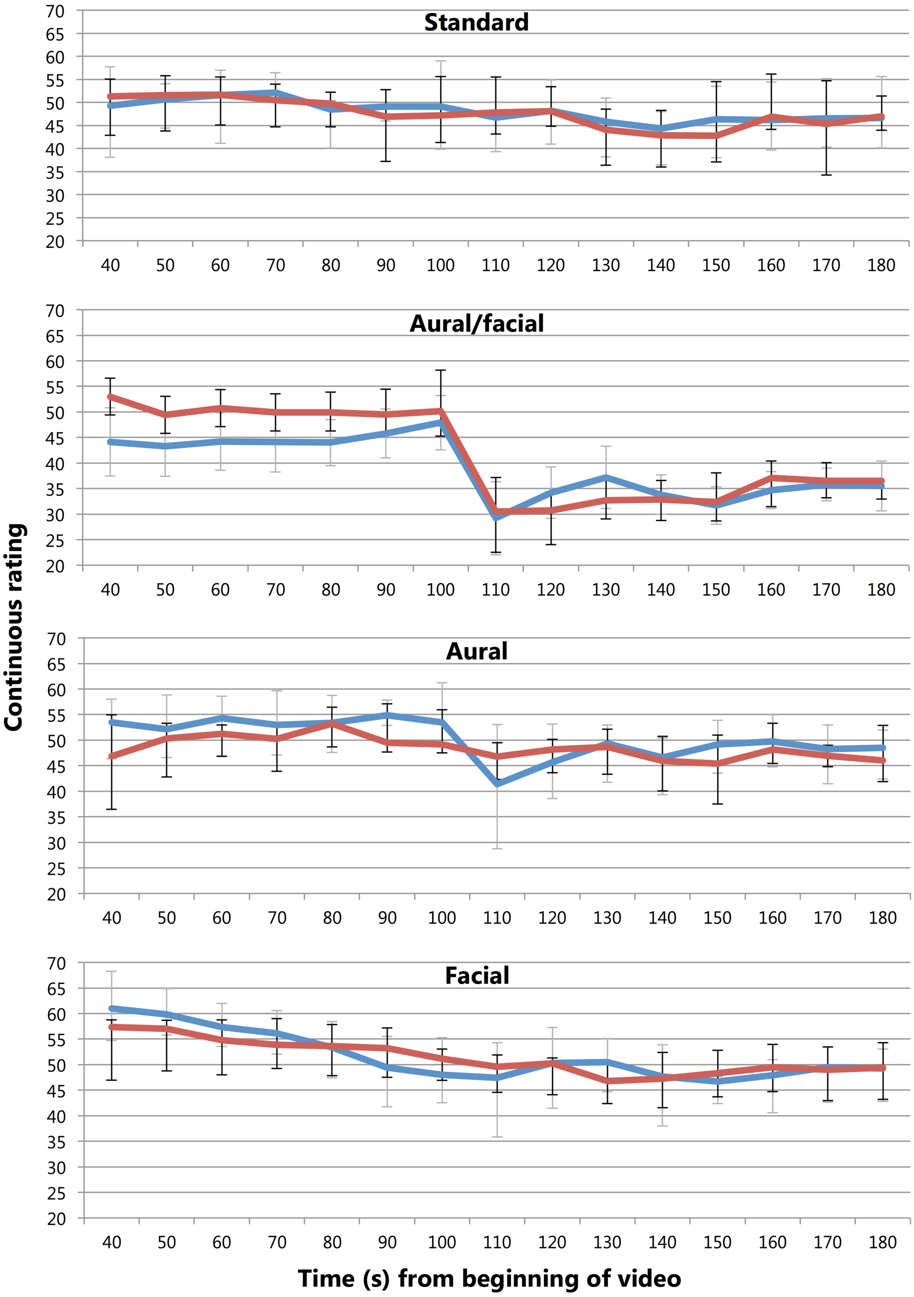


Figure 4.8. Mean participant ratings of musicians (blue; grey error bars) and non-musicians (red; black error bars) across the *standard*, *aural/facial*, *aural*, and *facial* conditions at 10-second intervals. Time in seconds from video opening – first note played at 25 seconds and error occurred at 100 seconds. Axes begin at t = 40 seconds to reflect the point at which most participants were supplying data, allowing for consistent representation of mean and error. A larger drop can be seen at the point of the error in the *aural/facial* condition, with a smaller drop in the *aural* condition by musicians only and no significant movement in the *standard* and *facial* conditions. Error bars show 95% CI adjusted for repeated-measures data.

revealing that both the *aural/facial* condition ($t_{(80)}$ = -19.02, p < .001, r = .90) and the *aural* condition ($t_{(80)}$ = -7.25, p < .05, r = .63) showed significant drops in comparison with the *standard* (M = -19.20, SD ± 13.87; M = -7.43, SD ± 7.24; and M = -0.18, SD ± 8.07, respectively), but no such movement was seen in the *facial* condition (M = -0.90, SD ± 7.24; see Figure 4.8).

Hypothesis 3 posited that musicians would react to the performance error more severely than non-musicians. A comparison confirmed this in the *aural* condition where musicians made a significantly larger drop ($t_{(19)}$ = -2.12, p < .05, r = .44) during that period, with musicians lowering their score by a mean 12.00 points (SD ± 12.69) and non-musicians by 2.40 points (SD ± 6.81) out of the total 70 over that 10-second period (see Figure 4.8). However, as shown by the R_2 and R_3 scores above (see Section 4.3.2) this penalisation by musicians was not reflected in their overall ratings. No such difference was found in a similar comparison within the *aural/facial* condition.

To summarise, when the *aural* error was presented alone, the musicians reacted with a significantly lower immediate decrease in scores to the non-musicians, although this penalisation was not reflected in the final scores. When the *facial* error was presented alone, no immediate or overall effect was shown, regardless of experience. When the two errors were juxtaposed in the *aural/facial* condition, however, both experience groups showed an immediate drop in continuous quality rating that was reflected in the final (R_2 and R_3) ratings.

4.3.5 Work familiarity, likeability, and typicality

Participants' ratings of how much they liked and knew the composition (likeability and familiarity), how typical the performance was, and the appropriateness of the performer's behaviour were tested for correlations (Kendal's tau, due to the large proportion of tied ranks within the 7-point scales) with T_1, T_2, R_1, R_2, and R_3. After controlling for multiple comparisons, no significant relationships with the time to form their decisions (T_1 or T_2) were found, and only the appropriateness of the performer's behaviour significantly correlated with the overall rating, R_3 (τ = .28, p < .05), although its correlation with R_2 was not significant and therefore should be interpreted with caution. A significant correlation between participants' familiarity with and liking of a composition was found ($\tau = .37$, $p < .01$).

4.4 DISCUSSION

The present study sought to examine the temporal nature of musical assessment as it is affected by visually-communicated behaviours of the performer. It employed continuous measures methodologies to reveal previously unexamined immediate and overall effects on the decision-making process of extra-musical variables that could be defined by their having occurred prior to (i.e. the stage entrance) or at a specific point during (i.e. the error) a performance. To achieve this, a recorded performance of Chopin's *Aeolian Harp* etude was manipulated to vary in appropriateness of the stage entrance or in the incidence of an aural performance error and/or corresponding negative facial reaction. The effect of experience was examined by comparing response differences in musicians and non-musicians. The continuous ratings were able to show effects of these variations that the standard *post hoc* measurements would not have revealed, addressing the four hypotheses set out in Section 4.1.3.

4.4.1 Hypotheses 1 and 2: Effects of the stage entrance

Where the inappropriate stage entrance did not have an overall effect on final ratings, the continuous data showed a significantly shorter time to first decision across experience groups and a lower initial rating by musicians that quickly recovered, confirming both hypotheses 1 and 2. In the discussion of their results, Thompson et al. (2007) questioned the generalisability of their finding that initial decisions were made within an average of 15 seconds following the first note, particularly in situations outside of their audio-only condition. The present research not only supports those findings, in that initial ratings across the four groups without the inappropriate stage entrance were made in approximately 18 seconds, but suggests that the presence of visual information relating to the performance, including the performer's behaviour as they take the stage, does not alter this process to a great degree so long as the entrance is deemed 'appropriate'. When stage entrances betrayed the expectations of their audience that decision was made earlier, and occasionally before the first note was played as was hypothesised by Thompson and colleagues.

This study also found that experience did not play a role in the speed at which the judgement was formed, implying that the heightened expectations and knowledge of the material to be performed neither increased nor hindered the rate at which judges could form their decisions (or, at least, were willing or able consciously to record their first decision). However, experience did play a small role in the height of the first rating, where musicians reported a significantly lower initial score than non-musicians for the inappropriate entrance. Here, their greater experience with, and thus expectations of, the protocols of stage entrance behaviour in the Western classical tradition may have caused them to penalise the performer more harshly. However, this judgement did not last long. When Platz and Kopiez (2013) demonstrated that the appropriateness of a violinist's entrance correlated positively with their anticipation of the performance's start, they wondered how sustainable the positive motivational effect might be were the performance to continue. While it is unclear what the specific effect of a *positive* impression might be in the current study, due to the finding that the average group ratings did not significantly differ from final ratings in the *standard* condition, it was shown that the *negative* impressions recorded by the musicians in the *entrance* scenario had dissipated (i.e. ratings had returned to the baseline of the *standard* rating) within 25 seconds of the first note. This aligns with the findings of Wapnick and colleagues (2009) where the visual effect of heightened attractiveness on higher quality ratings for female performers appeared in 25-second excerpts but not in longer ones. It is perhaps promising news for musicians; while the standard finding from the general evaluation literature is that negative first impressions are more resistant to change than positive impressions (e.g. Ybarra, 2001), in this case a negative first impression was quickly forgiven based on the quality of the performance that immediately followed. While stage entrance behaviour made an impression on performance quality ratings, the impression of the musical content itself took precedence once it began. Future studies can examine the effect of an appropriate or inappropriate stage entrance on an initially poor musical performance, the latter of which was found in Study 1 of this thesis to drastically alter the initial and final ratings of a performance.

4.4.2 Hypotheses 3 and 4: Effects of the error and facial reaction

Regarding the errors, overall written scores showed that only the performance error with corresponding facial reaction (i.e. the *aural/facial* condition) led to a lower rating (hypothesis 3), but the continuous measures data again demonstrated a more complex process at work. Musicians penalised then forgave a performance error on its own, only providing a lower overall score when the error was paired with a negative facial reaction. Non-musicians were significantly less harsh in their initial judgement of the aural error alone (hypothesis 4), although behaved just as their more musically experienced counterparts when the facial reaction was juxtaposed. Neither group reacted to the negative facial reaction on its own.

That the negative impression of performance errors in musicians was temporary replicates the finding of Study 1. There, errors placed at the midpoint of two audio-only recordings caused immediate but quickly-forgiven drops in continuous quality ratings, with no indication in the final written ratings that the errors made any lasting impression. The lack of response from the non-musicians may indicate that they simply did not perceive that an error had occurred, although the severity of the mistake makes this situation unlikely. In the optional comments section, several non-musicians rating the *aural* condition indicated that they were aware of the error, where one wrote that they "perceived a mistake at about two-thirds of the way through". Furthermore, the fact that non-musicians behaved in the same manner as the musicians in the *aural/facial* and *facial* conditions (i.e. reacting strongly to a performance error with negative facial response but having no reaction to the facial response on its own) indicates that they indeed perceived the aural difference. The facial reaction, then, may have instilled in the non-musicians the confidence to penalise the error which they lacked in hearing the aural error alone. However, the question remains why the facial reaction caused the error to be perceived as that much more detrimental to the overall performance, as when the negative expression was presented in isolation it caused no measured effect in either group. Put another way, it was not the behaviour inherent to the expression that was penalised; it was how the expression altered the impression of the performance error itself and its lasting effect on the final performance rating.

The ecological model of *emotion face overgeneralisation* may account for this, wherein those interpreting a facial expression infer information not only concerning affective state but also of generalised traits (Zebrowitz et al., 2008). Participants have rated people displaying sad faces as lower in trait dominance, while happy or surprised faces resulted in higher dominance and affiliation ratings (Montepare & Dobish, 2003). Thus, it could be expected that a musician's expression of frustration and anger at the committal of a performance error may result in the viewer regarding a trait tendency displaying general lack of control, instead of simply a performer who has, in that moment, lost control. Rather than being a musician momentarily making a mistake, they are perceived as musician *that makes mistakes*. This especially as the goal of music performance quality evaluations is often not only to rate the quality of the performance but, by extension, the performers themselves.

Both the findings relating to the stage entrance and to the facial expressions point to the interaction between aural and visual information, with the former taking some precedence. Tsay (2013, 2014) found that presenting visual information alone led to more accurate predictions of competition results than audio-only or audiovisual condition, though, crucially, participants were given extremely brief clips in which an immediate impression had to be formed. Here, a visually specific stage entrance caused an immediate reaction that was tempered after a period of aurally specific musical content, once participants were given time to process it. A visually specific facial reaction had no effect unless it supported an aurally presented musical error. While the visual element of performance still played a role, particularly in triggering immediate reactions, the aural information was dominant over time.

4.4.3 Directions for future research

Generalisability of the present study is limited by the nature of the experimental condition, as encountered in Study 1. While the use of genuine performance recordings and video manipulation to give the impression of a live performance was undertaken to maximise ecological validity, participants nonetheless made their judgements in artificial situations, wearing headphones while observing the performances take place on a laptop screen. While many music quality judgements indeed take place in this environment, whether in private listening to a recording or professional evaluation of a recorded competition submission, whether the processes of evaluation here studied are maintained *in situ* during live performances, surrounded by fellow audience or panel members, should be examined in future research. Furthermore, the use of multiple camera angles (necessary to hide the obvious asynchrony between the hands and music in the manipulations) maximised raters' view of the pianist's face at the point of the manipulated error in the relevant conditions. This provided ideal conditions for the effects of facial expression to manifest. While this framing is common in performance broadcasts, it is less likely to be viewable in single-camera or live performance settings and further study is required to determine whether the effects of facial expression are maintained in less ideal viewing conditions.

It should also be noted that the presentation of inappropriate stage entrances or performance errors were inserted into a performance of particularly high (although not perfect) overall quality. This juxtaposition was intentional in order to provide a clear experimental framework, and further study will be required to determine whether an audience's tendency to 'forgive' certain forms of performance error is maintained when the quality difference between those errors and the surrounding performance is not so stark. This also relates to the extreme severity of the performance error itself, where the performance momentary stopped. While common at amateur levels, this event is increasingly rare (but not unheard of) at such high ability levels. The current study demonstrates the effects of such a catastrophic mistake; further work could employ the same design with errors of varying nature and increasing subtlety.

Finally, it could be argued that use of the software interfered with participants' natural processes of performance evaluation, causing an increase in cognitive load that distracted from the final rating. The same could be said for the results of Study 1. Promisingly, when participants were asked following the present experiment whether using the software consciously affected their ability to deliver a quality judgement, only 11% reported that it made the process more difficult; 46% reported that the software made no difference, and 42% reported that it made judgements easier. Schubert (2013) found a test-retest reliability of approximately 80% when using a continuous interface to record perceptions of musical emotion. A significant amount of unreliability stemmed from the opening seconds of the performance, during which participants oriented themselves to the rating paradigm. The methodology employed in Studies 1 and 2 minimised this issue in that participants were asked not to begin recording until they had decided on their first response. Overall, this suggests that familiarity with such devices in musical experiments does not significantly affect participants' ability to focus on the task.

Overall, the present study has demonstrated a temporally dynamic process of music performance quality evaluation that can be measured to determine the effects of temporally specific musical and extra-musical factors. Visual information in particular plays a key role in the decision-making process, but in a more nuanced relationship with the aurally based musical content than previous research has been able to demonstrate. In particular, the pre-performance rituals of Western classical performance made a difference on quality ratings, both in terms of impression formation and perhaps in determining performer traits. Whether or not it has been a focus of study, the role of personal expression on musical impression formation has been acknowledged for some time in practice. George Grove, the first director of the Royal College of Music and author of the eponymous Grove Dictionary of Music, was struck by such an effect when he saw the pianist Franz Liszt perform in 1886. He wrote that he:

> was delighted (1) by his playing, so calm, clear, correct, refined–so entirely unlike the style of the so-called 'Liszt School'– (2) by his face. Directly he sat down he [sic] dismissed that very artificial smile, which he always wears, and his face assumed the most beautiful serene look with enormous power and repose in it. It was quite a wonderful sight (Graves, 1903, p. 311-312).

Grove was taken not only by the great pianist's performance, but the impression of Liszt's character; an impression that centred on the emotive capabilities of the face. Whether or not the visual aspect of Western classical performance has indeed been ignored in explicit practice and research, recent studies have moved it sharply into focus (e.g. Platz & Kopiez, 2012; Tsay 2013, 2014; Silveira, 2014; Krahe et al., 2015). Continued study of these extra-musical variables and their effects on evaluation can now tease apart the relation between and weighting of their myriad aspects, the points in time at which each is most influential, and the lasting effects they may have as musical decision-making unfolds.

4.5 SUMMARY

This chapter examined the effects of visually-communicated features of the performer (i.e. the stage entrance and negative facial reactions to a performance error) on the products and processes of performance quality ratings. 53 musicians and 52 non-musicians gave continuous quality evaluations of one of five randomly assigned videos, each manipulated to include an inappropriate stage entrance, aural performance error, error with negative facial reaction, or facial reaction alone. As in the previous chapter, the continuous measures methodology revealed insights into the evaluation products only apparent when examining the temporal process leading to them. Results showed that participants viewing the 'inappropriate' stage entrance made judgements significantly more quickly than those viewing the 'appropriate' entrance, and musicians' judgements started significantly lower in the former condition but quickly increased to match those of the latter. The aural error caused an immediate drop in quality judgements that persisted to a lower final score only when accompanied by the frustrated facial expression from the pianist; the performance error alone caused a temporary drop only in the musicians' ratings not seen among non-musicians, and the negative facial reaction alone caused no reaction regardless of participants' musical experience.

5 STUDY 3: THE EVALUATOR

5.1 INTRODUCTION

The two experimental studies presented thus far have focussed on musical and extra-musical variables relating to the repertoire and the performer, investigating their relation to the processes and products of music performance quality ratings. Following the overarching research questions set out in Chapter 1, the environment and evaluator remain to be examined. The evaluator is the focus of the study reported here. The first two experiments provided insight into the role and nature of the decision-maker in music assessment settings. Study 1 suggested a link between judges' familiarity with the repertoire and the process by which they assess a performance, where familiarity had no relation among the works of Chopin but a completely unknown work was judged via a significantly different trajectory. Study 2, in comparing musicians' and non-musicians' reactions to the performer, demonstrated some differences but great similarities in their reactions to a manipulated stage entrance.

Those studies considered the pre-existing knowledge and expertise of the evaluator, which are far from the only relevant factors in understanding how a judge approaches a musical performance. Human reactions to and perceptions of music are complex, comprising a range of affective, evaluative, behavioural, and autonomic responses. Each listener brings to the experience his or her own set of expectations and aesthetic preferences that inform judgements (Levinson, 1987), engaging not in a passive absorption of the performance but an active construction of opinion and experience (Cross, 2010). Unravelling these relationships is further complicated when the listener is in a live concert setting, outside of a formal or artificial situation in which the explicit task is the formation of a quality judgement and he or she is in a relatively restrictive setting, whether in a laboratory, judging an audition recording, or listening to a recording in the privacy of home. While the various features of the social and physical environment are examined in detail in Chapter 6, the present study engages with the complexity of the evaluator in the richness of a naturalistic setting. In particular, it focuses on two aspects of the evaluator as they relate to the formation of performance quality decisions: the affective state and the aesthetic judgement.

5.1.1 Affective states in music perception

The effects of mood and affective state on high-level cognition, interpretation, decision-making, and reasoning across domains has been well documented (see Blanchette & Richards, 2010, for a review). Particular focus has been given to the role of anxiety in increasing risk-avoidance behaviours and perceived negative outcomes (e.g. Yuen & Lee, 2003), and while few systematic patterns have been seen resulting from specifically emotional states such as happiness or sadness, some research has indicated their situationally-specific influence on decision-making (Bodenhausen et al., 1994; Park & Banaji, 2000). Thus, it seems straightforward to assume that one's state can influence music performance evaluation. McPherson and Schubert (2004) agreed with this view in their review of the music performance evaluation literature, speculating that "the mood of the assessor probably has some effect on his or her adjudication" (p. 72). At the time of their writing, they had no specific literature to cite directly supporting this idea, which very much remains the case. However, there are numerous examples in related literature to which one can refer (Schubert, 1996; Flôres & Ginsburgh, 1996; Glejser & Heyndels, 2001; Bergee & McWhirter, 2005; Dahl & Friberg, 2007; Chapados & Levetin, 2008; Juslin & Sloboda, 2009; Danzinger et al., 2011; Brattico et al., 2013; Juslin, 2013; Baltes & Miu, 2014; Quinto et al., 2014a). Schubert (1996) suggested that listeners will become bored and disengaged in listening to a series of musical performances lacking in variation. The serial effect in which later performances in a sequence are judged more favourably, or differences in judgements at varying times in the workday, has been suggested to be partly due to a reduction of mental resilience as mood and arousal fluctuate (Flôres & Ginsburgh, 1996; Glejser & Heyndels, 2001; Bergee & McWhirter, 2005; Danzinger et al., 2011).

Crucially, McPherson and Schubert (2004) mention the literature highlighting music's capacity to affect emotional state and perception as a strong reason to presume that a link could be expected between affect and evaluative decision-making in music. This area has seen considerable growth in the intervening time, in contrast to the evaluation literature (Juslin & Sloboda, 2009; Juslin, 2013). This has included how affective response continuously changes over the course of the performance (see Chapter 2 and Geringer et al., 2004, for a review), how visual information conveyed by a performer can affect perceptions of conveyed emotions such as happiness, sadness, and anger (Dahl & Friberg, 2007; Quinto et al., 2014b) and affective physiological responses (Chapados & Levetin, 2008), and how the interplay of pre-attentive and conscious neural processes are informed by individual tastes and aesthetic judgement (Brattico et al., 2013). Individual empathy, use of visual imagery, and mood have also been shown to influence emotional reactivity in a live operatic performance (Baltes & Miu, 2014).

An important distinction to make in this research is the difference between perceived and felt emotion (Gabrielsson, 2002; Kallinen & Ravaja, 2006), in which listeners are able to distinguish between the emotion they believe the music intends to evoke, and their subjective experience of emotions as a result of the music. This is typified by the experience of people experiencing happiness while listening to 'sad' music, and vice versa (Kawakami et al., 2013). Of the two, felt emotions tend to be stronger predictors of enjoyment ratings than perceived emotions, although the distance between the two constructs also plays a role in the formation of preference judgements (Schubert, 2007). One can then consider the third perspective of the composer's own intent in what emotion the music should evoke, achieved by common cues in musical content (Schutz, 2017) but not always corresponding to listeners' perceptions (Thompson & Robitaille, 1992). The Extended Lens Model (ELM) of musical communication posits a set of shared and idiosyncratic acoustic cues employed by composers and performers to trigger affective reactions in listeners

(Juslin & Lindström, 2010), cues that have been found to be linearly additive in their effects on perceived emotions (Eerola et al., 2013). This is itself complicated by the active expressive intent of the interpreter (Gabrielsson, 1996) and that different perceptions are reported by listeners of different ages (Stacho et al., 2013). In short, the relation between mood, emotion, and perception is exceptionally complex in musical contexts. This makes the examination of mood-related effects on music performance quality judgements particularly challenging, especially when compared with research examining decision-making in contexts where the stimulus is more neutral and does not automatically trigger affective reactions, such as when judging consumer products or perceptual stimuli. It is perhaps unsurprising that the relation between mood and quality judgements are not yet well understood. One must not only understand how affective state may influence a rating, but how the change of affective state caused by the very stimulus under scrutiny could play a role as well.

5.1.2 Aesthetic versus evaluative response

Related to the evaluative decision is the aesthetic one, a subset of which includes whether the listener enjoyed the performance. Gabrielsson and Lindstrom-Wik (2003) examined what they described as listeners' *Strong Experiences in Music* (SEMs). In their research, synchrony between music and mood was found to link strongly with such experiences, particularly when happiness was reported but less so for feelings of anxiety, discomfort, and sadness. A strong factor predicting these experiences was immersion and absorption in the experience, a focus of recent study in musical-emotional reactions that has been found to exist independently of an individual's musical training or empathy scores (Sandstrom & Russo, 2013). As mentioned above, Schubert (2006) found that felt emotions better predicted enjoyment responses than emotions perceived in the music. Cognitive factors can also play a role; Margulis (2010) found in her study of programme notes that providing listeners with text descriptions of the musical structure reduced their enjoyment rating of a performance. Research by Thompson (2006) examined audience experience in a live concert setting. Enjoyment and quality ratings by the performers correlated strongly, but audiences were nevertheless shown to be capable of separating their cognitive and affective response to the performances to some degree. The audience's familiarity with each work was not predictive of their enjoyment or quality ratings, although their liking for the composition did positively correlate. The overall enjoyment of the concert was also measured, and regression analyses found that enjoyment of the two performances, combined with liking of the concert venue, accounted for only 39% of the total variance in the total enjoyment score, indicating that other factors beyond the enjoyment of the specific works were contributing to the overall decision.

This relation between perceived quality and enjoyment judgements has been found in other research (e.g. Hargreaves et al., 1980) and suggests that these decisions, while related, represent distinct features that can be quantified and delineated. The nature of the causal relationship between these two factors remains unclear, however, and several authors have suggested that quality perception may exist as a subset of an aesthetic evaluation, but not vice versa (McPherson & Thompson, 1998; McPherson & Schubert, 2004; Thompson, 2006). A *quality-informing-enjoyment* causal route is not hard to postulate: one can easily imagine enjoying a well-executed performance more than a poorly-played one. On the other hand, an *enjoyment-informing-quality* model is conceivable but more complex. When an audience is split over the relative strength and value of a particular performer's ability (any controversial performer of choice can be inserted here), one can assume that personal differences between evaluators are at play. But are these differences stemming from the degree to which they are enjoying the performance? Or is a third variable, their preference or *liking* of a performance or performer, driving both quality and aesthetic judgements, as seen in the cross-correlations by Thompson (2006)? Unravelling these interrelations and the direction of causal influence is not simple, particularly due to the difficulty of experimentally manipulating the relevant variables. One can manipulate the objective quality of a performance without affecting the underlying aesthetic nature of the composition by introducing inconsistencies and errors in the performer's interpretation and technique. Changing an individual's preferences and tastes is another matter. Thus, correlational methods will continue to be used to examine this topic, and caution taken to avoid assuming an *enjoyment-informing-quality* model of causality.

5.1.3 Aims of the present study

The present research aimed to examine the nature of the evaluator in terms of his or her affective and physiological state when reporting enjoyment and performance quality. To provide an initial examination of how these relationships manifest in the complexity of a true performative setting, this study was conducted in the context of a live concert. It investigated three study-specific research questions (RQs) that expanded upon the fifth overarching thesis research question concerning qualities of the evaluator:

RQ1. Does a listener's self-reported affective state before a live concert, at the point of completing the final evaluation, or the change between the two predict his or her *quality rating* of a performance?

RQ2. What is the relationship between the likeability and familiarity of the composition and perceived quality of a live performance?

RQ3. What is the relationship between aesthetic and quality judgements of a live performance?

Finally, as the data required to answer the above questions would allow a comparative examination of the predictive value of affective state on qualitative versus aesthetic judgements:

RQ4. Does a listener's self-reported affective state before a live concert, at the point of completing the final evaluation, or the change between the two predict his or her *enjoyment* of a performance?

To investigate these questions, this study used an alternate approach to the experimental designs employed in the previous chapter. While the paradigms used thus far provided several benefits, most notably in increased control of sampling, stimuli, protocol, and experimental group randomisation, they took place in a laboratory setting. This study, by contrast, was conducted in a live setting in which no control could be exerted on the performance (neither the repertoire nor the performers), environment, or concertgoers who chose to attend. A self-report survey design was thus employed to capture the complexity of reactions in this setting, and to maximise the number of simultaneous participants. The latter goal was crucial to establishing statistical power; while no effect size precedents exist in the music evaluation research that could allow for comparative power analyses, it could be assumed that any effects of mood state in an uncontrolled setting would be subtle and thus a robust sample size would be required to determine the significance of and interaction between the various factors in question. As the concert audience was projected to be relatively large (~600), a simplified rating procedure was used with the goal of engaging with at least half of those in attendance. In this case, a single, post-performance written quality rating approach was used alone, without the continuous measures methodologies (see Chapter 2 for a discussion of the approaches) employed in the previous two empirical studies. The temporal and process-based theme of this thesis, first examined using continuous measures, was instead carried through by having the audience members complete the affective survey items before and after the performance. This allowed for examination not only of *how*, but of *when*, affect had an impact on the final quality judgement.

5.2 METHOD

The study took place at a professional choral concert by the Eric Whitacre Singers at Union Chapel, London. The programme comprised original compositions and arrangements, primarily by Eric Whitacre, including both a cappella and accompanied works and all employing a standard harmonic language.

5.2.1 Participants

Three hundred participants volunteered to take part in the study, drawn from an audience of approximately 560. This sample did not include approximately 100 questionnaires with three or more pieces of missing data or with a reported age below 18 years, both of which were rejected. The sample represented approximately 53% of the full audience in attendance, comprising 111 men and 188 women (one not reporting) with a mean age of 42.89 years (SD ± 16.31, range = 18 - 82). 241 (80.3%) reported having experience of playing a musical instrument or singing, among whom a mean 27.02 years of musical experience was reported (SD ± 17.13, range = 1 - 70). Overall, participants reported attending a mean of 6.53 “concerts like this one” per year (SD ± 11.67, range = 0 - 130) and gave a mean score of 6.10 out of 10 on overall familiarity with the music on the programme (SD ± 2.88, range = 1 - 10). The front page of the questionnaire explained that participation was voluntary and that, by completing the questionnaire, they were providing informed consent to take part in the research. Ethical approval was granted by the Conservatoires UK Research Ethics Committee and conducted according to the ethical guidelines of the British Psychological Society.

5.2.2 Materials

A custom survey was designed to capture audience mood states and perceptions of the performance (see Appendix 7 for the full survey). To maximise ease of distribution and participant response and to minimise disruption of the concert setting, the survey was designed to be as short as possible, printed on two sides of a single sheet of A4 paper. The first side carried instructions to be completed before the concert began, while the second was to be completed at the start of the interval. The pre-concert side collected demographic information including whether the participant played an instrument or sang, how many concerts they attended in an average year, and their general familiarity with Eric Whitacre's music. Participants then completed a series of 10-point scales assessing their current affective states from *not at all* (1) to *very* (10). Items were assembled in collaboration with parallel studies examining affective and biological responses to making music in cancer patients and carers (Fancourt et al., 2016) and in concert attendance by audience members at the present concert (Fancourt & Williamon, 2016). Seven mood items (*happy, sad, afraid*, *confused*, *angry, tired, energetic*) were adapted from existing psychometric scales (Arruda et al., 1996; Stern et al., 1997) by choosing those items corresponding to potential musical reactions, to which was added *connectedness to others* (Fancourt et al., 2016) to capture the social-affective element of being part of a live concert. Four items of anxiety were also included: *tense*, *relaxed*, *anxious*, and *stressed* (from Kim, 2008). These data were collected in conjunction with a larger project examining physiological responses of the choir itself (Fancourt et al., 2015) and of a separate sample of the audience (Fancourt & Williamon, 2016).

The second side was completed at the interval to maximise the number of respondents and opportunities to collect the questionnaire. It included the same mood scales as above. It also included three questions assessing affective perceptions of the concert (*stimulating*, *meaningful*, *enjoyable*). Finally, participants were asked to consider specifically the first work in the programme to allow a fine-grained examination of factors relating to the repertoire. In a simplified version of the questions from Thompson (2006), participants rated *quality of the performance*, *how much you enjoyed the performance*, *your familiarity with the piece*, and *how much you like this piece* from *low* (1) *to high* (10). All scales used the same 10-point system to ensure commensurability of the results.

5.2.3 Procedure

A survey and pencil were placed on each seat prior to the concert. Audience seating for the concert was unassigned. Ten minutes prior to the performance, the conductor took the stage to explain briefly the research being undertaken (Fancourt et al., 2015; Fancourt & Williamon, 2016; the present sample did not participate in the other projects) and procedure. Following the first half they were immediately reminded by the conductor to complete the second side of the survey, which were then collected over the interval and at the end of the concert. A team of researchers was on hand wearing distinctive clothing to collect the surveys and answer any questions of the participants. Figure 5.1 depicts a flow diagram of the research procedure and the points in time at which each set of data were collected.

5.2.4 Data treatment and analyses

Questionnaires with 1-2 missing data points were allowed, thus n-values vary slightly between tests (each of which were conducted excluding cases listwise) and are reported where appropriate. To examine the first research question, t-tests were used to examine initial changes in mood state across the concert, followed by principal axis factor analysis to reduce the 12 items to fewer dimensions for further analyses. Multiple regression was then employed using the *pre-concert* (X_{pre}), *interval* (X_{int}), and *changed* values (X_{Δ}) of the resulting factors to determine whether any served

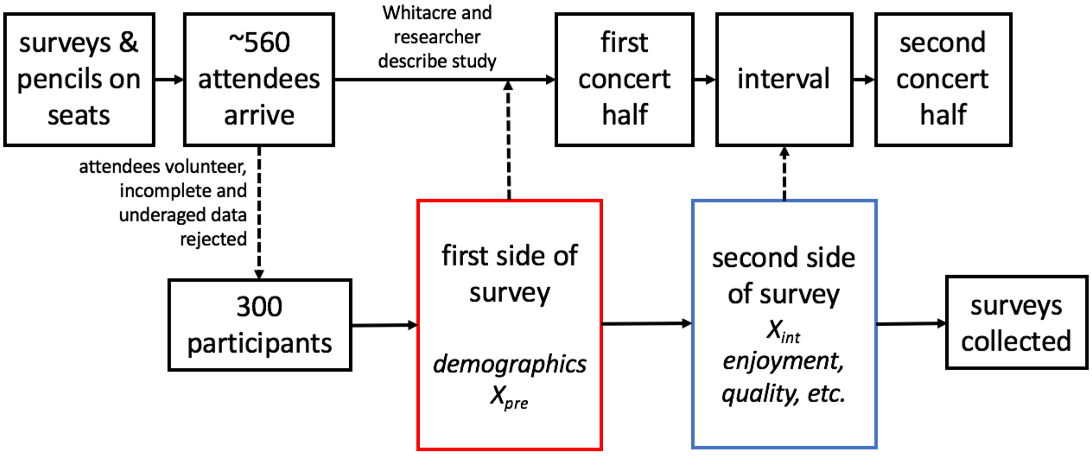


Figure 5.1. Flow diagram of the Study 3 research design. 300 participants responded to the surveys. The first half of the items were completed before the concert began (red) and the other half were completed at the concert interval (blue). See Appendix 7 for the complete survey.

as predictors of quality ratings of the first work. The second and third research questions investigating the relationship between likeability, enjoyment, familiarity, and quality ratings of the first work were examined using correlation and regression analyses specific to the central question. While mediation/moderation analyses or Structural Equation Modelling (SEM) were considered, they were not employed due to ambiguity in the direction of causation among the factors (Baron & Kenny, 1986; Ringle et al., 2012), particularly between enjoyment and perceived quality (McPherson & Thompson, 1998; McPherson & Schubert, 2004; Thompson, 2006). For the fourth research question, examining the relationship between affective states and aesthetic ratings, the same approach to research question 1 was used with the enjoyment rating as the dependant variable in the multiple regression analysis. Normality varied across the dataset, although as the sample size was large (300) parametric tests were used unless noted otherwise with appropriate caution given to the significance values and generalisability.

5.3 RESULTS

The mean quality rating of the first work, the principal focus of this study, was high (M = 9.23, SD ± 1.16, range = 4 - 10), thus highly negatively skewed and demonstrating a ceiling effect. The data demonstrated many tied ranks, precluding standard data transformation techniques. For this reason, all correlation analyses were conducted non-parametrically using Kendall's tau. Preliminary analyses found no significant relation (and correlation values of $\tau < .2$) between quality rating and participants' age, sex, musical experience, and concerts attended per year, thus these demographic factors were excluded from further analyses.

To address the first research question, descriptives for and changes in individual mood states are first reported, follow by factor reduction and multiple regression analysis. The sections that follow then address each of the remaining three research questions as described above in Section 5.1.3.

5.3.1 Descriptives and changes in mood state

Paired-sample t-tests were used to calculate differences between the pre-post mood scores. Results are shown in Table 5.1 and Figure 5.2. Participants' ratings of *energetic*, *tense*, *anxious*, and *stressed* dropped significantly ($p < .001$; due to multiple comparisons, only effects $p < .004$ were considered significant with a Bonferroni-corrected cutoff) by at least one half a scale point and with effect sizes above Cohen's *d* of 0.20 (considered to be a small effect size), while their scores for *relaxed* significantly rose by a mean of 1.23 points ($p < .001$). Their ratings of *sad, angry, tired, happy*, and *connected* also changed significantly (ps = .026 - .002) although effect sizes were small (Cohen's $d < 0.2$) and significance values of $p > .004$ should be interpreted with caution considering a Bonferroni control for multiple comparisons, thus leaving *sad* and *angry* with notable changes. Correlations within each pairing were significant and moderate, with *connected to others* showing the highest level (see Table 5.1).

Table 5.1. Means of and differences between *pre-concert* and *interval* mood scores. Measured using paired-sample t-tests. Significant differences with an effect size *d* > 0.20 are highlighted. Sample sizes and correlations within each factor are reported.

	Pre-concert		*Interval*		*Change*					
	M	*SD*	*M*	*SD*	*M*	*t*	*p*	*d*	*n*	*τ**
Afraid	1.30	0.87	1.28	0.96	-0.02	0.34	.733	0.02	295	.38
Confused	1.64	1.37	1.66	1.41	0.02	-0.22	.826	0.02	294	.55
Sad	1.84	1.42	2.13	1.67	0.29	-3.06	.002	0.18	295	.46
Angry	1.62	1.50	1.39	1.16	-0.23	2.98	.003	0.18	293	.52
Energetic	**4.53**	**2.09**	**3.42**	**1.82**	**-1.11**	**9.00**	**<.001**	**0.53**	**288**	**.43**
Tired	4.99	2.21	4.67	2.25	-0.32	2.26	.025	0.13	298	.42
Happy	6.77	1.82	6.59	2.06	-0.18	1.45	.149	0.09	295	.44
Tense	**3.32**	**2.26**	**2.34**	**1.77**	**-0.98**	**7.76**	**<.001**	**0.46**	**296**	**.44**
Relaxed	**5.62**	**2.36**	**6.88**	**2.23**	**1.26**	**-8.68**	**<.001**	**0.51**	**293**	**.42**
Anxious	**2.63**	**2.00**	**2.04**	**1.70**	**-0.63**	**5.57**	**<.001**	**0.33**	**292**	**.54**
Stressed	**3.46**	**2.46**	**2.27**	**1.84**	**-1.19**	**9.87**	**<.001**	**0.59**	**298**	**.56**
Connected	5.74	2.34	6.05	2.40	0.31	-2.70	.007	0.16	296	.65

*all significant at p < .001

Sex differences were examined using independent-sample t-tests, again controlling for multiple comparisons with a Bonferroni correction. The only notable significant differences were in the *connected to others* item, in which female respondents' *pre-concert* (M = 6.14, SD ± 2.24, n = 185) and *interval* (M = 6.37, SD ± 2.35, n = 186) scores were significantly higher (*pre-concert:* $t_{(294)} = -3.72$, p < .001, *d* = 0.43; *interval*: $t_{(295)} = -3.03$, p < .005, *d* = 0.35) than those of the men's scores (*pre-concert*: M = 5.11, SD ± 2.37, n = 111; *interval*: M = 5.50, SD ± 2.40, n = 111) by almost 1 point. The amount of change in *connected* over the course of the concert did not differ significantly between sexes, suggesting a higher overall trait value than differences in reaction to the concert.

5.3.2 Item reduction: Affective versus physiological

Inclusion of all 12 items as *pre-concert* (X_{pre}), *interval* (X_{int}), and *change* (X_{Δ}) variables in a single multiple regression would have resulted in 36 independent

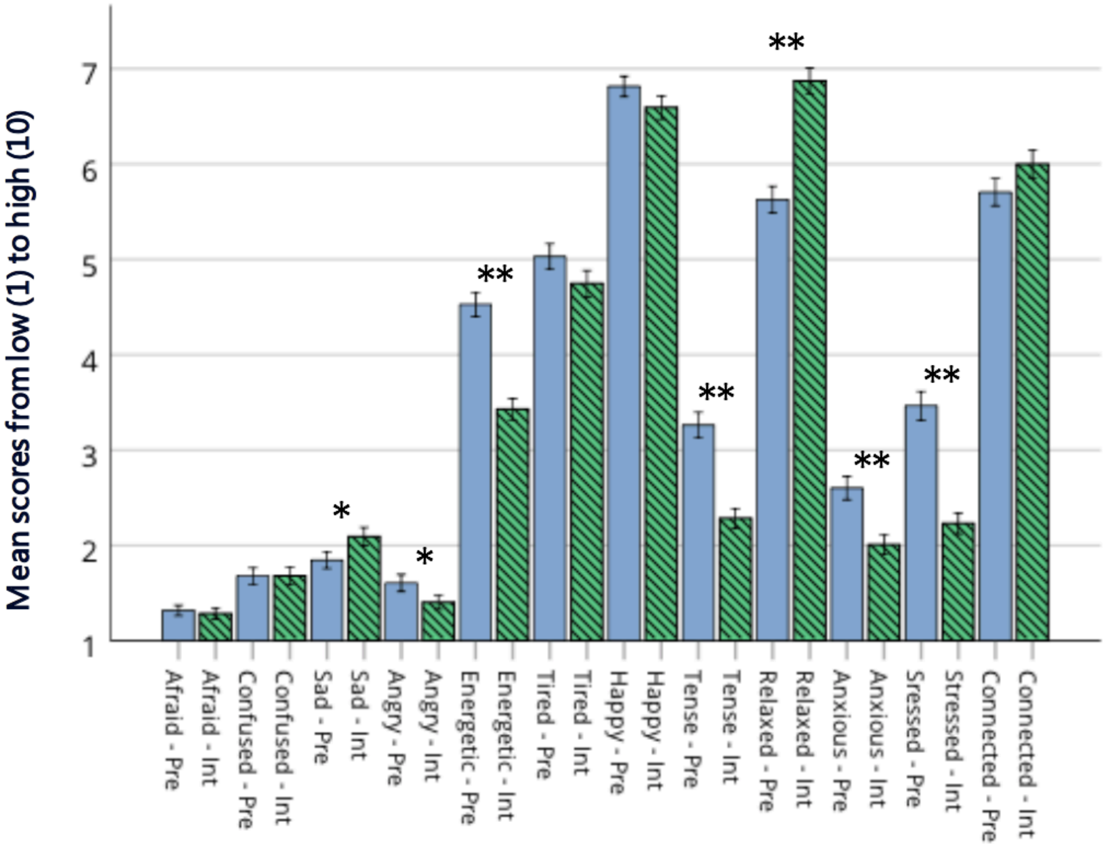


Figure 5.2. Means of and differences between *pre-concert* and *interval* mood scores. Blue = pre-concert scores; green = scores at the interval. Asterisks mark significant change across the two ratings; * = p < .005, ** = p < .001, as measured using paired-sample t-tests. Error bars show +/- 1 SE.

variables, an unwieldy number for the given sample size. Thus, an exploratory (principal axis) factor analysis was conducted on the mood scores taken prior to the concert start to reduce the number of factors for analysis. Principal axis factor analysis was chosen over other approaches (e.g. maximum likelihood analysis) due to the former's strength in exploring datasets with few indicators per factor (de Winter & Doudou, 2012), which would be likely with just 12 items overall. As the items and resulting factors were not assumed to be independent an oblique (direct oblimin) rotation was applied. A covariance matrix was used as each item was measured on the same 10-point scale, thus they were commensurable. The Kaiser-Meyor Olkin measure confirmed an adequate sample size (KMO = .77), and Bartlett's measure was significant (p < .001) indicating an acceptable minimum degree of correlation within the item matrix. The analysis revealed three factors with eigenvalues larger than 1, together explaining 51.48% of the variance. Examination of the scree plot confirmed a clear 3-factor structure. The factor loadings following rotation can be found in Table 5.2, which showed clusters around what could be described as self-reported physiological states (Phys: *tense*, *stressed*, *anxious*, *relaxed*, *tired*), positive affect (PosA: *happy*, *connected*, *energetic*), and negative affect (factor 3; NegA: *sad*, *afraid*, *confused*, *angry*). Reliability analyses were conducted with Cronbach's α, following a reverse-scoring of the *relaxed* item necessitated by its negative loading score within factor 1 (see Table 5.2), indicating moderate-to-high values (αs = .84, .65, and .72).

Table 5.2. Summary of principal axis factor analysis conducted on the pre-concert mood scores. A clear 3-factor structure was found.

Items	***Rotated factor loadings***		
	Self-reported physiological (Phys)	*Positive affective (PosA)*	*Negative affective (NegA)*
Tense	**.89**	.06	.02
Stressed	**.87**	.02	.03
Anxious	**.58**	.02	.24
Relaxed	**-.52**	.33	.03
Tired	**.41**	-.13	.06
Happy	.12	**.82**	-.19
Connected	-.07	**.54**	-.01
Energetic	-.08	**.52**	.09
Sad	-.07	-.18	**.80**
Afraid	.03	.05	**.57**
Confused	.10	.06	**.52**
Angry	.24	-.05	**.45**
Eigenvalues	20.22	6.01	3.84
Percentage of variance	44.50	13.23	8.45
α	.84	.65	.72

Note: Factors greater than .40 are highlighted in bold.

5.3.3 RQ1: Affective states as predictors of quality judgement

Following the establishment of the three mood factors – self-reported physiological (Phys), positive affective (PosA), and negative affective (NegA) – means were calculated for the *pre-concert* (e.g. Phys_{pre}), *interval* (e.g. Phys_{int}), and *change* values (e.g. Phys_{Δ}), again reverse-scoring the *relaxed* item, resulting in 9 total factors. A multivariate multiple regression analysis was conducted with these nine items as independent variables and with the quality rating of the first work (M = 9.23, SD ± 1.16) as the dependant variable. The correlation matrix revealed a degree of multicollinearity (unsurprising due to the interrelated nature of the items) although none were high (rs < .70). The analysis produced a significant model ($F_{(6,289)} = 6.00$, $p < .001$, $R^2 = .09$) accounting for 9% of variance in the quality rating. The analysis excluded all three *pre-concert* scores from the model due to a lack of predictive power (see Table 5.3). Of the remaining independent variables, the only significant predictor proved to be the aggregated positive affect score taken at the interval, when the quality rating was recorded (PosA_{int}; b = 0.13, *ß* = 0.19, p < .01) in which an increase of one point on the 10-point scale comprising *happy, relaxed,* and *energetic* predicted an increase of .13 points in the 10-point performance quality rating.

5.3.4 RQ2: Relationships between perceived quality, enjoyment, familiarity, and likeability of the work

To determine the interrelationship between enjoyment, familiarity with, and likeability of the first work with perceived quality, correlation analyses were first conducted. As with the initial *quality* rating reported above (M = 9.23, SD ± 1.16), participants gave high scores on *enjoyment* (M = 8.99, SD ± 1.43) of the performance. Liking of the composition itself was ranked highly (M = 8.69, SD ± 1.65) and familiarity with the work varied widely with a mean score of 5.64 (SD ± 3.73, range = 1 - 10). Ratings of *quality, enjoyment,* and *likeability* correlated strongly, with medium-to-weak correlations between *enjoyment*, *likeability*, and *familiarity* but no significant correlation between familiarity and quality when accounting for multiple comparisons (see Table 5.4).

Table 5.3. Regression model predicting performance quality rating of the first work. Only the positive affect score taken at the interval ($PosA_{int}$, highlighted in bold) significantly predicted the final quality score.

Item	*b*	*SE B*	*ß*	*p*
Constant	9.02	0.40		< .001
$Phys_{pre}$	-	-	-	-
$PosA_{pre}$	-	-	-	-
$NegA_{pre}$	-	-	-	-
$Phys_{int}$	-0.12	0.07	-0.14	.092
$PosA_{int}$	**0.13**	**0.05**	**0.19**	**.009**
$NegA_{int}$	-0.11	0.10	-0.10	.272
$Phys_{\Delta}$	-0.05	0.06	-0.06	.401
$PosA_{\Delta}$	-0.04	0.05	-0.05	.436
$NegA_{\Delta}$	-0.01	0.10	-0.01	.945

Table 5.4. Correlations (Kendall's tau) between *interval* questionnaire items relating to the first piece (n=299).

	Quality	*Enjoyment*	*Familiarity*	*Likeability*
Quality	-	**.71****	.12	**.53****
Enjoyment		-	**.24****	**.66****
Familiarity			-	**.39****
Likeability				-

** p < .001

As described in the introduction, previous authors have argued that the perception of quality represents a subset of a judgement of *enjoyment*, but not vice versa (McPherson & Thompson, 1998; McPherson & Schubert, 2004; Thompson, 2006). While the correlation value of .71 suggested some degree of independence, there was clearly a large degree of interrelation. For this reason, a multiple regression analysis was conducted to determine the predictive power of *familiarity* and *likeability* of the work on the *quality* rating, while excluding *enjoyment*. This provided a significant model ($F_{(2,296)} = 6.00$, $p < .001$, $R^2 = .35$) explaining 35% of the variance. The contribution of each of the individual predictors, however, was somewhat contradictory. While the *likeability* of the work provided a strong and significant contributor (b = 0.46, *SE* B = 0.04, *ß* = 0.66, p < .001) in which an increase of one point on the *likeability* scale predicted an increase of .46 points on the *quality* scale, the *familiarity* variable indicated a significant *negative* effect (b = -0.05, *SE* B = 0.01, *ß* = -0.16, p < .005) in which an increase of one familiarity point contributed a drop of only .05 points out of 10 in the *quality* rating. Taken with the borderline but non-significant correlation value, it may be fair to assume an over-powered statistical test and a non-meaningful relationship between familiarity with the work and the overall quality rating in this case.

5.3.5 RQ3: Relationships between quality and aesthetic judgements

To provide a broader perspective on the relationship between quality judgements and aesthetic ratings, the survey included three measures of aesthetic judgement (*stimulating*, *meaningful*, and *enjoyable*) referring to the entirety of the first half of the concert rather than the specific first work. While it cannot be assumed that the quality rating of that work was indeed independent of how they felt about the aggregate of the pieces they had so far heard, participants did respond to this overall *enjoyable* score (M = 8.04, Med = 8.00, SD ± 1.82) significantly lower (T = 15,953.50, p < .001, r = .39, n = 297; measured via a related-samples Wilcoxon signed rank test) than the *enjoyed* score for the first work (M = 8.99, Med = 10.00, SD ± 1.44). This implies that concertgoers were able to separate their perceptions of the first work from perceptions of the first half despite having heard additional works in the intervening time, at least in terms of *enjoyment*. Participants also rated the first half of the concert as highly stimulating (M = 7.03, SD ± 2.02) and meaningful (M = 6.97, SD ± 2.11).

Taken together these measures showed low-to-medium correlations (see Table 5.5). Notably, correlations between *quality* and the affective ratings, including *enjoyable*, were low (τs = .25 - .36). This is in contrast with the high correlation (τ = .71) between *quality* and *enjoyment* rated above. For this reason, one could speculate that the audience members were able to separate these overall aesthetic responses from the task of assessing the quality rating of the first piece and that these aesthetic feelings may contribute to how the act of committing a first rating was conducted. Multiple regression with the three aesthetic responses as independent variables and the first-work *quality* rating as the dependant variable resulted in a significant model ($F_{(2,296)} = 6.00$, $p < .001$, $R^2 = .35$), where *stimulating* and *enjoyable* appeared as significant predictors, but *meaningful* did not (see Table 5.6).

5.3.6 RQ4: Affective states as predictors of enjoyment

While not directly addressing the topic of quality judgement, the data collected allowed for an examination of the changes in self-reported affective state (i.e. self-reported physiological, positive affective, and negative affective) against the overall *enjoyable* score. This overall rating was chosen rather than the *enjoyment* score of the first work to give a more comprehensive view of the audience's judgement of the performance and, as reported above (see Section 5.3.4), this score showed a much lower correlation ($\tau = .36$ versus .71) with the quality rating of the first work. As such, any similarities in the findings of affective states and their relationships to enjoyment versus quality ratings could not be so easily attributed to multicollinearity of the two

Table 5.5. Correlations (Kendall's tau) between *interval* questionnaire items relating to aesthetic response to the entire first half and quality ratings of the performance of the first piece. N values are reported.

	Quality	*Stimulating*	*Meaningful*	*Enjoyable*
Quality	-	**.31****	**.25****	**.36****
Stimulating	295	-	**.57****	**.54****
Meaningful	296	296	-	**.55****
Enjoyable	297	296	297	-

** p < .001

Table 5.6. Regression model predicting quality rating of the performance of the first work based on aesthetic responses to the first half of the concert. Significant predictors are highlighted in bold.

Item	*b*	*SE B*	*ß*	*p*
Constant	6.71	0.27		< .001
Stimulating	**0.14**	**0.04**	**0.25**	**.001**
Meaningful	-0.02	0.04	-0.04	.580
Enjoyable	**0.21**	**0.05**	**0.33**	**< .001**

dependant variables. A regression analysis was conducted using the same approach as in research question 1 (see section 5.3.3). The correlation matrix revealed a degree of multicollinearity, although none were high (rs < .70). The model was significant ($F_{(6,288)} = 24.40$, $p < .001$, $R^2 = .32$) explaining 32% of the variance in the model (in contrast to the 9% found in the quality rating model). The contribution of the independent variables largely replicated the findings of the quality rating examination in that all three *pre-concert* scores (Phys_{pre}, PosA_{pre}, NegA_{pre}) were excluded from the model due to lack of predictive value, and positive affective state at the interval was a significant predictor (see Table 5.7) with every point of increase on the aggregate scale indicating an increase of .49 points on the *enjoyable* scale. In contrast to the quality rating, however, negative affective state was also a significant predictor, indicating a fall of .41 points on the *enjoyable* scale for every increased point on the accumulated negative affect scale. Finally, while neither pre-concert (Phys_{pre}) or interval arousal (Phys_{int}) states were predictive, the change in state (Phys_{Δ}) showed a significant negative relationship, with every point of decreased arousal across the concert indicating an increase of .33 points on the *enjoyable* scale.

Table 5.7. Regression model predicting the *enjoyable* rating of the full first half of the concert. Significant predictors are highlighted in bold.

Item	*b*	*SE B*	*ß*	*p*
Constant	5.77	0.54		< .001
Phys_{pre}	-	-	-	-
PosA_{pre}	-	-	-	-
NegA_{pre}	-	-	-	-
Phys_{int}	0.01	0.09	0.01	.904
PosA_{int}	**0.49**	**0.07**	**0.44**	**< .001**
NegA_{int}	**-0.42**	**0.13**	**-0.23**	**.002**
Phys_{Δ}	**-0.33**	**0.07**	**-0.27**	**< .001**
PosA_{Δ}	-0.05	0.07	-0.04	.528
NegA_{Δ}	0.17	0.10	0.08	.218

5.4 DISCUSSION

This study examined an audience's evaluative and aesthetic reactions to performance in a live concert setting and relationships with their changing affective state over the course of the performance. The researcher had no control over the venue, programme, or performers, nor whether the audience members chose to attend the concert. The audience was self-selecting in whether or not they chose to complete the provided survey, although over half of the audience in attendance did so and the demographics revealed a wide variety of age and musical experience, both in performance and concert attendance. This audience was asked to report mood state on 12 items before and after the performance and provided evaluative and aesthetic ratings of the performance at the interval. The results are discussed in terms of the research questions they addressed.

5.4.1 Research questions 1 and 4: Mood and judgement

Research questions 1 and 4 together examined whether self-reported mood states reported before and after one half of a concert performance, as well as the change between these scores, could predict quality ratings (RQ1) and enjoyment (RQ4). To determine this, analyses were conducted that reduced the twelve mood items to three underlying factors, termed in this study as physiological (*tense*, *anxious*, *stressed*, *relaxed*, *tired*) positive affective (*happy*, *energetic*, *connected to others*) and negative affective (*angry*, *sad*, *afraid*, *confused*). None of the pre-concert factors, nor changes within these factors, proved to be significant predictors of quality ratings or enjoyment. The only predictor was the positive mood state at the time of the evaluation. This time difference is meaningful as the work in question was at the start of the concert programme, with approximately 45 minutes of musical material following, thus the pre-concert mood state was recorded much closer to the time of hearing the performance in question, while the second set of scores describe the mood state at the time of the evaluation following the experience of not only hearing the first piece, but the ones that followed. Thus, audience members had to recall how they felt about the first piece when forming their judgements. The significant relationship between this quality assessment and positive mood prior to the performance could thus be accounted for by a sort of mood-based recency effect, akin to the memory-based recency effects show in other domains (e.g. Fredrickson & Kahneman, 1993; Varey & Kahneman, 1992) and discussed in Study 1. Rather than mood state at the time of hearing the piece (i.e. before the performance began), or the degree to which mood changed over the course of performance, it was the mood at the time of recalling the performance and forming a judgement that may have informed the evaluative process. Why then might positive affect, and not negative or physiological, have caused this? One line of research has found that increased state happiness can lead to greater adherence to stereotypical thinking in social judgements (Bodenhausen et al., 1994; Park & Banaji, 2000). Perhaps audience members experiencing increased positive affect were more likely to confirm to the stereotype that a world-class performance group was performing at a peak level. As a wealth of previous studies have found an effect on quality ratings of stereotyping based information given about the performer or composer (Duerksen, 1972; Radocy, 1976, Colley et al., 2003; North et al., 2003; Negut & Sârbescu, 2014), further research should examine whether this effect is moderated by states of positive affect.

No significant relationship on quality ratings was found with arousal in any of the pre-, post-, or change items. However, the descriptive analyses demonstrated that such scores were generally low, with mean scores for *tense*, *anxious*, and *stressed* in particular of less than 4 on the 10-point scale (*tired* scored approximately 5), and significantly reducing over the course of the concert to scores below 3 by the interval (*tired* did not significantly reduce). The act of judging the performance, whether passively or actively, was clearly not a stressful experience, nor was it predicted to be. In such contexts, performers perhaps need not worry whether their audience's arousal levels will influence their perceptions of performance quality. Performers may also benefit from the knowledge that, should their performance reduce their audience's perceived anxiety, this could lead to increased enjoyment of the experience, as was demonstrated in examining the fourth research question. Further study will be required to determine whether such effects generalise to other performance and evaluative situations, particularly the heightened scenarios of competitions and auditions where the evaluator may be under more pressure (see Chapter 7).

While it is tempting to assume the listener's increased positive mood states had a causal effect on the quality rating, without an experimental design such causality cannot be assumed. It is possible that those who perceived that the quality of the performance was higher were suitably impressed or enjoyed the performance more, in turn increasing their perceived happiness, connectedness, or energy. Further study could examine the question experimentally, although not without complication. One would need to manipulate the mood states of the evaluator. While methods exist to stimulate physiological arousal and anxiety, particularly using simulation (see Chapter 7 for a full discussion), inducing affective emotions can prove more challenging. Films and stories are the most effective so-called (and non-pharmaceutical) *mood induction procedures* (MIPs), with facial expressions, gifts, visualisation, social interaction, and of course music also used, although their efficacy can be limited in inducing positive emotions (Westerman et al., 1996). Also, as the results of the present study found that emotions following the performance, as opposed to those prior to, affected the rating, emotion would need to be induced following the performance but prior to the evaluation itself. An ecologically valid approach to this could be the social interaction between panel members or even with the performers themselves, which would in turn require simulation of that interaction to be experimentally controlled (see Chapter 7).

5.4.2 Research questions 2 and 3: Aesthetic judgements

Research questions 2 and 3 examined the relationship between evaluative and aesthetic judgements of the performance in terms of enjoyment, familiarity, and likeability of the first piece on the programme (RQ2) and whether the performance as a whole was considered enjoyable, meaningful, and stimulating (RQ3). Regarding the first piece, Thompson's (2006) results were supported in that familiarity had a negligible correlation with quality and enjoyment ratings. This was also in line with the findings of Study 1 of this thesis (see Chapter 3) in which the Chopin works of relative but varying familiarity showed no relation to the process by which the quality ratings were formed, which themselves correlated strongly while showing some independence. Likeability did show a strong correlation with both quality and enjoyment, again in supporting this research. While it cannot be discounted that audiences may decide they like a particular work more if they felt it was performed better, these results support a model in which the strong correlations between quality and enjoyment ratings are driven by a *quality-influencing-enjoyment* model and not the reverse (McPherson & Thompson, 1998; McPherson & Schubert, 2004; Thompson, 2006), with any apparent influence of enjoyment on quality ratings actually being caused by the evaluator's liking of the stimulus in question driving both judgements.

The findings of research question 3 further complicate this interaction. The regression analyses demonstrated that quality ratings show differing relationships between aesthetic judgements of the performance, where enjoyment of the overall concert (which showed low correlation with quality rating of the first work) was a moderate predictor of the quality rating. The degree to which the concert was stimulating was also significantly predictive of the quality rating, but the degree to which it was perceived as meaningful to the evaluator was not. Thus, just as quality ratings can be subdivided into constituent component segments (see Chapter 2) it should be remembered that aesthetic responses also represent a complex interaction of cognitive and perceptual experiences (Levinson, 1987; Thompson, 2007; Brattico et al., 2013), and future work should continue to engage with this complexity.

5.4.3 Directions for future research

While this study employed self-reported measures of affect and physiological state, future studies should also incorporate physiological measurements. Audience's continuous affective responses have been shown to correlate with the conductor's heart rate in live orchestral settings (Nakra & BuSha, 2014), and physiological responses collected in an audience subset in parallel to the present study found that concert attendance caused significant changes in hormones indicating a relaxation response (Fancourt & Williamon, 2016). Related work in the visual arts has found that the aesthetic judgements of visitors to an art gallery correlated with variability in heart rate and skin conductance (Tsacher et al., 2012). Examining how such physiological reactions relate to self-reported aesthetic, affective, and evaluative perceptions will be key to understanding the full role of the evaluator's states and traits in judging a musical performance.

The generalisability of the study is limited by the non-normal distribution of the primary dependant variable. The performance was considered to be of an outstandingly high quality, with a sizeable proportion of the audience providing the maximum possible quality rating. Such situations are also common in formal evaluative settings, where competition judges, audition panels, and examiners are asked to rate and differentiate between performers of the highest calibre (with varying degrees of success, as the research thus far discussed has shown). Nevertheless, the statistical models here presented must be interpreted with due caution, and further study will be required to determine whether they are generalisable to situations where the dependant variable of quality rating exhibits different (i.e. more normalised) distributions, and is not subject to the same ceiling effect. Performances varying in quality, as well as performer, composer, and genre, may help in this.

5.5 SUMMARY

This chapter focussed on the role of the evaluator in the process of forming music performance quality judgements, examining not only the products of evaluation but how they interacted with aesthetic judgements and the process of changing affective states. 300 audience members at a live professional choral concert self-reported their affective states prior to and following the first half of the concert, then provided quality and aesthetic judgements of the first work on the programme. Regression analyses found that positive mood following the performance, but not before or changing across, was predictive of the final quality rating. Change in arousal was also predictive of the enjoyment rating. Likeability of the work, but not familiarity with it, correlated with quality ratings, and a complex interrelation of evaluative judgements and enjoyment of the performance were demonstrated.

6 STUDY 4: THE ENVIRONMENT

6.1 INTRODUCTION

The previous study brought the examination of performance quality ratings, and the processes of forming them, to a live, naturalistic setting. Remaining in this domain provides an opportunity to give focus to the final element of the performance evaluation not yet examined: the evaluative environment. In comparison with the three features of *repertoire*, *performer*, and *evaluator*, the *environment* has been given comparatively little scrutiny. Each study thus far has considered the *repertoire* in terms of *familiarity* and *likeability*. The performer has been explicitly addressed in both Studies 1 and 2 in their execution of performance errors. The *evaluator* has been an active participant in each examination, with behaviour and decisions moderated by prior knowledge, musical experience, and affective states. The *environment*, however, has not been addressed.

In Studies 1 and 2 the ratings were conducted in a laboratory setting; participants judged audio and video on a laptop using custom-designed software and followed the instructions of the researcher present. While this method is able to isolate causal effects, it cannot speak directly to the situations in which evaluations are naturally conducted. Study 3, on the other hand, moved into a rich environment in which judgements of performance, whether passive or active, regularly occur. As opportunity to engage with those participants was limited, however, effort was focussed on examining features of the evaluator (i.e. their affective state and their aesthetic judgements). Thus, the present study examined a near-identical performance domain, this time concentrating on select features of the social and physical situation that could be examined without requiring experimental manipulation. It also took advantage of the methodological repetition to determine whether key findings of Study 3 could be replicated.

6.1.1 The social environment: Listening with others

In addition to the traits and states a listener brings to a concert, the social environment in which the evaluation takes place can also play a role. The nature and purpose of an assessment, whether in a competition, concert, exam, review, lesson, etc. is largely driven by the complex social agreements, expectations, and interactions between the people and groups taking part (McPherson & Thompson, 1998). While parallels can be drawn to related domains of sport and behavioural psychology to speculate that fellow concertgoers may have a direct effect on individuals' concert perceptions, links to music performance evaluation remain poorly understood (see Section 1.6.3.1). One study of particular note is the early work by Radocy (1976) in which, in addition to providing priming information relating to the performers as used by Duerksen (1972), he also gave a select group explicit (and manipulated) information that previous listeners had shown a preference for one of two works. Participants' evaluations tended to move in the direction of the manipulation, particularly in the context of piano recordings in contrast to those of trumpet or orchestral groups. Another key piece of research has found that hearing positive critical reviews of unfamiliar pop songs in a radio-listening paradigm increased participants' preference for the music (Silva & Silva, 2009). Of course, this effect required the explicit, verbal transfer of preference information from one hypothetical listener to another via the medium of the researcher, who was surely seen as an authority figure. It remains unknown whether such information could be conveyed nonverbally through the interactions on a social panel or in a live audience. Springer and Schlegel (2016) demonstrated a mixed effect in a laboratory setting, where applause of higher magnitude added to the end of concert band performances led to higher ratings of a march but lower ratings of a ballad.

A related area of music research examining the influence of fellow audience members' perceptions is that of so-called emotional contagion. Self-reported induced emotions have been shown to be mediated by knowledge of other listeners' responses in a large online survey (Egermann et al., 2009a), where presumed knowledge of others' responses caused participants to give relatively higher or lower valence ratings in the direction of the manipulation. In this case, information transfer was explicit. Simply listening to music in the presence of others versus alone has shown mixed effects. Some studies have found no effects on self-reported emotion (Sutherland et al. 2009; Egermann et al., 2011), one has found increased convergence of emotional responses in a live concert setting versus watching a recording of the same performance in a lecture hall (Coutinho & Scherer, 2017), and physiological measurements of skin conductance, associated with musically induced 'chills', were found to be more prevalent in solitary conditions than in groups (Egermann et al., 2011). As these results came from laboratory-induced social settings, it is unclear how they would translate to a live situation. A small selection of studies have examined these factors *in situ*, attempting to isolate trends and relationships in audience perception while maintaining ecological validity by studying audiences in genuine concert experiences. Pitts (2004) interviewed and provided questionnaires to concertgoers at a festival of British operatic music, finding that enjoyment of and immersion in the performances was intrinsically tied to feelings of allegiance to the genre and a desire for inclusion within a like-minded community. A subsequent study at a chamber music festival (Pitts, 2005) also highlighted the role of environmental factors in driving attendance and enjoyment, emphasising the need for the concertgoer to feel comfortable in the social and physical space. Recent research has looked at audience experiences of performances given by Australian performing arts companies (Radbourne et al., 2009; Radbourne et al., 2010a, 2010b; summarised in Radbourne et al., 2014). Through a series of focus groups and surveys, audiences were found to value live performances for their delivery of a shared experience with fellow concertgoers, for proximity to the performers, and for immersion in the experience. Recent research has continued to examine the pro-social benefits of and emotional-contagion within live performance settings (Ballantyne et al., 2014; Garrido & MacRitchie, 2018).

The research on emotional contagion is particularly salient to the discussion of performance quality judgement considering the findings of Study 3, in which an increase in positive affect was an indicator of increased performance quality ratings. Should emotional contagion be able to cause changes in an audience’s mood state in a live setting, and if the affect-quality relationship is causal, then one could expect a resulting change in quality rating. The applause at the end of each performance could serve as the mechanism by which information regarding emotional states and judgements is transferred, for applause represents an explicit communicator of appreciation which, as described above, has been shown to influence quality ratings. One of many questions that would need to be addressed in understanding this relationship is whether audience members have any explicit perception of how their neighbours or the audience as a whole are rating the performance, and whether they would assume this to be higher, the same as, or lower than their own assessments. As individuals may base such an assumption on differences in musical taste or relative perceptual ability, one could hypothesise that musical experience would play a role in this relationship.

6.1.2 The physical environment: Concert venues and extraneous variables

Despite the central role of the concert venue in musical performance, little research has been done on its effect on performance quality ratings. Just as a performer transfers information to audiences via aural and visual modes, the venue conveys its nature through its appearance and the degree to which it alters and changes the sound produced. While these acoustic influences on music production are easily perceived by laypeople and musical experts alike (Ueno & Tachibana, 2005; Galiana et al., 2012), no research has examined the effects of hall acoustic on performance evaluations. Related to both of these features would be seat location, as it alters the visual and aural perceptions of the work being performance (particularly in venues of poor or inappropriate design). Again, no research to date has examined this feature.

One tangential examination of the relationship between performance environment and audience perception can be found in Thompson’s (2007) examination of elements predicting enjoyment of performance. The researcher asked participants to rank the importance of factors taking place prior to and during a hypothetical concert as they impact their enjoyment of the performance. Principal component analyses outlined six underlying components relating what Thompson described as *music* (e.g. familiarity with, liking of, and anticipation for the music being played)*, self* (mood and relaxation states)*, environment* (familiarity with the performers and venue), *engagement* (e.g. absorption in the performance), *dynamic* (e.g. distractions, wrong notes, poor acoustics, a poor seat), and *background* features (e.g. whether the performers appeared nervous or were appropriately dressed). To varying degrees these features were considered to be important to the hypothetical concert enjoyment of the 264 respondents of mixed musical experience and the group of expert musicians that conceptualised the 22 test items. Considering the close relationship between aesthetic and evaluative judgements discussed in Study 3, it is worth considering how these features would predict the quality ratings in a live setting, particularly those relating to the physical environment and to performance features explored elsewhere in this thesis.

6.1.3 Aims of the present study

The present study aimed to examine the nature of the evaluative environment in terms of the physical venue and perceptions of concertgoers' responses. As enjoyment scores of the first piece differed significantly from the overall half in Study 3, it was determined that listeners were able to remember and separate qualities of the performance of an individual work to general feelings toward the performance as a whole. Thus, it was determined that an overall quality rating would be examined in addition to quality rating of the first piece to allow for overall predictors of quality to be measured separately from quality of the first work.

The first two research questions (RQs) built upon the fourth overarching research question of the thesis, which examined aspects of the social and physical environment as discussed above as they related to these performance quality ratings. They were:

RQ1. To what degree do perceptions and features regarding the physical environment (i.e. acoustic, seat location, and venue appropriateness) relate to quality rating of a live performance?

RQ2. Do concertgoers' performance quality ratings differ from their perceptions of how their fellow audience members rated the concert in a live performance, and is this mediated by musical experience?

While the environment was the principal focus, the study also examined the relationship between select variables of concert enjoyment posited by Thompson (2007). As these had only been examined with reference to enjoyment of a hypothetical concert, this study aimed to provide the first examination of these features in relation to the evaluative judgement of a live performance, with the features regarding the physical venue examined above included. Thus, a third research question was:

RQ3. To what degree do extraneous features of a performance (adapted from Thompson, 2007) predict evaluative ratings of a live concert?

Finally, application of the fundamental study design used in Study 3 allowed for a replication and expansion of several of the key findings. Thus, three hypotheses were posited:

H1. Self-reported mood state, but not physiological state, after the performance (but not before the performance or changing across the performance) would be predictive of quality ratings of both the first piece and of the cumulative first half of the performance.

H2. Self-reported mood state after the performance and change in perceived physiological state across the performance would be predictive of enjoyment.

H3. Regarding the first work in the programme, likeability of the composition would predict quality rating, but familiarity would not.

To investigate these questions and test the hypotheses, the study design employed in Study 3 was adapted, using the same performers (the Eric Whitacre Singers), repertoire (choral music), venue type (large sacred building) and methodology. It differed in a new audience sample in a different city, and the custom nature of the survey (described below). The pre-post survey design was again used to allow for certain extraneous features, such as anticipation of the concert and quality of the seat, to be assessed before the concert experience and evaluative process began to avoid influence of the performance itself or task of assessing it. It also allowed for a pseudo-replication (as some changes were made to the survey items; see 6.2.2 below) of the affective and arousal state investigations of Study 3.

6.2 METHOD

The procedure of the present study took largely the same form as that in Study 3. Again, the Eric Whitacre Singers performed a programme of a cappella and accompanied choral works with a harmonically standard tonal structure. The concert was held at Gloucester Cathedral (see Section 6.2.3 for further description of the venue)

6.2.1 Participants

As in Study 3, surveys with three or more pieces of missing data or with reported age below 18 were rejected, leaving a total of 433 complete datasets, representing an estimated 61% of the full audience in attendance. This comprised 159 men and 264 women (10 not reporting) with a mean age of 53.17 years (SD ± 15.59, range = 18 - 82). 260 (60.0%) participants reported having experience playing a musical instrument or singing (of whom 145 reported singing), among whom a mean 30.19 years of musical experience was reported (SD ± 18.95, range = 1 - 73). Overall, participants reported attending a mean 7.33 “concerts like this one” per year (SD ± 10.83, range = 0 - 100). The front page of the questionnaire explained that participation was voluntary and that, by completing the questionnaire, they were providing informed consent to take part in the research. Ethical approval was granted by the Conservatoires UK Research Ethics Committee and conducted according to the ethical guidelines of the British Psychological Society.

6.2.2 Materials

A similar survey design approach was taken to that employed in Study 3, although adapted to capture the audience mood states, traits, opinions on the concert venue, and perceptions of the performance necessary to examine the research questions and hypotheses outlined in Section 6.1.3 (see Appendix 8 for the full survey). Again, the survey was designed to be as short as possible to maximise participant response and minimise disruption of the concert event. Printed on both sides of a single sheet of A5 paper, the first side again carried instructions to be completed before the concert began, with the second to be completed at the start of the interval. In addition to the demographic information collected in Study 3, participants reported where they were seated (by section and row; see Section 6.2.3 below).

The pre-concert side asked audience members to respond to seven items, each on the same 10-point scale from 1 (*not true at all*) to 10 (*very true*). The first two items examined affective response relating to arousal and affect. Rather than the full battery of mood questions, participants responded to a reduced set of just two items: whether they were in a “good mood” (affect) and whether they felt “relaxed and rested” (arousal). This language was adapted from Thompson (2007) and conformed to the arousal/affect distinction identified in the factor loadings in Study 3 while allowing for the test of mood state relations with judgement posited in hypotheses 1 and 2 of this research. Although Study 3 found that affective state loaded onto positive and negative factors (see Section 5.3.2), examination of the descriptive values in Section 5.3.1 showed that the four component negative affect items (i.e. *sad*, *afraid*, *confused*, *angry*) all showed very low scores with relatively little variance across the sample and no significant change across the concert. Furthermore, negative affect did not significantly contribute to the regression model predicting the performance quality ratings in Table 5.3. Thus, a single, positively-framed affective mood state question was considered suitable for the present study, fitting the language established by Thompson (2007). Five items were then selected from Thompson (2007) that could be answered prior to the concert’s start, free of influence from the performance itself. Two were chosen to address research question 1 relating to perception of the physical environment (*I am in a good seat; this is a good venue for this concert*), and the remaining items, addressing research question 3, examined general familiarity (*I am familiar with Eric Whitacre's music*), anticipation (*I have been looking forward to this performance*), and likeability (*I like this kind of concert*).

The second side, to be completed at the interval, first included a repetition of the two mood items to allow for the testing of hypotheses 1 and 2. Again, five items were adapted from Thompson (2007), this time including those where responses required experience of the performance. To examine the physical environment in research question 1, the first item related to the acoustic (*the acoustics were good*). Also chosen were items questioning what the audience perceived of the performance itself (*there were wrong notes, the performers appeared anxious*), relating specifically to the issues of performance errors and perceptions of the performers' affective state (i.e. negative facial expressions) examined in Studies 1 and 2 of this thesis. One item questioned perceived features of the physical and social environment (*there were unwelcome distractions* [*e.g. coughing, traffic noise*]) and finally one item examined engagement (*I was absorbed by the performance*).

The remainder of the second, post-performance side of the survey addressed quality ratings, using 10-point scales from 1 (*low*) to 10 (*high*). First, they rated their perceived quality and enjoyment of the full performance thus far. To examine research question 2 and differences in perceived ratings of those around them, this section also asked participants to guess how their immediate neighbours and the audience as a whole judged the performance quality. Finally, allowing for the replication test posited in hypothesis 3, participants were asked to consider quality, enjoyment, familiarity, and likeability of the first work in the programme.

6.2.3 Venue

The venue for the performance was Gloucester Cathedral (Gloucester, UK), the nature of which requires some description due to the focus of this study on the physical environment. The venue was typical of an English cathedral, with high, stone ceilings, walls, and columns surrounding the audience and performance stage (see Figure 6.1). Cathedral acoustics have been found to vary considerably depending on the relative location of the sound source and receiver (Álvarez-Morales et al., 2014), and there is no reason to suppose that such variability would not be exhibited in this setting. Where seating in Study 3 was unassigned, in this instance audiences were seated by ticket and corresponding row in the Nave (the central area; see Figure 6.1) while seating in the flanking Aisles was unassigned and a handful of audience members took seats in the Quire behind the stage (only 2 of the participants reported sitting in this area). The aisles were located on the far side of large support columns and by their nature restricted the view and (in tandem with the lower ceiling to the sides) could be presumed to have acoustic variations. Taken with the great length of the cathedral, thus the large variability in seat location by row, it was assumed that a great deal of variability in the perceived nature of the seating based on location could be expected.

6.2.4 Procedure

Surveys and pencils were placed on each seat prior to the concert. Ten minutes prior to the performance, the conductor and a researcher took the stage to explain the research study and a coinciding project (i.e. Fancourt & Williamon, 2016; the present sample did not participate in that project). Following the first half, they were immediately reminded by the conductor to complete the second side of the survey. The surveys were then collected over the interval and at the end of the concert. A team of researchers wearing identifiable clothing was on hand to collect the surveys and answer any questions of the participants. Figure 6.2 depicts a flow diagram of the research procedure and the points in time at which each set of data were collected.

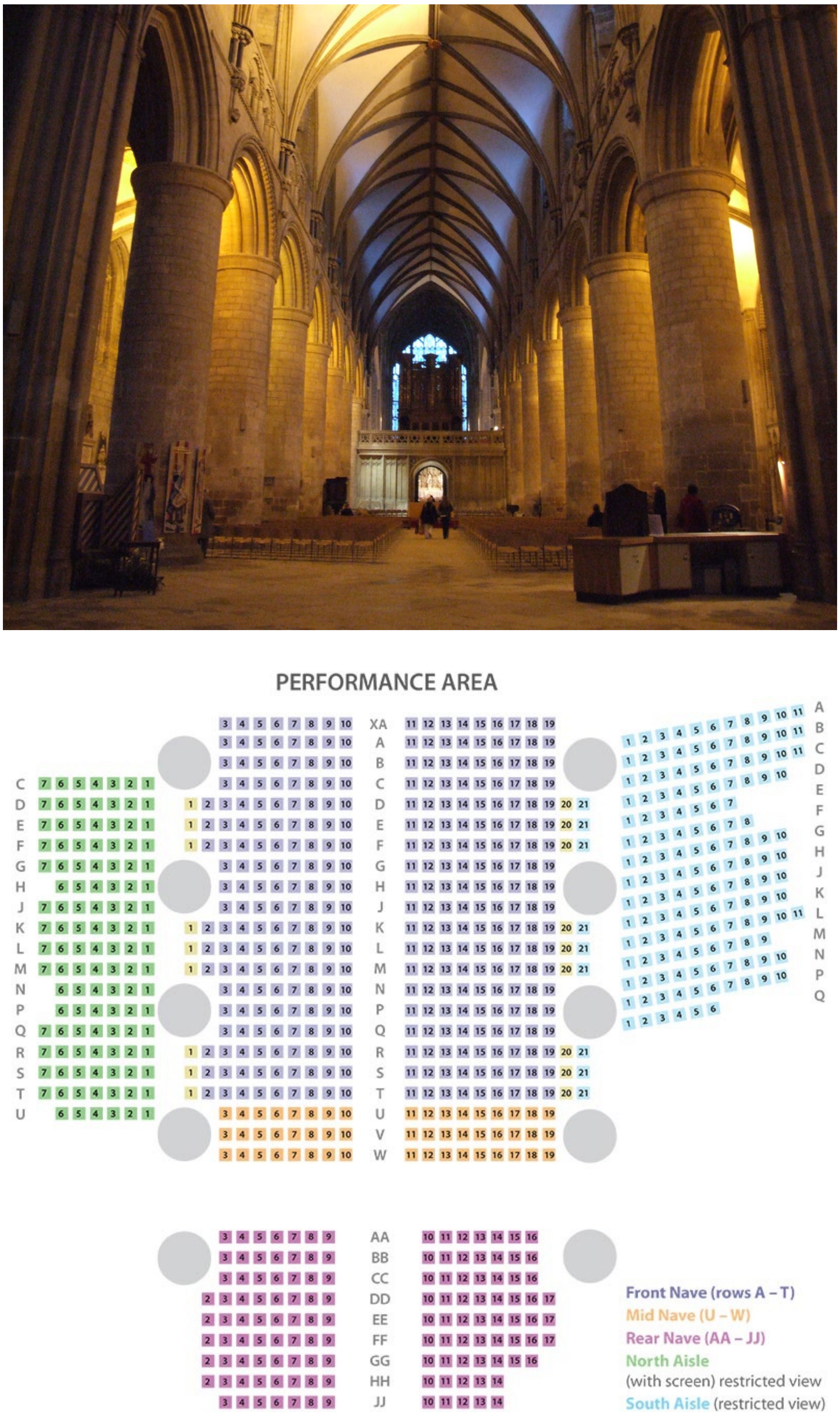


Figure 6.1. Gloucester Cathedral and its seating plan for the concert. The image was taken in the space between rows W and AA. Top source: Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Gloucester_cathedral_interior_001.JPG Below source: Gloucester Choral Society, http://gloucesterchoral.com/booking-tickets/gloucester-cathedral.

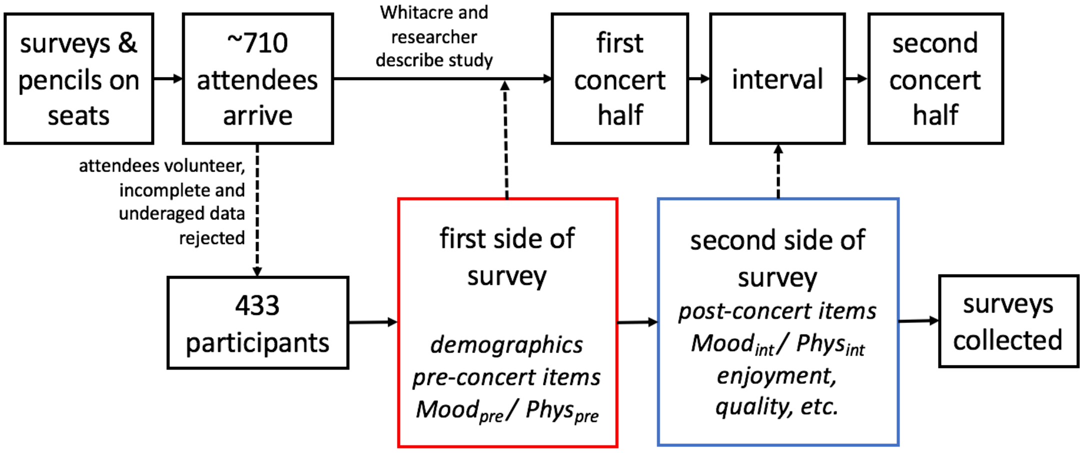


Figure 6.2. Flow diagram of the Study 4 research design. 433 participants responded to the surveys. The first half of the items were completed before the concert began (red) and the other half were completed at the concert interval (blue). See Appendix 8 for the complete survey.

6.2.5 Data treatment and analyses

Questionnaires with 1-2 missing data points were allowed, thus n-values varied slightly between tests (each of which were conducted excluding cases listwise) and are reported where appropriate. To examine the first and third research questions regarding predictors of performance quality and aesthetic judgements, correlations and multiple regression were used with the corresponding survey items. For research question 2, repeated-measures ANOVA was used to examine differences in own, neighbour, and audience quality ratings. As hypotheses 1-3 comprised replication tests of the models established in Study 3, the same multiple regression approaches were conducted. To accomplish this, change scores were calculated for the "good mood" (affective) and "relaxed and rested" (self-reported physiological) items as they differed between the *pre-concert* and *interval* measurements, resulting in six predictor variables (Mood_{pre}, Mood_{int}, Mood_{Δ}, Phys_{pre}, Phys_{int}, Phys_{Δ}) with quality and enjoyment ratings of the first work and of the first half as the dependant variables across four regression analyses. Considerations of sample size and normality in Study 3 were similar, in that variations in normality were present. In conjunction with the large sample size (> 400), parametric tests were used unless noted otherwise where the highly negatively skewed nature of and number of tied ranks within the quality and enjoyment ratings called for non-parametric alternatives, again with appropriate caution given to significance and generalisability.

6.3 RESULTS

As in Study 3, preliminary analyses found no significant relation (and correlation values of $\tau < .2$) between quality rating and participants' age, sex, musical experience, and concerts attended per year, thus these demographic factors were excluded from further analyses. The mean quality ratings of the first work and the first half of the concert, again as in Study 3, were high (first work: M = 9.18, SD ± 1.22, range = 2 - 10; overall: M = 9.26, SD ± 0.98, range = 4 - 10), thus were negatively skewed and demonstrated a ceiling effect. The same was found for the two respective *enjoyment* ratings (see Table 6.1). Repeated-samples Wilcoxon signed rank tests were conducted to compare both quality and enjoyment ratings of the performance of the first work with their respective overall scores, with no significant differences found. However, Kendall's tau values across the same comparisons tests showed only medium significant correlations (quality: $\tau = .57$, $p < .001$; enjoyment: $\tau = .57$, $p < .52$). The reason for this can be seen by examining the relatively symmetrical degree to which the sample changed their quality scores, where 263 (61.0%) made no change, 81 (18.7%) rated the first work higher than the overall, and 88 (20.3%) reported the first work lower than the overall. Thus, these quality scores were taken as generally comparable in that the majority of the audience did not feel the first piece had been performed with any difference in quality worth reporting, with some degree of non-systematic inter-rater variability among those who reported differently. Therefore, quality of the overall concert was used as the primary dependant variable in the analyses that follow unless indicated otherwise.

Table 6.1. Descriptive responses to the Study 4 survey. Pre- and post-concert questions were rated from 1 (*not true at all*) to 10 (*very true*); questions on quality and enjoyment were rated from 1 (*low*) to 10 (*high*).

Item	*M*	*SD*	*n*
Pre-concert			
I am in a good mood	7.69	1.86	432
I feel relaxed and rested	6.66	2.16	432
I am familiar with Eric Whitacre's music	4.84	3.25	433
I have been looking forward to this performance	7.92	2.05	430
I like this kind of concert	7.82	1.92	420
I am in a good seat	7.04	2.50	431
This is a good venue for this concert	8.71	1.47	422
Interval			
I am in a good mood	8.60	1.46	429
I feel relaxed and rested	8.30	1.67	428
The acoustics were good	9.17	1.37	429
There were unwelcome distractions	2.80	2.32	432
There were wrong notes	1.51	1.52	424
The performers appeared anxious	1.57	1.46	429
I was absorbed by the performance	8.58	1.73	429
Quality/enjoyment overall			
The quality of the performance of the first half	9.26	0.98	432
How I think those sitting next to me would rate the performance's quality	8.88	1.18	417
How I think the general audience would rate the performance's quality	8.94	0.97	426
My enjoyment of the performance of the first half	8.98	1.27	432
Quality/enjoyment first piece			
The quality of the performance of the first piece	9.18	1.22	432
My enjoyment of the performance of the first piece	8.89	1.61	432
My familiarity with the first piece	2.18	2.48	432
How much I like the first piece	8.52	1.99	431

6.3.1 RQ1: Relationship between the physical environment and ratings

The first research question related to those elements in the survey related to the physical environment in which the performance took place and their relation to the quality rating of the full performance. This included the *location*: i.e. the central Nave (n = 333), flanking Aisles (n = 83), or Quire behind the stage (n = 2; these participants were excluded from this analysis to provide two variables in the nominal variable, thus allowing it to be included in the correlational analyses). Participants were also asked to report their *row* (only n = 324 of N = 433 reporting: 245 from the Nave, 66 from the Aisles, and 13 without *location* reported) where a higher number indicated a relatively greater distance from the stage, and presumably a less desirable location. They also self-reported *seat quality*, the appropriateness of the *venue* for the concert, and the quality of the *acoustic.* Correlations (Kendall's tau) of each of these items with quality rating, along with enjoyment rating for comparison, are reported in Table 6.2. To account for multiple comparisons between the variables contributing to relationships relating to the physical environment and the quality and enjoyment ratings, only p values below .002 were considered significant following a Bonferroni correction.

Results showed that the location of one's seat, whether in terms of general area (Nave versus Aisle) or row number (distance from stage) showed no correlation, whatsoever with quality or enjoyment ratings. This was not necessarily due to lack of seat preference, as a significant and medium negative correlation was shown between *row* and perceived *seat quality* (τ = .48) wherein rows closer to the stage were perceived to be better. However, those sitting in the Nave did not feel their location was any better or worse than those in the Aisles. While perceived *seat quality* showed significant but negligible (τ < .2) correlations with *quality* and *enjoyment* ratings, ratings of appropriateness of the *venue* and quality of the *acoustic* showed significant

Table 6.2. Correlations (Kendall's tau) between quality and enjoyment ratings and indicators of the physical environment. N values are given in the bottom left**.**

	Quality	*Enjoy*	*Loc*	*Row*	*Seat qual*	*Venue*	*Acoustic*
Quality	-	**.68****	.02	-.01	**.13****	**.32****	**.41****
Enjoy	*432*	-	-.02	-.02	**.14****	**.29****	**.41****
Location	*415*	*415*	-	.04	-.03	.02	.08
Row	*323*	*323*	*311*	-	**-.48****	-.10	-.11
Seat quality	*430*	*430*	*414*	*322*	-	**.32****	**.18****
Venue	*421*	*421*	*405*	*316*	*422*	-	**.34****
Acoustic	*428*	*428*	*412*	*322*	*427*	*418*	-

** p < .001

small and medium positive correlations, respectively. However, some degree of multicollinearity should be noted, as the three items of *seat quality*, *venue,* and *acoustic* also showed significant small correlations between them.

To determine the accumulated predictive power of these items on the quality rating, a multiple regression analysis was performed with quality rating as the dependant variable and *location*, *row*, *seat quality*, *venue*, and *acoustic* as predictor variables. The correlation matrix revealed a degree of multicollinearity in line with the results of table 6.2, although none were high (rs < .50). The analyses produced a significant model ($F_{(5,294)} = 12.76$, $p < .001$, $R^2 = .16$) accounting for 16% of variance in the quality rating (see Table 6.3). Of the independent variables, the only significant predictors proved to be the perceived appropriateness of the venue reported before the concert began (b = .18, *ß* = .30, p < .001) and quality of the acoustic reported after the performance (b = .14, *ß* = .21, p < .001).

6.3.2 RQ2: Perceptions of fellow concertgoers' quality ratings

Research question two examined an aspect of the social evaluative environment, and how concertgoers perceived their own ratings to differ from those sitting immediately next to them and the audience as a whole. It was thought that musical experience may affect the degree to which participants felt their judgements differed from their peers, thus a 2x3 mixed ANOVA was used in which musical experience (i.e. whether or not the participant reported playing a musical instrument) was entered as a between-groups independent variable and the three quality ratings

Table 6.3. Regression model predicting performance quality rating of the first half of the concert based on environmental variables. Only appropriateness of the venue and quality of the acoustic proved significant (highlighted in bold).

Item	*b*	*SE B*	*ß*	*p*
Constant	6.36	.51		< .001
Location	.05	.12	.02	.951
Row	.00	.01	-.04	.663
Seat quality	-.01	.03	-.02	.798
Venue	**.18**	**.04**	**.30**	**< .001**
Acoustic	**.14**	**.04**	**.21**	**< .001**

(*own*, *neighbour*, and *audience*) entered as the within-subjects factor. Mauchly's W indicated a violation of sphericity (p < .05), thus Greenhouse-Geisser corrections were used. The ANOVA revealed a significant and medium main effect of rating type ($F_{(1.95,782.74)}$ = 41.30, p < .001, η_2 = .09; see Figure 6.3), with post-hoc Bonferroni tests demonstrating a that participants' own ratings (M = 9.28, SD ± 0.99) were significantly higher than their perceptions of their neighbours' (M = 8.89, SD ± 1.17; p < .001) and the full audiences' (M = 8.93, SD ± 0.98; p < .001) ratings. No significant difference was found between the two hypothetical ratings. Also, no significant main effect of or interaction with musical experience was shown.

6.3.3 RQ3: Relationship between performance features and quality rating

The third research question concerned the relationship between the combined extraneous variables relating to the concert and the overall quality rating. A

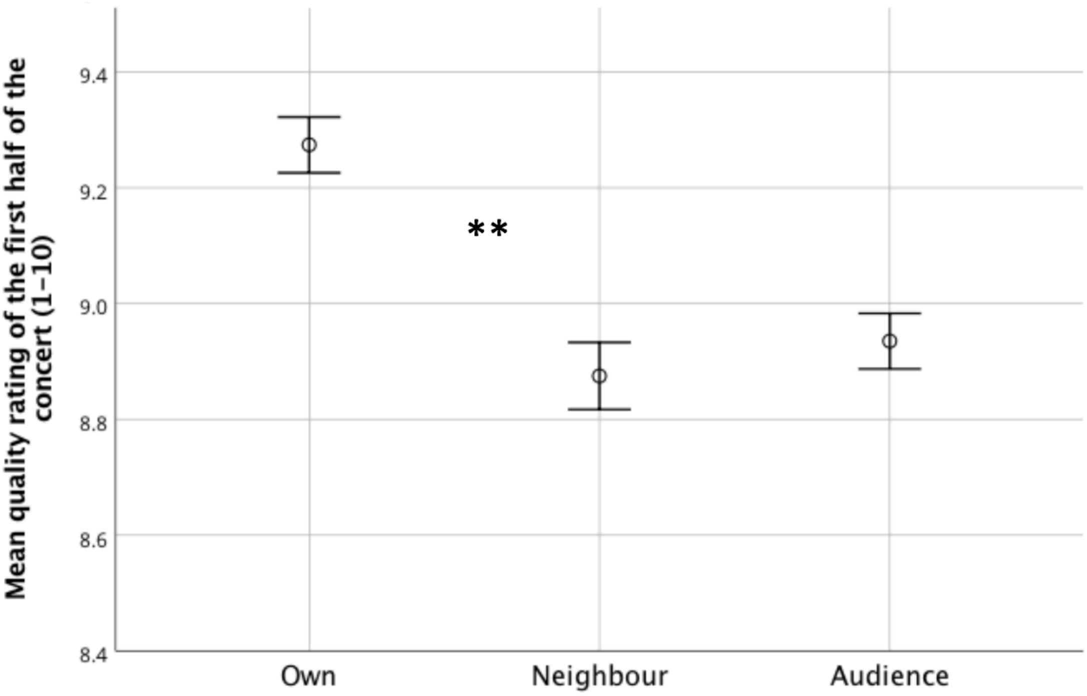


Figure 6.3. Differences in *own* quality rating, assumed quality rating of the immediate *neighbour*, and assumed quality rating of the entire *audience*. ** = p < .001, as measured using a repeated-measures ANOVA and post-hoc Bonferroni tests. Error bars show +/- 1 SE.

hierarchical multiple regression model was used to examine this. As those items relating to the physical environment (*location*, *row*, *seat quality*, *venue*, and *acoustic*) have already been examined, finding only appropriateness of the venue and quality of the acoustic to be predictive, just these two items were included in the first step of the model. The second step entered the remaining items, both those recorded before the concert began (*familiarity* with Eric Whitacre's music, *anticipation* of the performance, and the degree to which they *like* this kind of concert) and those recorded after (whether participants perceived unwelcome *distractions*, wrong *notes*, or that the performers were *anxious*, and whether they were *absorbed* in the performance). Finally, the third step of the model incorporated the four items of affective state prior to and at the interval of the performance, both in terms of mood ($Mood_{pre}$, $Mood_{int}$) and whether they were relaxed ($Relaxed_{pre}$, $Relaxed_{int}$). This formed a partial replication of Study 3 in the examination of whether affective state prior to or following a performance was more predictive of quality rating, and how this may interact with other extraneous performance features and perceptions. The correlation matrix revealed a degree of multicollinearity, although none were high (rs < .70).

The first step of the analysis, including *venue* and *acoustic*, produced a significant model ($F_{(2,387)} = 46.21$, $p < .001$, $R^2 = .19$) accounting for 19% of variance in the quality rating (see Table 6.4). Both entered variables were significant at this step. The second step of the model, with the majority of the perceived variables, produced a significant increase of variance explained by the model ($F_{(7,380)} = 26.69$, $p < .001$, $\Delta R^2 = .27$). This accounted for a further 27% of the variance, where the degree to which participants liked this type of concert and were absorbed in the performance both contributed significantly. Appropriateness of the concert venue maintained its significance at this step, although perceived quality of the acoustic fell out as a significant predictor. Finally, the third and final step, comprising the pre- and post-performance affect states, accounted for a small but significant 2% increase in the model ($F_{(4,376)} = 3.72$, $p < .01$, $\Delta R^2 = .02$; total $R^2 = .48$), where only affective (but not arousal) state following (but not prior to) the performance contributed as a significant predictor, supporting the findings of Study 3.

Table 6.4. 3-step hierarchical regression model predicting perceived performance quality rating of the first half of the concert.

Item	*b*	*SE B*	*ß*	*p*
Step 1				
Constant	5.96	0.35		< .001
Venue	**0.14**	**0.03**	**0.21**	**< .001**
Acoustic	**0.23**	**0.04**	**0.31**	**< .001**
Step 2				
Constant	5.20	0.31		< .001
Venue	**0.07**	**0.03**	**0.10**	**.025**
Acoustic	0.06	0.03	0.08	.088
Familiar	0.00	0.01	-0.01	.911
Anticipation	-0.02	0.03	-0.04	.545
Like	**0.09**	**0.03**	**0.17**	**.003**
Distraction	-0.01	0.02	-0.02	.571
Notes	-0.05	0.03	-0.07	.081
Anxious	-0.04	0.03	-0.06	.181
Absorbed	**0.30**	**0.03**	**0.51**	**< .001**
Step 3				
Constant	**4.93**	**0.32**		**< .001**
Venue	**0.06**	**0.03**	**0.09**	**.047**
Acoustic	0.04	0.03	0.05	.232
Familiar	0.01	0.01	0.02	.721
Anticipation	-0.04	0.03	-0.08	.192
Like	**0.08**	**0.03**	**0.16**	**.006**
Distraction	-0.01	0.02	-0.03	.451
Notes	-0.04	0.03	-0.07	.097
Anxious	-0.03	0.03	-0.05	.251
Absorbed	**0.27**	**0.03**	**0.46**	**< .001**
Mood_{pre}	0.02	0.04	0.03	.649
$\text{Relaxed}_{\text{pre}}$	-0.05	0.03	-0.10	.116
$\textbf{Mood}_{\textbf{int}}$	**0.13**	**0.05**	**0.19**	**.016**
$\text{Relaxed}_{\text{int}}$	0.00	0.04	0.00	.968

6.3.4 Hypothesis 1: Affective state post-performance as a predictor of quality judgement

The previous section has supported a key finding of Study 3 in that mood state following the performance significantly predicted performance quality ratings in a multiple regression model where mood state prior to the performance and arousal state at either point did not, even in combination with other features of the performance. To further strengthen this finding, the data collected in the present study allowed for a closer replication of that model by including two new variables representing the degree to which the *mood* and *relaxed* items changed over the performance (*Mood*$_{\Delta}$ and *Relaxed*$_{\Delta}$). Furthermore, this could be tested twice with both the quality rating of the first work and of the whole performance as dependant variables. Two multiple regressions were run using these features (see Tables 6.5 and 6.6). The correlation matrices revealed a degree of multicollinearity in both, although none were high (rs < .70). Both provided significant models ($F_{(4,419)} = 20.29$, $p < .001$, $R^2 = .15$; $F_{(4,419)} = 37.58$, $p < .001$, $R^2 = .26$, respectively), with the affective states accounting for 15% and 26% of first-work and full-performance quality ratings, respectively. The contributions of the predictors in both models supported the hypothesis, thus replicating the findings of Study 3 (see Section 5.3.3), with pre-concert items being

Table 6.5. Regression model predicting performance quality rating of the first work.

Item	*b*	*SE B*	*ß*	*p*
Constant	6.44	0.33		< .001
Mood$_{pre}$	-	-	-	-
Relaxed$_{pre}$	-	-	-	-
Mood$_{int}$	**0.38**	**0.06**	**0.45**	**< .001**
Relaxed$_{int}$	-0.08	0.06	-0.11	.139
Mood$_{\Delta}$	0.01	0.05	0.01	.815
Relaxed$_{\Delta}$	0.08	0.04	0.12	.067

Table 6.6. Regression model predicting performance quality rating of the first half of the concert.

Item	*b*	*SE B*	*ß*	*p*
Constant	6.34	0.25		< .001
Mood$_{pre}$	-	-	-	-
Relaxed$_{pre}$	-	-	-	-
Mood$_{int}$	**0.31**	**0.05**	**0.46**	**< .001**
Relaxed$_{int}$	0.02	0.04	0.04	.595
Mood$_{\Delta}$	0.04	0.04	0.06	.338
Relaxed$_{\Delta}$	0.03	0.03	0.05	.401

dropped from the model due to lack of contribution, and only mood following the performance providing a significant contribution.

6.3.5 Hypothesis 2: Affective states as predictors of enjoyment

The data collected in the present study also allowed for a replication of the finding in Study 3 that change in arousal state as well as the post-mood state also contributed to the *enjoyment* rating of the performance (see Section 5.3.6). The same independent variables were used, and again this could be tested twice with both the enjoyment rating of the first work and of the whole performance as dependant variables. Two multiple regressions were run using these features (see Tables 6.7 and 6.8). The correlation matrices revealed a degree of multicollinearity in each, although none were high (rs < .70). Both provided significant models ($F_{(4,419)} = 24.47$, p < .001, $R^2 = .18$; $F_{(4,419)} = 69.81$, p < .001, $R^2 = .39$, respectively), with the affective states accounting for 18% and 39% of first-work and full-performance enjoyment ratings,

Table 6.7. Regression model predicting enjoyment rating of the first work.

Item	*b*	*SE B*	*ß*	*p*
Constant	4.95	0.43		< .001
Mood_{pre}	-	-	-	-
$\text{Relaxed}_{\text{pre}}$	-	-	-	-
Mood_{int}	**0.53**	**0.08**	**0.47**	**< .001**
$\text{Relaxed}_{\text{int}}$	-0.09	0.07	-0.01	.210
Mood_{Δ}	0.01	0.07	0.01	.842
Relaxed_{Δ}	0.11	0.06	0.12	.053

Table 6.8. Regression model predicting enjoyment rating of the first half of the concert.

Item	*b*	*SE B*	*ß*	*p*
Constant	4.56	0.29		< .001
Mood_{pre}	-	-	-	-
$\text{Relaxed}_{\text{pre}}$	-	-	-	-
Mood_{int}	**0.45**	**0.06**	**0.51**	**< .001**
$\text{Relaxed}_{\text{int}}$	0.04	0.05	0.05	.418
Mood_{Δ}	**0.12**	**0.05**	**0.13**	**.011**
Relaxed_{Δ}	**0.08**	**0.04**	**0.11**	**.045**

respectively. The contributions of the predictors in both models did not fully align with or replicate the findings of Study 3, although similarities could be seen. In both, the pre-concert items were dropped from the model due to lack of contribution. In the first model examining the enjoyment rating of the first work, mood following the performance was significantly predictive but change in relaxation was not, although the latter's significance value of p = .053 is worth noting. In the second model examining the enjoyment rating of the full performance, not only were mood following and change of relaxation significant predictors (themselves a replication of Study 3), the degree to which mood changed across the performance was also significant.

6.3.6 Hypothesis 3: Relationships between perceived quality, enjoyment, familiarity, and likeability of the performance of the first work

Regarding the first piece, correlation analyses revealed a similar pattern to Study 3 in which quality, enjoyment, and likeability of the work correlated while familiarity did not (see Table 6.9). Here, the relationship was even more pronounced, with stronger correlations (τs = .7 - .8) found between the former and no significant relationships with familiarity.

6.4 DISCUSSION

This study investigated an audience's evaluative and enjoyment ratings of a performance in a live concert setting, and how these judgements related to self-reported perceptions of extraneous performance features contributing to the social and physical environment and the participants' affective state. As in Study 3, the sample represented a diverse range of ages and musical experience, with participants

Table 6.9. Correlations (Kendall's tau) between *interval* questionnaire items relating to the first piece.

	Quality	*Enjoyment*	*Familiarity*	*Likeability*
Quality	-	**.80****	.00	**.70****
Enjoyment	432	-	.03	**.80****
Familiarity	432	432	-	.03
Likeability	431	431	431	-

** p < .001

completing a custom questionnaire immediately before and after a professional choral concert they had chosen to attend. Results of the data collected are discussed with respect to the three research questions and three hypotheses posited in Section 6.1.3.

6.4.1 Research question 1: Seat location and acoustics

The first research question examined the interaction between factors relating to audience members' perceptions of and location within the physical environment, and their evaluative ratings of the performance. Participants provided their location and seat number and rated the quality of their seat, the appropriateness of the venue, and quality of the acoustic. Only appropriateness of the venue and quality of the acoustic contributed significantly to the multiple regression model. Of course, causality of this relationship cannot be assumed. Participants may well have experienced better acoustics in certain parts of the cathedral and, as a result, heard the performance differently than their peers and provided a commensurate performance quality rating. Studies in psychoacoustics have demonstrated that listeners, while exhibiting individual variability in their preference for and language used to describe acoustic features (Ueno & Tachibana, 2005), are able to make relatively consistent assessments of acoustic quality and preference (Galiana et al., 2012). Alternatively, listeners' judgements of acoustic and quality may have been linked by differences in baseline rating outcome – i.e. someone more likely to think highly of performance quality will also think highly of the acoustic aspect. Three findings support this view: (1) acoustic ratings saw a near identical significant correlation with both quality and enjoyment ratings (τs = .41), (2) seat location saw a negligible correlation with acoustic quality, and (3) acoustic quality was dropped as a significant predictor in the hierarchical regression model examined as part of research question 3. Judgement of venue appropriateness may have interacted in a similar way. Thus, further research is required in which acoustic is experimentally manipulated while listening to the same stimulus to determine whether a true effect exists. Until this point, the results presented here should be interpreted with due caution.

The lack of significant correlation between seat location and quality rating was surprising, especially considering the virtually null correlation values (τs < .02) and beta values in the multiple regression (bs = < 0.01). This is not to say that the seats themselves were all of equal perceived value, with a significant medium correlation seen between row number (i.e. distance from the stage) and the perceived quality of the seat. Worth noting is that the standard pricing model was in effect in which seats closer to the stage were priced higher than those to the rear or in the unreserved aisles. Perhaps, then, the perceived value of the more affordable seats was counteracting any loss of enjoyment stemming from the less optimal distance from the stage. If this result is replicable and generalisable, those who design concert spaces and market to audiences may be interested to know that their patrons may still be able to enjoy the performance fully, and appreciate the performer's ability, from the back row.

6.4.2 Research question 2: Assumptions of peer judgement

The second research question examined audience members' perceptions of how others rated the performance, revealing that they believed that their own ratings were significantly higher than those of their peers by approximately one third of a point – a relatively large amount considering the standard deviation of less than 1. Two possible explanations, although not mutually exclusive, immediately present themselves. Individuals may have tended to assume that they recognised greater ability in the performers than their peers. Alternatively, they may have assumed that others would be more critical and discerning in their judgements. It was thought that this trend may be moderated by participants' musical experience, perhaps assuming heightened musical ability could make a judge consider themselves more discerning, although no evidence for this was found. Nonetheless, it would not have been surprising to find that audience members assumed their perceptions of the performers' ability matched that of their peers. Further research should examine the underlying motivations of this effect, and whether they are driven by external stimuli (e.g. applause) or internal assumptions.

6.4.3 Research question 3: Predictors of performance quality

The third research question examined whether extraneous features of the performance, adapted from Thompson (2007), were predictive of performance quality. The step-wise regression model supported earlier findings (Thompson, 2006; Study 3 of this thesis) in that likeability of, but not familiarity with, the musical material was predictive of quality ratings. It was also found that absorption in the performance was predictive and demonstrated that factors found to show significant relationships with quality rating – appropriateness of the venue and mood state following the performance – were robust to inclusion in a more complex model. Thus, the hypotheses posited in the previous study (see Section 5.4.1) relating to the role of mood state in judging performance quality can be revisited considering the replication: that of a mood-based recency effect in which the mood at the time of forming the judgement, rather than the mood experienced at the point of hearing the piece or stimulated due to hearing it, was most predictive. That perceived performance anxiety, wrong notes, and distractions were not predictive of quality ratings in this case was perhaps unsurprising. All three received very low descriptive ratings (less than 2, 2, and 3 respectively on scales from 1 to 10), as the performance was of very high standard with no obvious performance errors or audience-wide distractors apparent to the researcher.

6.4.4 Hypotheses 1 - 3: Replications of Study 3

Hypotheses 1, 2, and 3 aimed to replicate the key findings of Study 3, all three of which were fully or partially supported. Hypothesis 1 was confirmed twice in that mood state following the performance showed a significant relationship with quality judgement in both quality ratings (first-piece and full performance), while initial mood state, change between, and all states relating to arousal were not. Taken with the similar effect found in the hierarchical regression model discussed in the previous section, a strong case is presented for the robustness of this effect within the common features and population of the performances and environments in question. Hypothesis 2 was supported in that similar models predicting enjoyment ratings saw inclusion of the change in arousal state as a significant predictor in one of the two models (with both hovering close to p = .05) and the addition of change in mood state appearing as a significant predictor in the second. Suffice to say, the role of mood and arousal state was found to be more complex and variable in predicting enjoyment than quality ratings. Finally, hypothesis 3 sought to replicate the finding that likeability, but not familiarity, correlated with enjoyment and quality ratings. This hypothesis was confirmed, with even stronger correlations found between likeability and the two judgements and familiarity showing a true random correlation of 0 with regards to the first piece on either rating.

6.4.5 Directions for future research

This study highlighted the challenges of engaging with the complexity of the physical and social evaluative environment in a naturalistically rich setting. Thus, this research remains exploratory, with suggestions that appropriateness and acoustic features of the venue may play a role in quality evaluation, and that evaluators hold assumptions of the rating decisions of their peers. A great deal remains to be understood as to how an evaluator is affected by the place they are in and the people with whom they interact when conducting a performance evaluation. Researchers require new approaches to investigate this complexity. For this reason, the following chapter engages with this issue by describing a novel approach to the study, and subsequent training, of the act of performance evaluation.

As in Study 3, the nature of a naturalistic survey design has limited the degree to which assumptions of causality can be made across the relevant research questions, the details of which have been considered across the previous sections of this discussion. Generalisability is also limited by the nature of the performance studied, which was chosen specifically to match Study 3 in terms of its performance quality (high), instrumentation (choral), genre (contemporary, harmonically tonal), venue (large sacred spaces), audience size (600-800), and performers (the Eric Whitacre Singers). This was done to allow direct comparison between and replication of the findings, although further research will be necessary to examine the degree to which the models can be applied to other performance conditions.

6.5 SUMMARY

This chapter examined the context of the physical and social environment in the formation of music quality judgements with a sample of 433 audience members at a live choral concert. A custom survey examining perceptions of extraneous performance features, along with mood states adapted from the previous study, were collected in conjunction with evaluative and aesthetic ratings of the performance and assumed ratings of fellow audience members. Analyses found that perceptions of acoustic quality and venue appropriateness, but not physical location, related to quality ratings, although causality was indeterminate. Participants tended to assume that their peers gave significantly lower ratings of performance quality. The key findings of Chapter 5 relating to relations between affective and arousal states, quality and enjoyment ratings, and work likeability and familiarity were replicated.

7 SIMULATING EVALUATION

7.1 INTRODUCTION

The chapters presented thus far have highlighted the limitations of existing methodologies of performance evaluation research, relating primarily to issues of ecological validity in laboratory settings and of determining causality in live settings. While the use of continuous measures and employment of temporally-specific surveys has given greater insight into the processes that lead to evaluative products, the complex nature and innumerable factors comprising musical performances, and evaluations of them, remain difficult to capture fully in traditional methodological paradigms. As a result, full understanding of the evaluative process remains incomplete. The current chapter presents the theoretical basis for and development of a new methodological tool to study and train the act of music performance evaluation. It begins by reframing the literature presented in Chapters 1 and 2, which reviewed existing studies examining the process of forming music performance quality evaluations in terms of (1) the factors they examine and (2) the models of performance measurement used. Following this reframing, the present chapter identifies a methodological gap where a new approach could allow for enhanced study of the evaluative environment in a an experimentally-controlled setting. The chapter then considers the parallel domain of training evaluative skills in musicians, presenting evaluation as a form of performance to be taught and demonstrating a similar gap in opportunities for trainees to develop evaluative skills under the heightened environments of live assessment scenarios. The concepts of Immersive Virtual Environments (IVEs) and distributed simulation are outlined, highlighting their use in training and research other performance domains. Taking this model as a starting point, the Chapter presents the development of an *Evaluation Simulator* as a new tool to study and train performance evaluation. Potential applications of this tool in academic and pedagogical settings are then discussed.

7.2 REFRAMING THE EVALUATIVE RESEARCH

Chapters 1 and 2 reviewed the literature that has examined the process of forming music performance quality evaluations. Chapter 1 surveyed the findings in the context of the factors they examine, and whether they provided insight into the repertoire, performer, environment, or evaluator and their respective effects on the evaluative process. Chapter 2 considered the tools used to conduct the evaluation itself, contrasting the scales and rubrics used in post-hoc assessments with the continuous measures methodologies more commonly used in affective response that offer new ways of examining the evaluative process. The subsequent four chapters then examined features and used tools from across these categorisations, with particular focus on the repertoire and performer using continuous measures in Chapters 3 and 4 and on the environment and evaluator using post-hoc assessments in Chapters 5 and 6.

A particular feature of the survey studies in Chapters 5 and 6 was that, in order to examine the evaluative responses of audience members as they related to their affective states and reactions to the performance environment, they were conducted in a live performance environment. The limitations of this approach were discussed in the chapter and highlighted the conflict between maintaining ecological validity and collecting rigorous quantitative or qualitative data. When examining the specific nature of studying performance evaluation, these can be grouped into four general categories:

1. ***Artificial Evaluation/Artificial Setting/Artificial Stimulus (AAA)***: those studies that set up artificial evaluative situations in laboratory settings in which participants rate pre-recorded performance material that has been created for the purpose of the experiment. In these, aural and visual variables of the performance are carefully controlled to suit the nature of the study, often

comprising trials in which a particular feature of the performance is experimentally manipulated while the rest remain constant. Examples include the experimental methods used in Chapters 3 and 4, in which customised aural and visual stimuli were recorded and manipulated to vary by stage entrance and performance error. Much of the literature in this field has used this method (e.g. Williamon, 1999; Griffiths 2008, 2010; Elliott, 1995; VanWeelden, 2004; Wapnick et al., 1998, 2000; Ryan & Costa-Giomi, 2004; Kopiez et al., 2017), allowing for casual conclusions to be drawn about the effects of extra-musical factors.

2. ***Artificial Evaluation/Artificial Setting/True Stimulus (AAT)***: those studies that set up artificial evaluative situations in which participants rate recorded performances not initially intended to be studied in a laboratory setting. Tsay's (2013) study provides a good example, where she presented the audio and video recordings from genuine piano concerto competitions to participants to evaluate in a laboratory setting (see Chapter 4 for a full description). This method was also used in Ryan and colleagues' (2006) study of the effects of performer attractiveness and gender using footage from the Eleventh Van Cliburn International Piano Competition, and Platz and Kopiez's (2013) research on stage entrances using recorded video from the Joseph Joachim International Violin Competition.
3. ***Artificial Evaluation/True Setting/True Stimulus (ATT)***: Those studies that set up artificial evaluation situations in live performance environments. The survey methods used in the Chapters 5 and 6 provide examples of this, where concertgoers were asked to provide evaluations that they would not have conducted otherwise of a public performance they had previously planned and paid to attend. A limited range of studies have been conducted in live concert settings (e.g. Thompson, 2006). This category could also include 'mock' auditions in which a live performance is used as the stimulus for amateur or expert assessors to judge for the explicit purpose of the research or student

training (e.g. the student assessments in Bergee & Cecconi-Roberts, 2002; Daniel, 2004).

4. ***True Evaluation/True Setting/True Stimulus (TTT)***: those studies that analyse data taken from genuine evaluation situations, such as Flôres and Ginsburgh's (1996) examination of existing judges' data from the Queen Elisabeth Music Competition, or Davidson and Coimbra's (2001) observations of vocalists and examiners in conservatoire exam panels.

These four permutations of true/artificial evaluations, settings, and stimuli are the only ones logically possible, for if 'true' represents an act that would have occurred in a genuine evaluative environment, an artificial stimulus by definition does not have a corresponding true setting, and an artificial setting by definition could not host a true evaluation. However, the lines between the categories listed above can be blurred. A type 1 artificial stimulus could be manipulated in such a way that perceivably replicates a genuine performance, as was attempted in the Chapter 4 experiment, giving the participant the impression of a type 2 situation with the goal of eliciting a response in line with the more generalisable construct. By the same token, a situation in which a live performance and evaluation was organised for the sole purpose of an evaluation study, by definition type 3, could give the effect of a type 4 true evaluation if the evaluative environment were constructed in such a way that the judge (i.e. study participant) behaved as he or she would in a real-world setting, with the associated motivations, thoughts and behaviours. Experiments falling into type 1 maximise control of extraneous and experimentally manipulated variables, increasing replicability between individual participants. What is gained in control, though, is lost in generalisability to real world performance and evaluative situations. Experimenters may create stimuli that give the impression of genuine performances, or use stimuli taken from such performances as those studies falling into type 2, but this still leaves participants in isolated, laboratory-style conditions that may not reflect the concerts, competitions, or examinations in which evaluations take place.

Alternatively, researchers may move down the spectrum towards types 3 and 4, thus increasing ecological validity. This, however, sacrifices control over the situation. A human performer cannot perfectly replicate a live performance, reducing study replicability and minimising the opportunity to perform true experiments in which specific variables are controlled and causality can be determined. In some cases, naturalistic experiments may be conducted if the appropriate conditions are set and data are available, the latter of which can be a rare occurrence given the often confidential nature of assessors' judgements in institutional evaluative settings. The studies surrounding the Queen Elisabeth Competition (Flôres & Ginsburgh 1996; Glejser & Heyndels 2001) provide good examples of such a scenario, wherein researchers were able to analyse the published scores across decades of competitions to demonstrate how the sequence in which competitors performed correlated with their likelihood to win. Such research opportunities are rare, however, requiring a convenient conflux of a randomly assigned independent variable (in these cases, performance order in the semi-final round), decades of archived data from the judges, and an organisation willing to open those data (and by extension their own evaluative practices) to public scrutiny. Davidson and Coimbra's (2001) examination of academic assessment practices required even more access, where students, assessors, and administrators permitted the researchers to audio- and video-record the performances and evaluative discussions and alter the standard procedure of the mid-term recitals of a London conservatoire in order to obtain the qualitative and quantitative data the researchers required. While this provided unprecedented insight into the process by which examiners reach their decisions, the authors' observation of and intervention with the assessment undoubtedly affected the process by which the assessments were made (a caveat highlighted by the researchers). It also required a significant degree of cooperation with the institution, not only in allowing their own practices to be investigated but also in accepting the risk that the students' academic evaluations would be affected to an unknown degree.

These challenges of validity and control are by no means restricted to the study of assessment, or of music performance in general; they are fundamental to the social sciences. Music performance and its evaluation are simply domains that are particularly difficult to study in a naturalistic setting due to the complexity of the performative act being evaluated, the myriad social and psychological forces driving the behaviour of the evaluators and the interactions between panel members, performers, and institutions, and the potentially significant and sensitive consequences of the outcome. The question then remains as to whether new methodologies for capturing and examining the act of performance assessment are available. This chapter addresses precisely this question by considering evaluation as a skill to be trained and performed, and how this compares with the domain of general performance. It outlines how the tools of experiential learning and simulation have been used across performance domains, and specifically in music performance, to address the challenges of learning, improving, and researching performance. It then proposes the use of simulation to provide a new methodology in the training and research of the evaluative process and outlines the creation of an *Evaluation Simulator* at the Royal College of Music. Finally, potential uses and implications of such a simulator for skills training and research are discussed.

7.3 THE SKILL OF EVALUATION

Evaluation is surely a skill. Good evaluations can be defined, and good evaluators distinguished. At least, this is the assumption on which any formal assessment scheme incorporating an ‘expert’ assessor is based (Thompson & Williamon, 2003). However, the concept of the skilful, professional evaluator is not one to be taken for granted. Previous studies have questioned the value of an evaluator’s expertise in delivering reliable and consistent judgements (e.g. Fiske, 1975, 1977; Winter, 1993), as was seen in Chapter 4 wherein experienced musicians showed only minor deviation from non-musicians in their continuous evaluation processes and no significant differences in the nature of their final judgements. A review of 86 articles examining the abilities of music teachers in classroom or instrumental settings found a high degree of variability in the nature and effectiveness of their feedback, even within a single lesson (Duke, 1999). This is not to say that there does not exist an evaluative skill, or that such a skill is not valuable, but simply emphasises the point that one’s ability as a musical performer does not automatically translate to ability as an effective judge. Indeed, the profession of the instrumental music teacher (and, by extension, music examiner or competition judge) is populated primarily not by those with significant training in evaluation but rather by those who have demonstrated significant ability in the specialist area on which they are passing judgement, i.e. performance.

This is not due to lack of effort by those evaluators, or those who have assigned them. Rather, it indicates the lack of opportunities for this training, and the assumptions underlying what comprises an expert evaluator. The celebrated violinist Joseph Szigeti noted this in his autobiography, speaking of the challenges faced by the expert music judge and critic:

> This comparison of performances (whether of those by the same player spread over a given length of time, or of performances of the same works by about equally qualified players, massed within a short period) should be one of the self-imposed tasks of all conscientious critics. I don't quite know how they could manage it; perhaps by attending contests, examinations, and the like, taking a kind of post-graduate course in performance-criticism. As far as my own experience goes, my duties as member of the jury at the Paris Conservatoire contests and at the Brussels Concours International provided me with invaluable object lessons in the field of critical listening. On an active practitioner such lessons are wasted, of course, whereas for a critic…. (Szigeti, 1947, p 254, ellipses in original)

In this context Szigeti is referring to the "critic" in the sense of a critical reviewer, one publishing written reports and reviews of public performances. However, the translation can be made to the evaluator, as critics must also deconstruct the salient aspects of the performance (e.g. technique, artistic style, control, interpretation, etc), make comparisons across performances, and translate this to a form of feedback that provides a desired outcome for a particular audience/reader (Alessandri et al., 2014, 2015). With this in mind, Szigeti makes several salient points in the quotation. First, he addresses the challenge of making consistent and reliable comparisons between performances separated by time or between interpretations. The research literature has emphasised this difficulty, most notably in studies demonstrating how experienced listeners can often mistake the same performance played twice as two distinct interpretations (Duerksen, 1972; Anglada-Tort & Müllensiefen, 2017). Second, Szigeti struggles to identify a programme by which one could develop this skill, suggesting experience through exposure and a hypothetical course of advanced study, although seemingly unaware of whether such a programme or degree exists. Even if he is speaking of the specific skill of published performance criticism, a course on performance evaluation would seem to be a clear analogue. He confirms this view in his third point, where he highlights his role as jury member for a number of internationally prominent panels as his own lessons in criticism. Thus, he learned to assess by undertaking the assessment of others, in the process contributing to decisions having considerable ramifications for those assessed without any specific education in how to conduct them. He concludes by suggesting that such lessons are wasted on an "active practitioner" (meaning performer?), but have value for the critic.

This quotation by a prominent musician from the relatively recent history of the Western classical tradition highlights the degree to which the skill of evaluation has been given far less attention than the skill of performance. It suggests that those in positions of evaluative power are chosen not for their ability as judges, but for their prominence in a related domain. Such a view would be in line with the history of skill assessment. Centuries earlier, the apprenticeship model of developing skilled crafts once favoured social class in determining who held the power to assess and determine worth, a trend that shifted in 19th century Europe with the rise of competitive assessment, individualism, and a gradual (and unfinished) transition from a hierarchy based on class structure to one of meritocracy (Eggleston, 1991). It is notable, therefore, that the method of training modern musicians, at least those in the Western classical tradition, remains based largely upon the master-apprentice model (Gaunt, 2017). Conservatoires heavily favour the training of performance skills (Perkins, 2013), while the skill of performing effective evaluations receives far less attention. This despite the fact that the ability to diagnose and deliver useful feedback upon performance is central to the career of the modern portfolio musician, in which musicians are likely to have multiple roles as performer, assessor, and teacher (Bennett, 2008).

A few exceptions to this can be found. The Associated Board of the Royal Schools of Music (ABRSM), for instance, requires training, professional development, and monitoring for its 700 examiners through a three-day introductory course and subsequent four days of sessions that emphasise learning through the conducting of mock or true evaluations under the guidance of those more experienced evaluators (Stewart, 2011). Examiners are also periodically moderated, during which a second examiner remains in the room for the full session. Such practices have also been piloted and employed in higher education settings, examples of which are discussed below, although the practice is not widespread.

The practice and skill of evaluation delivery has been given greater attention, at least in terms of research and discussion, in the domain of classroom-based and higher-education teaching. In Chapter 1, four types of assessment were defined (see Section 1.2.1); (1) placement, in which performances are ranked or chosen; (2) summative, in which a performance evaluation is used to summarise ability or a period of learning; (3) diagnostic, used to pinpoint learning and technical deficiencies; and (4) formative, to determine whether development has taken place and to foster continued learning (Goolsby, 1999). Research and practice in evaluation in the wider educational context has focussed on the third and fourth categories in their role in enhancing student learning and development. Nicol and Macfarlane-Dick (2006) identified seven principles of good practice in the delivery of formative assessment. They encouraged feedback that:

1. helps clarify what good performance is (goals, criteria, expected standards);
2. facilitates the development of self-assessment (reflection) in their learning;
3. delivers high quality information to students about their learning;
4. encourages teacher and peer dialogue around learning;
5. encourages positive motivational beliefs and self-esteem;
6. provides opportunities to close the gap between current and desired performance;
7. provides information to teachers that can be used to help shape teaching.

These principles share close ties with those of self-regulated learning, which theorises that effective learning happens when learners deliberately plan, execute, and review their practice, working towards concrete goals while maintaining a metacognitive awareness that allows them to monitor and adapt their cycle of learning depending on their individual and subject-specific challenges (Zimmerman, 1990; Jørgensen, 2004, 2008). This can foster practice that is considered and deliberate, features critical to achieving peak performance outcomes (Ericsson et al., 1993). Paris and Winograd (1990) proposed that regular self-assessment of learning processes and outcomes promotes more effective monitoring of progress, facilitates the identification and correction of mistakes, and enhances feelings of self-efficacy, which is the belief in one's ability to perform domain-specific skills (Bandura, 1997; McCormick & McPherson, 2003; McPherson & McCormick, 2006; Ritchie & Williamon, 2011). and has been linked to improvements in practice (Ritchie & Williamon, 2012). Reciprocally, increased self-efficacy has been found to lead to higher self-evaluations, which themselves become increasingly underconfident as performance ability increases (Hewitt, 2015). In general, self-assessments are found to be higher than those of third-party experts (Hewitt, 2002, 2005). Such optimism in self-assessment has been linked to higher performance achievement and persistence in comparison with students displaying more realistic or pessimistic tendencies (Bonneville-Roussy et al., 2017a). Effective feedback, especially feedback that motivates and facilitates self-assessment, allows learners to close the cycle of self-regulated learning and enhance their performance practice most effectively. If this practice is performing the skill of assessment, then one must learn to self-assess one's ability to assess.

This ability to self-regulate feedback delivery forms a subset of what Medland (2015) defines as assessment literacy. In a study of external examiners in UK higher education she found that, while subject literacy was consistently high, their assessment literacy was being overlooked in their training, selection, practice, and in the research literature, with substantial variance across the examiners. This deficit was distributed over six subtopics: (1) *community*, or degree to which examiners had knowledge of and participated in groups sharing good practice; (2) *standards*, or the knowledge of and adherence to institutional and national policies; (3) *dialogue*, or the role and methods of engaging with students in their feedback and fostering peer-to-peer dialogue; (4) *self-regulation*, or the ability to demine and improve the quality of their own feedback; (5) *programme-wide approach*, or knowledge of and integration with the wider institutional and learning context for the material being taught and assessed, and (6) *knowledge and understanding*, or familiarity with the underlying pedagogical and psychological principles of effective assessment. Medland found a significant emphasis on *standards*, especially relating to the consistency, transparency, and appropriateness of the assessment policies in place. Such focus on procedure and policy invokes the danger of what Ferm Almqvist and colleagues (2016) defined as 'deformative' assessments, where over-assessed learning can promote a culture of criteria compliance rather than individualised self-regulated learning practices. Emphasising this, Medland found the category of *self-regulation* to be, by far, the least-mentioned component in her cohort. Responses relating to *dialogue* also highlighted an emphasis on one-directional feedback delivery or 'monologue' rather than constructive and formative interaction between instructor and student or, indeed, between external examiners, programme leaders, and lecturers. The importance of the methods of feedback delivery should not be overlooked. Not only do they provide new opportunities for formative learning, but the assessor's style and language can have a greater effect on the students' perceived value of the criticism and resulting self-confidence than the pedagogical content itself (Bonshor, 2017). It is here that the 'performance' of an effective evaluation is crucial.

7.4 EVALUATION AS PERFORMANCE

While performance evaluation can be conceptualised as a unique skill to be developed, there is value in considering it as an act of performance in itself. Like the musical performance it seeks to quantify, it calls upon specialist knowledge. It takes place in specific settings, often involving interaction with a team of familiar and/or unique experts that may or may not share a specific sub-specialism. It can take place in front of an audience (as in public competitions), one that can be critical of the outcome. The results of the act have consequences, not only for those being assessed, but for the evaluative performer in its effects on their reputation, standing, and employability as an evaluator. And, as has been the theme throughout this thesis, it is a process that unfolds in a fixed sequence over a fixed amount of time, often limiting or outright preventing opportunity for pause, repeat, or reflection, and including distinct periods of pre- and post-performance activities. To examine evaluation through the lens of performance allows us to consider its treatment anew. Evaluation is not just a tool to summarise, diagnose, and develop performance; it is an act whose quality and efficacy can itself be summarised, diagnosed, and developed through the same means.

Taking this view, the skills involved in executing a skilful evaluation now become a form of meta-assessment; how does one deliver formative assessment of a formative assessment? If considering evaluation as a performance, one can apply the seven principles of evaluation listed above (Nicol & Macfarlane-Dick, 2006) not just to the assessment of performance, but to the assessment of assessment itself. When reframed in this manner, good formative evaluation:

1. helps clarify what good feedback is (goals, purposes, expected outcomes);
2. facilitates the development of self-assessment (reflection) in the feedback given;
3. delivers high quality information to students (i.e. future assessors) about the quality of their assessments;
4. encourages teacher and peer dialogue around providing feedback;
5. encourages positive motivational beliefs and self-esteem;
6. provides opportunities to close the gap between current and desired performance (of feedback delivery);
7. provides information to assessors that can be used to help shape assessment.

With the role of self-regulated learning again at the core of this philosophy, the opportunity to execute the skill to be practised and improved becomes key. This focus is emphasised in the theory of *experiential learning*, which posits that learning is most effective when students create knowledge through a process of engagement, interaction, and conflict with rich and holistic experiences (Kolb & Kolb, 2005). If one is to take these two perspectives together – i.e. that evaluation is a skill to be not only learned but also performed – then existing methods of performance training that incorporate experiential learning provide a framework from which new forms of evaluation training and study can be adapted.

The classic form of simulated performance training in music is the dress rehearsal, in which a performance is conducted with every component in place save the audience themselves, thus allowing the performers (and in the case of larger productions, the off-stage support) to ensure that the extra-musical aspects of performance are in place. While this can include testing the practical components of performance – timings, clothing choices, the functionality of electronic or mechanical elements – the performers themselves also have the opportunity to check the technical, physical, and psychological aspects of their craft. Crucially, the dress rehearsal offers the possibility of dealing with the heightened physiological arousal inherent to performance, and its potential to have a maladaptive influence on outcomes should performers interpret this arousal as the manifestation of performance anxiety (Kenny, 2011; Nieuwenhuys & Oudejans, 2012; Endo et al., 2014). This applies not only to the on-stage experience, but also to the period of time spent backstage prior to the performance where performance-related physiological arousal has been found to be at its highest (Williamon et al., 2014; Chanwimalueang et al., 2017). Research has also suggested that the act of video-recording these sessions can also induce anxiety in student performers, again providing an opportunity to simulate the stress of a true performance (Daniel, 2001).

Assessment has been used as a form of experiential learning in educational settings. Indeed, the act of providing self- and peer-assessments as a part of the learning process has seen increased use across higher education, with one meta-analysis demonstrating a trend of strong correlations between peer- and faculty evaluations so long as global criteria are being used (Falchikov & Goldfinch, 2000). In the musical domain, pedagogy classes will investigate theories of teaching and modes of feedback delivery. These may include mock lessons conducted within the classroom or recorded for review by the instructor, which requires sourcing willing students for such experimental teaching. A traditional approach can be also found in the masterclass or studio class, in which the expert musician works with one or more musicians in front of an audience (i.e. the masterclass) or other students (i.e. the studio class; Gaunt, 2017). This basic template can be adjusted to accommodate multiple experts, students taught by their own or other teachers, or, crucially, opportunities for students to critique each other's performance in a controlled setting (Long et al., 2012). While the master/studio class offers obvious benefits for performers (further feedback from a variety of sources, opportunities to perform in public) and for teachers (opportunities to gain exposure as a master teacher, to reach and recruit new students, and to hone their own evaluative skills), those where student feedback is incorporated provides a platform in which musicians can test and develop their skills of attentive listening and viewing, of performance diagnosis, and of effective feedback delivery (Hanken, 2008, 2010; Taylor, 2010; Long et al., 2012; Haddon, 2014; Gaunt, 2017).

Whether a masterclass or studio class provides specific opportunity to examine the quality of feedback delivery depends largely on the focus and time mandated by the teacher. Otherwise the act of providing an evaluation serves more to enhance reflecting on the performative skill, rather than the evaluative. Studies examining the act of conducting peer- and self-assessments of video-recorded performances highlight performance-focussed feedback (e.g. Bergee, 1993, 1997; Robinson, 1993; Johnston, 1993). Daniel (2001) examined video-assisted self-assessment with 35 undergraduate music students at an Australian university, finding in a preliminary questionnaire that fewer than half of the students reviewed audio or video recording of their own performance with any kind of regularity.

Several studies have examined the act of having students conduct peer-to-peer feedback as part of their training, often examining live pilot programmes. Hunter and Russ (1996) worked with an Irish university to develop and monitor a seminar on peer assessment over several years. Students received training in the university's assessment procedures and assembled into panels of students with a variety of instrumental experience, a self-elected leader, and a supporting member of staff who had provided the initial procedural training. In post-evaluation discussions among the students, several extra-performance biases and complications were explicitly raised that have been revealed through subsequent research, including recognition that it was socially and emotionally difficult to provide a low mark despite a weak performance, that assessors playing the same instrument as the performer were harsher in their criticism than those without the specific expertise, that marks assigned often reflected pre-existing expectations of a particular performer (i.e. the so-called halo effect), that the relative relation between the assessor and performer (i.e. whether they were of the same or a different year group) coloured feelings towards providing and receiving the feedback, and panel disagreements were often unresolved due to expedience and a lack of discussion (see also Chapter 1).

Searby and Ewers (1997) examined the use of a peer assessment scheme within courses across a UK university's music programme, starting with an initial pilot in composition and expanding to areas including music performance, business, technology, and theory. In each setting students determined the criteria for assessment, gained initial experience through the evaluation of previous years' work, paired off for peer assessment to be moderated by the lecturer, and received 20% of their final mark for the quality of the written feedback they provided. The process for peer-assessing musical performance was conducted with performances of a different year group rather than previously documented work. With each subsequent year the groups negotiated a new set of evaluative criteria, which follow-on discussion with the students showed to be a critical component of their taking ownership of the evaluative process and thinking critically about creating their own work to be assessed. This feedback on the process also revealed that students were happy with receiving peer feedback and felt that it was a valuable learning tool. (Despite hopes that peer-assessment would reduce the evaluative workload of the faculty members, operating the programme did not lead to a significant reduction in their efforts.)

Following two studies demonstrating students' inconsistency in their self-and peer-assessment abilities compared with faculty-generated scores (Bergee 1993, 1997), Bergee and Cecconi-Roberts (2002) assembled experimental groups of 3-5 undergraduate music majors to perform for one another in four video-recorded sessions, after which they reviewed and discussed the performance footage while completing self- and peer-assessments using fixed rubrics. After self-evaluating recordings of their final jury recitals, these were compared with the evaluations by the jury examiners. No significant difference in ability to self-evaluate was shown based on year or performance level, and correlations between self- and faculty assessments were modestly higher among the experimental group compared with a control group who had not completed the peer assessment discussion sessions. However, a great deal of variability remained in the scores, especially in ratings of tone and interpretation. A follow-up experiment that included greater discussion of the evaluative criteria and their application to two sample scores also showed moderate to no effect of the treatment on alignment of self- and peer-assessments with faculty assessments, with the authors suggesting that the interventions had not fully engaged with the social and environmental complexities of performance self-assessment.

Daniel (2004) had 36 students who were involved in weekly performance seminars provide feedback on fellow student performances in the form of short evaluative comments and as detailed grades using a segmented scheme. Reflective questionnaires showed that students preferred the structured approach and that a trend of students to be too reserved in their critical judgements improved over the course of the sessions.

In Blom and Poole's (2004) research, 16 third-year music students were asked to evaluate second-year performances in an Australian university. Having completed self-assessment tasks in their first year and paired peer-assessment critiques in their second, they were tasked with grading recorded performances of their second-year peers using the same criteria employed by staff, providing written critiques to be read by the performers, assigning grades, and providing a self-reflective commentary on the process. Students struggled to cope with the variety of instrumental specialties they were asked to assess, the prospect of delivering harsh feedback when they already had a personal familiarity with the performer, adhering to a pre-existing set of criteria, and their ability or 'authority' to provide such assessments to their peers. As Hunter and Russ (1996) demonstrated, the students found the exercise to be helpful in not only developing their abilities and confidence in assessment but also how they might adjust their performance for assessment. Further research also followed on Hunter and Russ' use of student-chosen evaluation criteria, finding that students placed focus on 'soft' skills in assessing rehearsal quality – personal, interpersonal, and organisational skills – and 'hard' skills in assessing performance quality: technical, analytical, and musicianship skills (Blom & Encarnacao, 2012).

Lebler (2007) described the establishment of a 'master-less studio' in the execution of a course on popular music production at an Australian university in which students self-directed their learning strategies, outcomes, and outputs in collaboration with their peers. This included a structured method of peer evaluation in which recordings were shared and written commentary posted on a course website, amounting to over 180,000 words of feedback on 292 recorded tracks in one semester. Course conveners monitored whether the feedback conformed to good standards of constructive criticism, highlighting instances of overly authoritative tone or lack of appropriate detail, although specific instruction or focus on effective feedback production was not provided.

Latukefu (2010) examined a scaffolded peer-assessment framework among undergraduate vocal students at an Australian university. Adapting the model set by Searby and Ewers (1997), student focus groups established the assessment criteria and processes before the programme was implemented across a cohort. Following dissemination and discussion of the criteria to a class on contemporary performance practice, panels of three students performed peer evaluations. An open-ended survey found that students recognised the benefits of peer evaluation in improving their abilities to reflect upon their own performances, as well as developing skills important to their future work as evaluators. They highlighted the difficulties in conducting these evaluations with peers and friends, citing awkwardness and social influences preventing objective discussions of performance and assessment.

The Centre for Excellence in Music Performance Education at the Norwegian Academy of Music established peer learning and group teaching as a 'principal instrument study' (Hanken, 2016). Several approaches were employed, each a variation on a teacher-supervised studio classes in which students engaged in discussion of performance and feedback. One approach employed Lerman and Borstel's (2003) Critical Response Process, which comprises an initial discussion of what components of the performance are meaningful, the performer asking questions on which they would like feedback, the evaluators asking neutral questions of the performer, and finally the evaluators asking permission to give opinions on specific aspects of the performance, only delivering those opinions if asked. This study found that, in the most effective uses of the method, the fourth stage became redundant as the performer had already reached the relevant conclusions through the dialogue. Hanken also highlighted the role peer learning can play in continuing professional development of music teachers through seminars and discussion, combatting the isolation that can be inherent to music instruction through the nature of working practices.

More recently, Mitchell and Benedict (2017) employed peer-to-peer examination as a teaching tool during auditions at an Australian university. Rather than having the students provide evaluations in genuine grading scenarios, they rated live performances with or without a blinding screen in front of the stage, as well as recorded performances in audio only, visual only, and audiovisual scenarios to confront directly the issues of audio/video interaction inherent to music performance evaluation (as discussed in Chapter 4). The student judges felt more confident when rating performances in audio-only conditions and were prompted to reflect on the role of their appearance and stage presence in their own performances.

Finally, Dotger and colleagues (2018) adopted methods used in medical education to train physicians, targeting a specific form of feedback delivery in music teachers. Where a doctor may interact with a mock patient, the researchers had 13 trainee music teachers interact with a mock parent, herself coached to question the teachers as to why her daughter had not been successful in a recent (hypothetical) audition, the validity of the assessment itself, and whether her daughter had 'the look' (i.e. whether she conformed to the presumed stereotypes of performer appearance discussed in Section 1.6.2). Trainees had not been given prior instruction in how to navigate the interaction, thus their responses were highly variable. Several were able effectively to incorporate a combination of personal experience, acknowledgement of the parents' concerns, and specific advice for further development into their conversations.

In reviewing these approaches, several similarities can be seen. Each embraced experiential learning, not only giving students the ability to take part in the act of evaluation but in several cases also taking control over the terms and goals of the process. Those that measured outcomes found positive responses from the students and educators. However, simply providing learners the opportunity to evaluate others is not so simple a proposition, with several of the studies highlighting the workload costs of administering such training and acknowledging that many still felt unprepared to face the pressures of genuine evaluation situations. It is here that the gap is highlighted between artificially constructed assessments among familiar peers and settings and the heightened competitions, auditions, exams, and masterclasses in which the students will be called upon to make impactful decisions. This mirrors the discrepancy in research methodologies outlined at the beginning of this chapter, wherein artificial evaluative situations created to study assessment cannot capture the complexity of authentic assessments and the generalisability of the knowledge or skills gained. Alternatively, allowing learners (or researchers) access to true evaluative situations robs them of control of the situation and risks affecting the outcomes of those to be evaluated, especially if the evaluators in question are novices.

What is needed, therefore, is a way to recreate the complexity of a true or mock evaluation while maintaining control over the stimulus and setting to be evaluated. In the mock-parent study by Dotger and colleagues (2018), the authors describe the approach as a form of simulation, differentiating it from a role-playing exercise in that those taking part were told that the mock parent would never break from their character, and that the interaction could not be stopped or tried over. An existing approach embracing the concept of simulation can be found in the use of Immersive Virtual Environments (IVEs).

7.5 SIMULATING PERFORMANCE

7.5.1 Immersive Virtual Environments (IVEs) and distributed simulation

IVEs comprising some combination of projected visuals, aural and acoustic simulation, interactive physical environments, and closed narrative loops have now seen decades of use in both medical and social psychological settings (Blascovich et al., 2002a; Sanchez-Vivez & Slater, 2005). The simulation of performance as a training tool has seen considerable use in non-musical domains, including the development of pilots (Hamman, 2004), athletes (Miles et al., 2012), and firefighters (Bliss et al., 1997). A particularly fruitful domain has been that of medicine, where shrinking opportunities to gain experience with patients in consultation and surgery, the unending and exponential growth of clinical techniques to be learned, and increased pressure to reduce the amount of practising skills on patients is driving a shift to learning through simulation (Kneebone et al., 2010). While their efficacy was initially contested (Blascovich et al., 2002b), simulations can offer insights into issues of human perception and social behaviour, and their functionality has increased with the rapid growth in computational power and projection techniques. Furthermore, their ability to simulate risk while providing the operator with complete control over the environment has demonstrated their efficacy as a therapeutic tool to combat, for example, posttraumatic stress (Difede et al., 2007), and fear of flying (Rothbaum et al., 2000), spiders (Bouchard et al., 2006), and public speaking (Slater et al., 1999).

One branch of this work has been the advancement of distributed simulation, wherein alternatives to the advanced, complex, expensive, and/or immobile architectures that often typify simulation environments are developed that emphasise affordability, accessibility, and portability (Kneebone et al., 2010). In Kneebone and colleagues' example, a surgical theatre is reproduced in an affordable, inflatable room; expensive equipment is represented through life-size, high-fidelity photographs; lightweight versions of surgical lighting provide the intensity of a lit operating table; speakers recreate the genuine sounds of the operation space; a combination of affordable prosthetics and human actors provide the social, visual, and tactile experience of engaging with a patient. This approach emphasises recreating the function, rather than the structure, of the true environment, with particular focus on the aural and visual stimuli peripheral to the central task, and has been found to be an effective and adaptive form of training (Kassab et al., 2011). The affordable and portable nature of this approach, in particular, lends itself to the musical domain, where space and funds are regularly in short supply in music education institutions.

7.5.2 Simulating music performance

Several approaches to simulated performance training through Immersive Virtual Environments have been employed in music research. Orman (2003, 2004) employed a head-mounted display in which she simulated an empty practice space and seated audience of familiar peers, faculty members, or the head of bands performing an audition. Tests with eight saxophonists showed some evidence of increased heart rate in several participants, although results were inconclusive due to lack of correspondence with physiological scales and lack of experimental control. Bissonnette and colleagues (2011, 2015) had nine guitarists and pianists perform several sessions in a virtual environment comprising a classical music audience and/or panel of three judges giving a variety of reactions and interjections presented via four large screens in a three-dimensional arrangement, speakers, and stage lights. When state anxiety scores were taken following public performances before and after these sessions, participants with high trait and initial state anxiety showed a reduction in state anxiety across the two performances significantly greater than those of a control group who had not experienced the virtual environment. Significant increases in third-party-assessed performance quality were also noted in the experimental group. Further study tracked changes in reported anxiety within each of the six one-hour sessions, finding a decrease in anxiety provoked by the simulation in subsequent sessions so long as similar musical material was being presented (Bissonnette et al., 2016).

A different immersive approach to the simulation of musical performance can be seen in the development and operation of Williamon and colleagues' (2014) *Performance Simulator*. The platform recreates an intimate concert recital with 24 audience members or an audition for a panel of three expert judges. To create the audience, eleven participants were filmed via green-screen performing typical random movements of concert viewing, as well as providing specific responses (e.g. mild applause, booing, a standing ovation, etc.). Accompanying audio was recorded separately. This footage was then compiled into a digitally constructed representation of a concert space, which was itself embedded into a software programme that allows the operator to trigger the various reactions via a keyboard, in addition to cuing coughs and mobile phone rings intended to test the performer's concentration. For the audition simulation, three professional actors were recorded while seated at a table recreating the effect of an audition panel. Following a neutral greeting to the performer, they can be triggered to provide an overtly positive, neutral, or negative mode in their passive listening, conveyed through eye contact, facial expression, and body language, and in their final response.

Following Kneebone and colleagues' (2010) goals of distributed simulation, the goal of the *Performance Simulator* was to replicate not only the panel or audience, but also the surrounding environment. In addition to the stage lights as used in previous simulations (Bissonnette et al., 2011, 2015), curtains were placed alongside the screen and a darkened, stage-light atmosphere replicated in the room. A backstage area was also recreated including dimmed lighting, music stands, seating, audio bleed from the stage comprising indecipherable chatter for the audition panel and the sound of an audience taking their seats for the concert setting, the latter of which was also featured on CCTV footage of a comparable performance space and audience. An operator played the role of a 'backstage assistant', guiding the performer through the experience while operating the virtual panel or audience. Crucially, this actor interacted with the performer as though the event were a genuine performance, and the performers themselves were expected to come wearing concert dress and to allow themselves to be caught up in the experience. Examination of electrocardiographic and self-reported state anxiety data among seven violinists demonstrated that the simulation provoked anxiety responses comparable to a live audition, and further qualitative research found that students perceived the simulation to be an effective tool to provoke and train for performance anxiety (Aufegger et al., 2017).

This work was followed by Glowinski and colleagues (2015) in which the projected audience comprised fully-digitised audience avatars standing in loose formation in a large, simulated concert space and projected in an immersive, three-dimensional configuration. As the audience members were rendered in real time it allowed the operators to manipulate the audience's behaviour; in this case, the audience's 'engagement' was manipulated via altering the proportion of avatars fixing their eye gaze on the performer versus those whose gaze moved randomly and disinterestedly through the space. Using this, the researchers were able to demonstrate through motion tracking how four violinists' performance movements were altered, although not consistently, under different audience conditions.

Based upon these existing simulation approaches, this chapter will now present the novel conceptualisation and development of a tool to apply the concepts of Virtual Immersive Environments and distributed simulation to the practice and study of music performance evaluation.

7.6 THE EVALUATION SIMULATOR: A NEW TOOL FOR RESEARCH AND PRACTICE

There is a clear need for further approaches to study the act of live performance evaluation in a controlled environment and to improve and expand the delivery of assessment training. Musicians require access to skilled evaluators to provide feedback on their own performance and to develop skills as assessors to prepare for portfolio careers and enhance their self-evaluative abilities. Teachers and educational institutions have a duty to ensure they are preparing their students for careers that include teaching and assessing and to ensure that the evaluations they provide of their students are fair and robust. And researchers require new means to investigate and control experimentally the myriad social and environmental factors that influence the act of decision-making.

While numerous approaches have been described that apply the tenets of experiential learning and simulation through mock experience, none have embraced the possibilities of IVEs or distributed simulation in recreating the surrounding and intensifying stimuli of the true evaluative experience. This is akin to the pianist experiencing a ‘performance’ in a closed room with their peers, minus the time backstage, the concert dress, the darkened hall, the stage lights, the unfamiliar audience, and the true pressure of a live performance. It is these features that music performance simulations have sought to replicate. A genuine performance evaluation, as discussed above, can come with the same pressure of performance. Increased arousal can limit the ability to attend to and process information (Hanoch & Vitouch, 2004), which is also central to the act of performance assessment. Thus, the goal of the present work was to develop an immersive simulation that stimulated the heightened pressure of performing an evaluation to allow for immersive and experiential training while providing a controlled setting to facilitate experimental research.

To address these issues, the *Evaluation Simulator* was developed to allow for the recreation of the following scenarios in research and training:

1. evaluating an expandable set of replicable stimuli;
2. evaluating alone or as part of a panel;
3. evaluating in a heightened setting, such as in a live competition or masterclass, where the judges themselves are a focus of attention;
4. having to evaluate a performance of good or poor quality;
5. having to deliver summative, diagnostic, and/or formative evaluation directly to the performer immediately and verbally;
6. having to deliver that feedback to a performer who is in a variety of emotional states.

7.6.1 Development

A primary question in developing the simulation was in the fundamental mode of stimulus presentation – that is, how the performance would be immersively visualised. The music performance simulation literature presented three existing approaches: (1) a head-mounted virtual display (Orman, 2003, 2004), (2) a projected visualisation of 3D rendered avatars (Bissonnette et al., 2011, 2015), or (3) a projected visualisation of looped video recordings (Williamon et al., 2014). The head-mounted display, while offering perhaps the most 'immersive' of the approaches, was discounted due to the difficulties in engaging multiple people simultaneously with the simulation and the relative complexity and cost in developing and operating the platform. A system employing a large display or screen and projector typical to education settings was thus determined to be the most appropriate for the intended use cases.

With regard to artificially-rendered avatars, they provide several advantages: (1) they allow for complete control over audience behaviour, reactions, and appearance, theoretically providing infinite variety in audience conditions; (2) they provide the opportunity to generate audiences that are dynamically reactive to the performer, altering their behaviour as a true audience might in response to the quality and expressiveness of the performer (a stated objective of Glowinski et al.'s 2015 research); and (3) they theoretically allow for seamless transitions between presentation modes (e.g. a stationery to an applauding audience) as transitions can be rendered in real time, where use of video often necessitates noticeable transitions or 'jumps' between sets of pre-recorded footage. However, such an approach comes with drawbacks. Despite exponential advances in the ability to create lifelike human avatars and repeated demonstration that they can provoke realistic responses, they tend to remain across the 'uncanny valley' that separates them from being perceived as true human representations (Kätsyri et al., 2015; de Borst & de Gelder, 2015). This has particular salience in music performance evaluation considering the highly influential role of the performer's behaviour and appearance in performance evaluation (see Chapter 4). The use of pre-recorded video loops eliminates this problem and allows for photorealistic performers. With a carefully controlled protocol and instructions, it offers the possibility of convincing users that they are interacting with a genuine audience or auditioner via a videoconferencing system.

Considering the limitations of these technologies and of existing practice described throughout this chapter, 10 qualities were determined as crucial in development of the *Evaluation Simulator*. These were as follows (and are summarised in Table 7.1):

1. **Experimentally replicable:** replicability was the primary goal of the simulator, i.e. providing experiences that could be duplicated within and across students/study participants. This would not be possible in mock or true performances, and while assessing lone recordings allows for replicability of the evaluative experience, an IVE is necessary to immerse the judge in a stimulating environment.
2. **Immersive:** the experience must be free from extraneous distraction and provide a full sensory experience of the evaluation. Mock evaluations offer potential here, if a suitable environment is created, although IVEs specifically tailor this experience.
3. **Heightened arousal:** the immersion should seek to increase the arousal experienced in completing the evaluation, mirroring the risk of the true situation. Again, mock evaluations have the potential to recreate this, although examples in the literature are lacking.
4. **Risk-free for performer/organisation:** conducting genuine evaluations defined by real impact on the grades/standing of the performer introduces risk for those being evaluated. A simulation should recreate this tension while avoiding the need to influence actual assessment procedures.
5. **Photorealistic:** due to the importance of visual performance features, looped recorded video within an IVE would be ideal as used in Williamon et al.'s (2014) *Performance Simulator*.
6. **Allows solo and group evaluation:** the simulator should allow a panel of evaluators to interact in a genuine physical environment. This is a particular challenge for VR applications, which naturally isolate the user within the head-mounted display.
7. **Inexpensive to create:** to determine an approach that could be widely adapted following the goals of distributed simulation, the complex computing expertise

and equipment required to generate immersive VR or computer-generated avatars precluded their use in this simulator.

8. **Inexpensive to operate:** the equipment required for the employment of VR simulation is not readily available in most music learning environments. Mock evaluations have the potential to incur great expense if performers/actors need to be hired.
9. **Adaptable:** true performances are restricted by nature. Mock evaluations and simulations rendered in real time offer infinite adaptability. While video simulations are more restrictive in their adaptability, multiple scenarios could be filmed in advance and combined to allow an exponential number of possible use cases in combination with variations in the environment.
10. **Portable:** the experience must be operable in a wide variety of physical locations, with minimal effort and cost required in transporting it.

Table 7.1 summarises these points and the degree to which traditional evaluative environments used in research and teaching (assessing recorded videos, mock evaluations, and true evaluations) and the options for IVEs described earlier (VR displays, 3D rendered displays, and looped video displays) meet the demands. As a result of this summary, it was determined that Williamon et al.'s (2014) *Performance Simulator* provided the best model upon which the *Evaluation Simulator* would be based. To achieve this, performance footage would need to be recorded, combined in an interactive software framework, and presented within an artificially created physical and social environment. This process is outlined in section 7.6.2.

Table 7.1. The qualities of traditional and immersive virtual environments (IVEs) in the training of evaluative skills and in research.

Needs of the *Evaluation Simulator*	*Traditional Environments*			*Immersive Virtual Environments*		
	Video review	Mock	True	VR Display	3D Display	Looped Video
Experimentally replicable	Yes	No	No	Yes	Yes	Yes
Immersive	No	Potential	Yes	Yes	Yes	Yes
Heightened arousal	No	Potential	Yes	Yes	Yes	Yes
Risk-free for performer	Yes	Yes	No	Yes	Yes	Yes
Photorealistic	Yes	Yes	Yes	No	No	Yes
Allows solo and group eval	Yes	Yes	Yes	No	Yes	Yes
Inexpensive to create	Yes	Yes	Yes	No	No	Yes
Inexpensive to operate	Yes	Potential	Yes	No	Yes	Yes
Adaptable	Yes	Yes	No	Yes	Yes	Yes
Portable	Yes	Yes	No	Yes	Yes	Yes

7.6.2 Recorded video

7.6.2.1 Stage and setup

The stage setting was designed to be ambiguous in the size of the space in which the performer was appearing, allowing the simulation to be physically displayed in a variety of settings without creating visual conflict. To achieve this, the video was shot against a black-curtained backdrop without side walls or ceiling visible, leaving the size of the space ambiguous. Carpeted floor was also chosen to maximise transferability to alternate spaces, as this could be interpreted as a rug placed over the local flooring. A tight camera framing was used, maximising the size of the performer in the shot while ensuring his entire body remained in frame at all times. This served several purposes: (1) guaranteeing the whole body could be seen without cut-off to give the strongest impression of a performer in the room with the evaluator; (2) allowing the assessor to judge the full range of body movement; (3) maximising the size of the instrument and hands to facilitate instrument-specific technical assessment; (4) maximising the size of the performer's face to facilitate social cues; (5) allowing the performer to be projected as close to life-size as possible on a standard, stand-mounted projector screen to facilitate the simulation; and (6) minimising the perceived distance from the performer to allow for a more socially intense setting.

Professional studio lighting and audio/video capture equipment was used to maximise the veracity of the videos and facilitate the simulation. The performer was asked to wear semi-formal clothing appropriate for a high-level orchestral audition (see Figure 7.1).

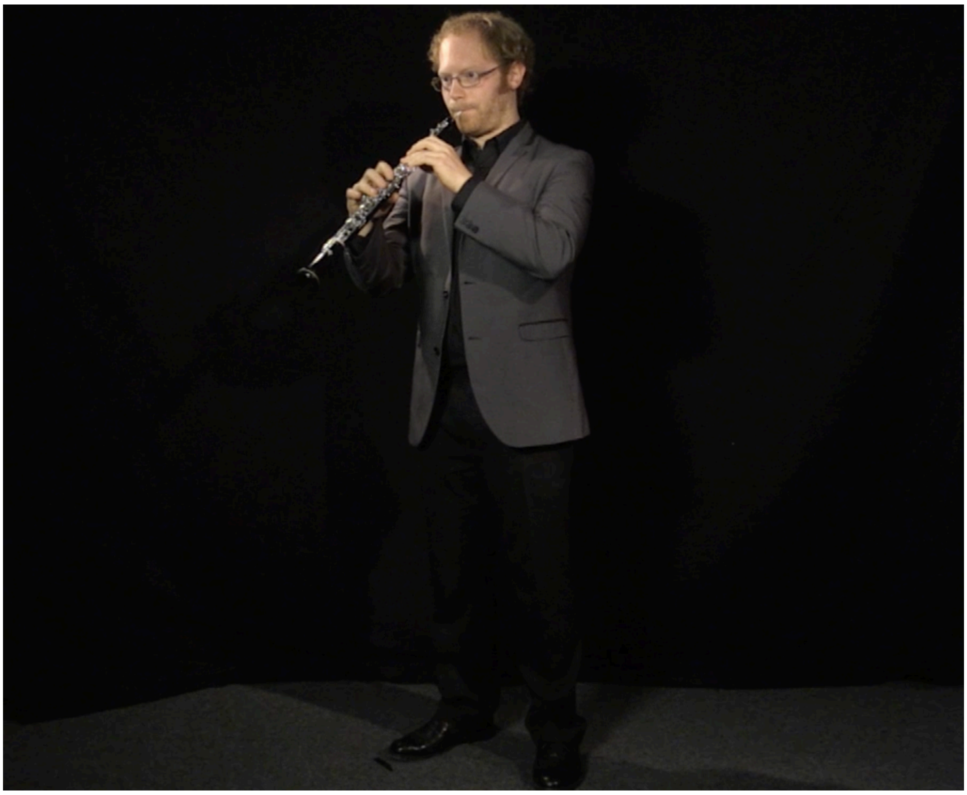


Figure 7.1. Framing of the performer in the recorded video. The size of the performer in the scene was maximised to enhance the effect of the simulation.

7.6.2.2 Performance footage

The performer, a semi-professional oboist, was asked to prepare two excerpts of standard orchestral repertoire typical of a professional audition. The excerpts were chosen to vary in tempo and style: a relatively fast work emphasising articulation, ornamentation, and rhythmic drive, and a relatively slow work to demonstrate melodic phrasing and breath control. Respectively, these were the oboe solo opening of the *Prélude* of Maurice Ravel's *Tombeau de Couperin*, bars 1-14, and the oboe solo opening of the second movement of Tchaikovsky's *Symphony No. 4, Op. 36*, bars 1-21 (see Figure 7.2). For each work the performer delivered two performances for a total of four: a 'good' performance of high playing standard, and a 'poor' performance in which he struggled with intonation, tempo, and tone and displayed mild facial frustration.

7.6.2.3 Extra-performance footage: Entrance, feedback, and exit

The beginning of each of the four recorded performances opened with the empty stage, followed by the performer walking in and standing on a mark facing the camera. In each case, the performer was asked to face the hypothetical judging panel, wait approximately three seconds to leave time for a brief welcome and indication to start, give a nod of acknowledgement, then begin performing. The same activity was recorded ahead of each of the four performances.

Following the performance, the oboist was asked to face back towards the panel to receive feedback. At this point, three modes of feedback reception were filmed: (1) *confident*, in which the oboist was instructed to appear resolute and stoic, ready to receive positive or negative feedback in stride with direct eye contact and occasional nods of understanding; (2) *frustrated*, in which he was asked to appear disappointed in his performance and to not give the panel his full attention, avoiding eye contact and punctuating his reaction with subtle eye rolls, sighs, and grimaces; and (3) *distraught*, in which he was told to appear in a poor emotional state following the performance, looking at the floor and giving the impression of holding back tears with the expectation that poor or harsh feedback would be given (see Figure 7.3). Each feedback scenario was recorded for 60 seconds, with the performer instructed not to

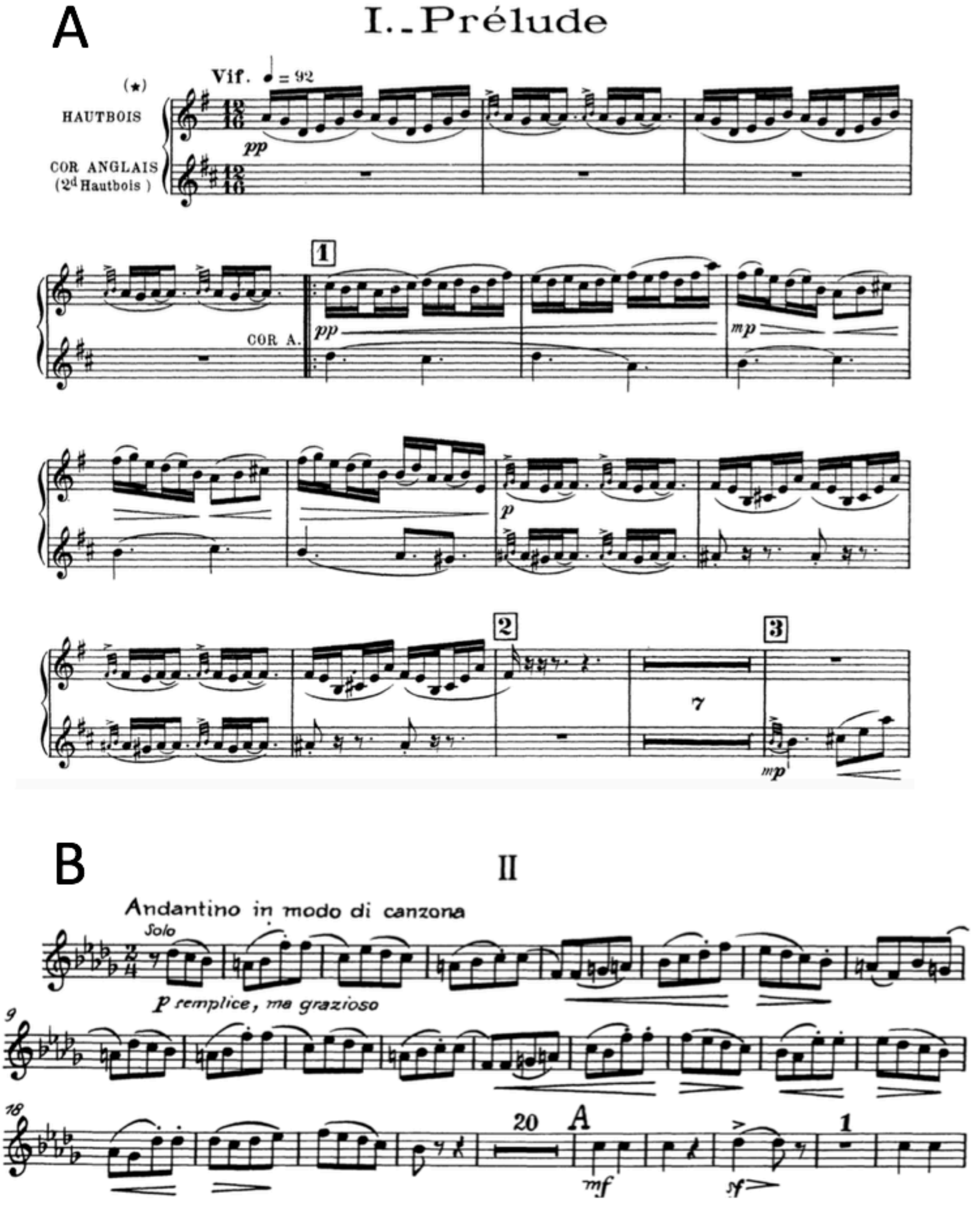


Figure 7.2. Musical excerpts recorded for the simulation. Top panel A: oboe solo from the *Prélude* of Maurice Ravel's *Tombeau de Couperin*, bars 1-14 (Ravel, 1919, p. 1); bottom panel B; oboe solo from the second movement of Tchaikovsky's *Symphony No. 4, Op. 36*, bars 1 - 21 (Tchaikovsky, 1946, p. 6).

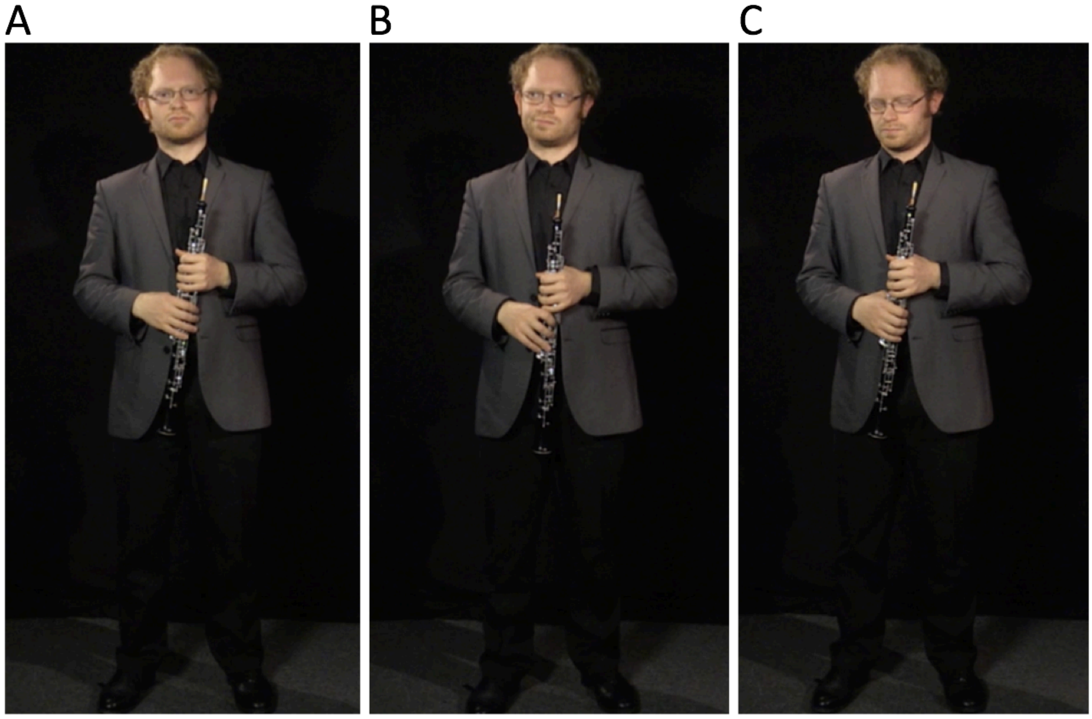


Figure 7.3. Screenshots of the performer's three reaction modes. Panel A: *confident*. Panel B: *frustrated*. Panel C: *distraught*. These reactions can also be seen in the video files in Appendix 9.

change standing position and minimise torso movement to allow the segment to be looped (described further below). Each of the three feedback scenes was concluded by the performer saying "thank-you very much" or "thanks" to the panel in the style of each setting – confident and gracious, brief and dismissive, barely audible and distraught – and walking out of frame in the direction he entered.

A summary of the seven pieces of video footage collected can be found in Table 7.2, and the video files themselves can be found in Appendix 9. Screenshots of the three performance reactions (*confident*, *frustrated*, and *distraught*) are shown in Figure 7.3.

Table 7.2. Summary of the video footage collected. Stage entrances, performances, reactions, and exits were recorded in grouped sequences as described by codes A - D. In the *Evaluation Simulator*, any "Entrance & performance" may be paired with any "Reaction & exit", allowing for 12 possible permutations. The video files can be found in Appendix 9, where the three reactions and stage exits are paired with three of the performances following codes A - D.

Video	Code	Category	Description
Video 1	A	Entrance & performance	Ravel (fast), good quality
Video 2	B	Entrance & performance	Ravel (fast), poor quality
Video 3	D	Entrance & performance	Tchaikovsky (slow), good quality
Video 4	C	Entrance & performance	Tchaikovsky (slow), poor quality
Video 5	A	Reaction & exit	*Confident*
Video 6	B	Reaction & exit	*Frustrated*
Video 7	C	Reaction & exit	*Distraught*

7.6.3 Software

Figure 7.4 outlines the interaction mapping of an Adobe Flash-based software interface developed to manipulate the videos using keyboard commands. Upon opening the program (and setting to full-screen view), the software holds a still image of the empty stage. By pressing keys 1-4 the operator triggers one of the four recorded performances (i.e. Ravel versus Tchaikovsky; good versus bad), which triggers the stage entrance and performance. Following the performance, the neutral reaction is then triggered by default with a dissolve transition between the two consecutive videos; the operator can trigger the *frustrated* or *distraught* reactions by pressing the 'B' or 'C' keys at any point following the beginning of the performance. The last key pressed triggers the corresponding reaction, and the 'A' key returns the reaction to *confident*. Once one of the reaction videos have been triggered, it remains on a continuous loop until the operator closes the session by pressing the space bar, which triggers the corresponding "thank you" and the performer's exit sequence.

The interface can also be operated using a standard USB presentation remote. In this case, the equivalent of a slide advance triggers the 'good' Ravel performance with a confident reaction, and another click triggers the stage exit. This can also be used to end any of the reaction loops if they had been triggered by the computer keyboard.

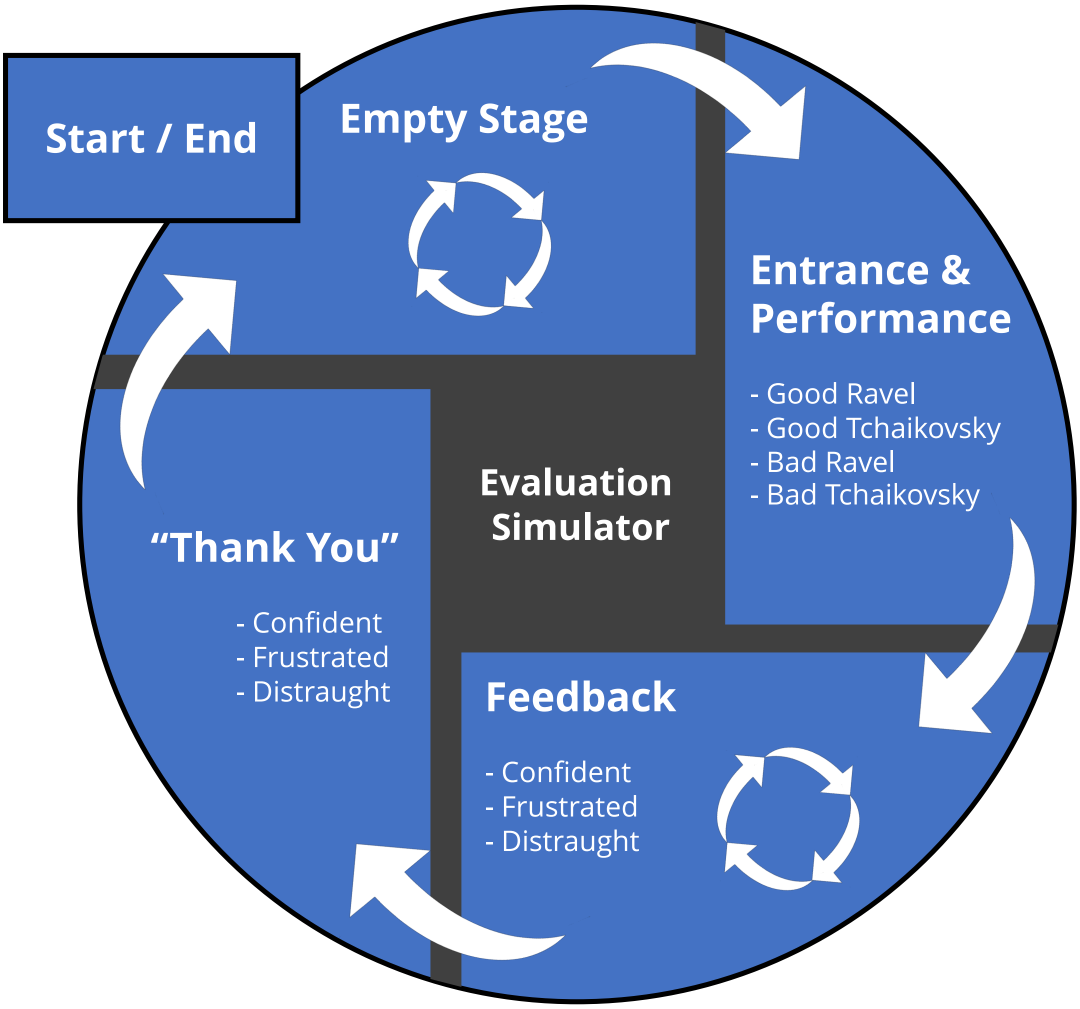


Figure 7.4. Process mapping of the software interface. Following a hold of the empty stage shot, pressing keys 1-4 triggers the stage entrance and respective performance. During this performance, selecting the 'b' or 'c' keys prompts the eventual transition from to the appropriate reaction (which otherwise goes to the default *confident*). Once the looped feedback reaction is no longer needed (which can be quickly skipped if verbal feedback would not be appropriate for the scenario), the space bar triggers the stage exit and returns the software to the original stage, ready for another evaluation.

7.6.4 Physical environment

While the recorded video and software interface provides the core simulator experience, it is augmented by features of the physical environment in which it was designed and into which it can be set up. The configuration used here mirrors that of

Williamon et al.'s (2014) *Performance Simulator*. The projection screen (or large monitor) is placed against a wall and flanked by heavy curtains, giving the impression of a stage space extending beyond the physical room. Where possible, the screen is placed at floor level to give the impression of the performer standing in the room; where the screen must be raised, the gap at the bottom can be blocked to give the impression that the performer is standing on a raised platform or stage. The curtains and screen are topped by remote-operated stage lights, directed back at the panel to heighten the feeling of attention and pressure on the decision-making process. The room is best left darkened to maximise the effect of both lights and projection. High-quality speakers are placed are placed as close to the projection as possible. A table and chairs for the panellists is placed at the centre of the room, to which props can be added that are common to a judging experience (e.g. glasses of water, clipboards, judging rubrics, desk lighting; see Figure 7.5).

7.6.5 Operation

A crucial component of the simulation is the human operator and the supporting theatre he or she provides; the operator must treat the situation as a genuine performance and not allude to the artificial nature of the environment, emphasising the role of simulation over role-play (Dotger et al., 2018). The details of the operator's role can alter based on the specific setting, but generally comprises a welcome and introduction, briefing on evaluation protocols, orally calling in the mock performer (with accompanying triggering of the stage entrance sequence and desired performance sequence), triggering the desired reaction sequenced if not the default, triggering the stage exit at the appropriate point, and providing the closing and debriefing of the user. The operator may be serving alongside a researcher, teacher, and/or one or more mock panellists performing their respective roles, or they may be serving these roles themselves.

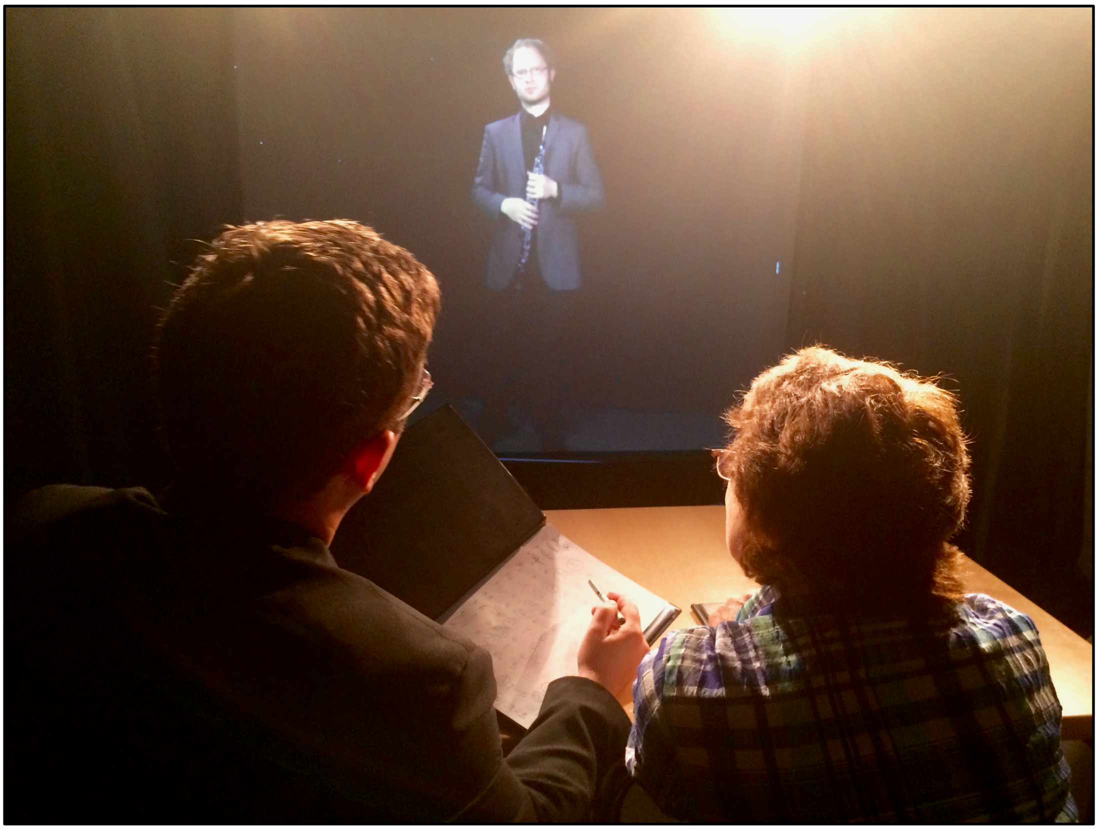


Figure 7.5. Two evaluators delivering performance feedback in the *Evaluation Simulator*. Stage lights illuminate a user and a facilitator in the environment, delivering feedback to the performer in the *confident* feedback mode.

7.6.6 Initial piloting

The simulator was piloted at the 2015 Cheltenham Music Festival, where it was set up as a public engagement event to allow festivalgoers to experience the heightened effect of performing as a competition judge akin to those popularised by the *Idol, X Factor,* and *…'s Got Talent* series. This also provided an opportunity to test the simulator's functionality as a piece of distributed simulation in whether it could be set up quickly in a space not designed for such use and provide an effective simulation. The collapsible lights, curtains, and projection screen and portable projector were assembled in a darkened storage room, with table and chairs locally sourced. Three operators facilitated the event: one to greet, brief, and debrief guests on their experience, one to act as a fellow panellist to the guest and prompt them to provide feedback to the performer, and one to operate the simulation from backstage. Public response was positive, with guests highlighting the intensity of the experience and several questioning whether the performer in question had been videoconferenced in due to his coincidental 'reactions' to statements they had made in their feedback. While further validation is required, this pilot suggested the goals of immersion, increased arousal, adaptability, portability, and cost-effectiveness to operate was achieved.

7.7 APPLICATIONS

The benefits of IVEs and distributed simulation have already been seen in the domains of medical and music performance training, providing new avenues to promote experiential learning and provide a platform to conduct performance research in controlled environments. The *Evaluation Simulator* provides the first opportunity to apply these benefits to the study and training of music performance evaluation. As the adaptability of the software and surrounding social environment provides a variety of permutations, potential applications can be posited for its use in research and teaching.

Before addressing these possibilities, it is important to highlight a central limitation of the simulator at this stage. While it was created with the goal of stimulating heightened arousal, a full efficacy study will be required to demonstrate whether the simulator is truly capable of evoking similar physiological responses to genuine evaluative settings, as was demonstrated with the *Performance Simulator* (Aufegger et al*.,* 2017). Such work, however, would be complicated by a lack of knowledge of the real-world analogue. While much is known about the anxiety experienced by musical performers (e.g. Kenny, 2011; Nieuwenhuys & Oudejans, 2012; Endo et al., 2014; Williamon et al., 2014; Chanwimalueang et al., 2017), no work to date has examined the physiological experience of the music examiner or competition judge. A major line of research is required to achieve this aim, one in which the *Evaluation Simulator* could play a central role. A second limitation is the range of performances available for evaluation: while quality and response can be varied across the two performances for a total of twelve evaluation scenarios from the video alone, they are nevertheless restricted to one performer on one instrument with two pieces of standard repertoire. However, the existing conceptual and software framework could be expanded with relative ease, requiring only the collection of new video footage with different performers, instruments, and repertoire while following the same script of entrance, performance, feedback, and exit footage. Over time a library of performances could be assembled, and even shared between groups or institutions.

7.7.1 In research

In its current state, the simulator offers numerous possibilities as a tool for research. By giving controlled, replicable stimuli for evaluation in a heightened setting, it provides an ideal tool to examine the causal relation of environmental and social factors on evaluation procedures. At a fundamental level, studies could be conducted comparing the evaluation of pre-recorded audio and/or video in laboratory conditions (i.e. watching the provided videos on a computer screen: AAA evaluation studies as described at the opening of this chapter) with varying degrees of heightened environmental arousal. Variations could include computer screen only, full-sized projection, or with or without pre-evaluation waiting period, performer stage entrance, or intense lighting. Social features could also be adapted, including informing the participant that the performer is being broadcast live via videoconferencing with possible real-world implications of the evaluation, or by providing additional information about the performer's experience and history.

The variety of pre-programmed responses could be used to examine differences in quantitative and qualitative feedback as affected by the performer's state, including whether a distraught performer triggers empathic reactions and more forgiving evaluations, especially when paired with the good versus the poor performance. The role of facial features in evaluation as explored in Chapter 4 could be expanded here to see whether a frustrated or distraught reaction following the performance affects how the musical component is remembered and contextualised. In addition to evaluators' written and oral responses, their behaviour (e.g. hand gestures, eye contact, rate and pitch of speech, etc.) and physiology (heart, respiratory, skin conductivity, etc.) could be monitored to determine differences across time, especially as they relate to the nature and speed the of feedback given as defined by time to first and final decision in Chapters 3 and 4.

As the simulator is conducive to panel judgements, it also offers the possibility of examining elements of intra-panel conformity and social response, such as furthering the celebrated conformity studies of Asch (1956) and examining whether artificially positive or negative evaluations from one or more actors playing the role of assumed fellow panellists affect subsequent judgements by the participant. This interaction could be examined at all points of the evaluation: the time spent before the evaluations when 'insider' information or initial impressions might be shared; the time during the performance where a variety of non-verbal cues might be used to indicate positive or negative response; direct responses of the actor(s) to the performer; and the time spent after the performer has been dismissed but before the final assessment is provided. In each case, a combination of continuous measures methodologies (see Chapters 2, 3, and 4) and written ratings could be used to capture changes in perception at varying points in the evaluation.

7.7.2 In pedagogy

The same features highlighted in research use can also be applied to training situations, allowing students to experience the intellectual complexity of delivering effective summative assessments and diagnostic feedback while navigating the procedures of an audition-like process and contending with the heightened social situation.

Care must be given in how best to employ the simulator in pedagogical settings. Through a review of studies in the medical domain, Issenberg and colleagues (2005) outlined 10 good practices in using simulation in training settings. They highlighted how (1) feedback should be given during the learning experience, (2) learners should practise their skills repetitively, (3) simulators should be integrated into the overall curriculum rather than used in extra-ordinary circumstances, (4) learners should practise with increasing levels of difficulty, (5) simulators should be used with a variety of learning strategies, (6) simulators should capture a variety of contexts, (7) learning should occur in a controlled environment, (8) learners should be provided with individualised experiences, (9) clear outcomes and benchmarks should be provided, and (10) the validity of simulators should be demonstrated. In its current form the *Evaluation Simulator* fosters repetition (2), a range of difficulty (4; i.e. the differing performance qualities and responses) and the controlled environment (7). The need to validate the simulator (10) has already been discussed, as has the possibility to expand the simulation to a wider variety of contexts beyond what is already possible through variations in the software interface, social, and environmental factors (6). The use of varying strategies (5) while providing individualised learning (8) will be up to the instructor, who can vary the use of group size or use of instructor-versus-peer led settings. For example, a lesson might have students enter alone, with the instructor as a panel leader, with a panel of peers, or with a panel of strangers, depending on the experience most needed by a particular student or group. The use of benchmarks (9) and ongoing feedback (1) will also require creative thinking as to what constitutes an effective assessment, drawing on the criteria adapted from Nicol and Macfarlane-Dick (2006; see section 7.4) to establish when feedback given is effective and informative and using peer- and video-stimulated approaches to provide *feedback on the feedback*. Finally, adoption into the curriculum (3) will require support not only from students and teachers but programme leaders, facilities managers, and administration. The use of distributed simulation to ensure the *Evaluation Simulator* is as cost-effective and adaptable as possible might help this adoption and lead to lasting change.

7.8 SUMMARY

This chapter highlighted a gap in the methodological and pedagogical approaches to the study and training of performance assessment. By considering the processes by which expert performance is trained, practiced, and simulated, particularly the concepts of Interactive Virtual Environments and distributed simulation as employed in the domains of medicine and music performance training, it has outlined the development of the first *Evaluation Simulator* to allow for musicians and researchers to control and engage with the process of performance assessment in an ecologically rich environment. Much remains to be done in understanding the full experience and process of conducting a performance assessment, thus the intention of the approach presented here is to provoke and facilitate the next generation of innovation in performance evaluation understanding and practice.

8 DISCUSSION AND CONCLUSIONS

8.1 INTRODUCTION

This thesis set out to examine the processes and products of evaluative decision-making in music performance, and this chapter brings together the research conducted across its component studies. The findings from each chapter are considered against the fundamental research questions posed in Chapter 1 and situated in the relevant literature. Implications of the research for musical practice are then discussed, considering first the music competition and audition, then moving to guidance for educators and performers, and finally implications for audiences and concert programmers. Domains beyond music are then examined, including how the findings align with and can augment theory and strategies for effective assessment and training in other areas. Finally, limitations of the research are presented, avenues for further work suggested, and overall contributions to knowledge stated.

8.2 THE RESEARCH QUESTIONS

In Chapter 1, a review of the literature led to the framing of a process model of performance quality evaluation (see Figures 1.4 and 8.1). That model posited a series of nested factors categorising the features inherent to a performance evaluation. It also focussed on the act of evaluation as a process rather than an individual product, prompting an investigation of how the musical and extra-musical components of a performance might affect and interact with the evaluative process at differing points at it unfolds. This led to the formation of five research questions, prompting a series of four studies which investigated the act of evaluation from numerous vantage points. Those research questions were:

RQ1. When are decisions made and adjusted while assessing the quality of a musical performance?

RQ2. How is the process of music performance evaluation affected by variables relating to repertoire?

RQ3. How is this process affected by variables relating to the performer?

RQ4. How is this process affected by variables relating to the environment?

RQ5. How is this process affected by variables relating to the evaluator?

Following these research questions, four empirical studies were designed and conducted, and a new theoretical approach and methodological tool for considering performance evaluation introduced. Study 1 (Chapter 3) focussed on the repertoire with an experimental study of manipulated audio tracks. Study 2 (Chapter 4) used a similar experimental methodology, examining effects of the performer with manipulated video performances. Study 3 (Chapter 5) employed a survey design in a live, professional concert setting to examine the relationship between the audiences' affective state and their decision-making. Study 4 (Chapter 6) used a similar approach to examine the relationship between these ratings and aspects of the physical and social environment. Finally, Chapter 7 considered the complexity of the evaluative environment and the evaluator's skillset to set out a new paradigm for research and training in performance assessment. Each of these studies, by definition, included the evaluation by an evaluator of a piece of repertoire performed by a musician in an appropriate environment. However, each focussed on one or more particular components by nature of the methodology, stimuli, and questions asked. While the individual results have been presented in each respective chapter, results across the four studies are now discussed as they relate to the five research questions.

8.2.1 RQ1: The evaluative process

The first research question and Studies 1 and 2 focussed on the temporal features of forming music performance quality assessments: specifically, when decisions are first made, when they are finalised, and when they might drastically change. Existing work by Thompson and colleagues (2007) provided the first quantitative data in music on the time taken to form first decisions. They demonstrated that musically experienced listeners began reporting their initial judgements approximately 15-20 seconds from the first note. The study used error-free performances of short (3-minute) works by Bach and Chopin presented in an audio-only format following four seconds of silence, and the finding was consistent whether musicality, technical proficiency, or overall quality were being rated. The experimental Studies 1 and 2 served as a replication and extension of these findings across different conditions. Study 1 found the same effect when presenting several works of Chopin of 2 - 3 minutes in length in the same audio-only format and using the same software, with first ratings being made at a mean of approximately 13 to 17 seconds depending on the work. Study 2 extended this further with an audio-video presentation of a 3-minute Chopin performance preceded by a (confident) stage entrance with a newly created capture tool and with a group of raters with little-to-no musical experience. Again, a mean of approximately 18 seconds was found to the first rating, unaffected by musical experience. Both studies also found that these times did not correlate with the value given at the first or overall ratings, suggesting the decision-making *process* in these conditions was separate from the actual decision taken. These findings support those of Thompson and colleagues (2007), in that the time to first decision seems to be a relatively stable phenomenon that is dramatically shorter than the several minutes suggested by the early assessment literature examining first impressions in interviews (Tucker & Rowe, 1977; Tullar et al., 1979; Buckley & Eder, 1988).

Studies 1 and 2 also provided the first experimental demonstrations that this window of time to first decision could be significantly altered by manipulating specific performance features. This included evidence of a shorter mean time to first decision with an inappropriate stage entrance (Study 2: within 8 seconds), performance errors in the opening seconds (Study 1: within 6 - 8 seconds), and participants being told in advance that the work was just 30 seconds in length (Study 1: within 11 seconds). The time to first decision could also be significantly lengthened when presented with a work by an unfamiliar composer in an atonal style (Study 1: with 35 seconds). Thus, while these studies provided supporting evidence that the range of 15 - 20 seconds to first decision can be taken as a model for performances of standard repertoire of several minutes in length, it suggested that this is not a fixed process but is affected by other features (discussed further in sections 8.2.2 and 8.2.3).

Regarding the time to *final* decision, Thompson and colleagues (2007) found that, when comparing mean quality ratings at 15-second intervals, the participants' aggregate score did not differ significantly from the final rating at approximately 60 seconds into the performance. This method was used in Studies 1 and 2 using 10-second increments, finding a much shorter time of 30 seconds to the final decision, even in the presence of the inappropriate stage entrance, with a longer time of 90 seconds found for the unfamiliar Caprice in Study 1. However, as noted in Thompson and colleagues' work, individuals continued making changes to their decisions after the aggregate agreement had been reached. Thus, in Study 2 the time to final decision was measured directly by noting the point at which individuals made no further adjustments to their ratings, with a mean time of 128 of the total 180 seconds that did not significantly differ across all five conditions, even those containing a poor stage entrance or a performance error at the 100-second mark. No correlations were found between this and time to first decision, individual differences in quality ratings across the performance, or familiarity with the composition.

This thesis also demonstrated that the final continuous rating is indicative of the overall written score, consistent with previous work (Thompson et al., 2007). No significant difference was found between the final and written ratings across any conditions in Studies 1 and 2, even those in which the presence of a major performance error caused a significant change within the performance. While this could be interpreted simply as a recency effect, reactions to the errors in these studies support what Thompson and colleagues posited may be an "evolving process of preference formation" (p. 27; discussed further in Chapter 3), although it is in this case a complex one. An error part-way through a performance was strong enough to trigger a temporary reaction but not enough to leave a lasting impression; when the error was moved to the opening seconds (Study 1) or juxtaposed with a negative facial reaction (Study 2) it altered the rating of everything that followed. The former result is particularly salient when considering the evaluative process. It highlights that identical musical material can trigger a different reaction depending on its temporal location within the act of conducting a performance evaluation.

8.2.2 RQ2: The repertoire

Study 1 in the third chapter provided a direct examination of the repertoire's nature and its effect on the evaluative process by varying its familiarity, likeability, and length. Works of varying familiarity and likeability had no significant effect on the time taken to form a first decision when they were taken from the relatively familiar canon of Chopin, but when the work was entirely unknown and in an unpredictable, atonal style, the time to form a first decision doubled and the time to reach consensus on the final decision trebled in comparison to the most unfamiliar Chopin work. In Study 2, the degree to which participants liked and were familiar with the Chopin piece being evaluated showed no correlation with the ratings or the temporal processes forming them. The survey studies (3 and 4) also touched on the nature of the repertoire in that concertgoers were asked the degree to which they liked and were familiar with the works being performed, and in the case of Study 4, the composer's repertoire in general. Here, no meaningful relationship (i.e. correlations, if significant, were $\tau < .20$) was found between familiarity with the repertoire and either reported performance quality or enjoyment of the performances, supporting the above finding. However, strong significant correlations were noted between likeability of the work and performance quality regardless of musical experience, contradicting Studies 1 and 2 but supportive of previous findings linking work likeability with performance enjoyment (Thompson, 2007) and extending them to reported performance quality. As the finding is correlational, one could hypothesise that listeners struggled to separate the concepts of the quality of the performance and the degree to which they liked the composition, or that those who felt the musicians gave a stronger performance were more likely to reflect favourably upon it.

Overall, familiarity was found to have a negligible influence on quality ratings across studies except in the outlying case of the unfamiliar and atonal Caprice. That unfamiliarity with the work or composer of the relatively unknown but tonally stable

Chopin Tarantelle (Study 1) or the Whitacre works (Studies 3 and 4) had no effect suggests that it may be a process of orienting to an unpredictable musical language, although further study will be required to generalise the effect. Likeability showed a correlational relationship with quality ratings in Studies 3 and 4. As discussed above, length of the repertoire was also suggested to affect the evaluative process in that the short length of the 30-second prelude in Study 1 prompted a time to first decision significantly shorter than the benchmark set by other 2-3-minute works.

8.2.3 RQ3: The performer

The role of the performer's extra-musical behaviour on performance evaluation was figuratively and literally showcased in Study 2 with the use of an audio-video presentation highlighting both the pianist's stage entrance and his facial expression when committing an error. As discussed in Section 8.2.1, both affected the evaluation. The inappropriate stage entrance, modelled after the features identified in previous research (Platz & Kopiez, 2013), resulted in a shorter time to first decision as well as a lower initial judgement by musicians when compared with that of the non-musicians. However, that this deficit disappeared within the first 30 seconds of the performance suggests that the effect of such extraneous visual information can be overwritten by a convincing performance, given enough time. This contextualises previous research that has suggested a strong influence of the performer's behaviour and appearance on excerpts of short length (Wapnick et al., 2009; Tsay, 2013, 2014), highlighting that a performance should be evaluated in its full context to understand the complete effect of temporally specific performer behaviours.

Study 2 also demonstrated that such behaviours can provide a lasting impression in the right context. When the performance error was accompanied by a negative facial reaction, it triggered the dramatic drop in performance rating by both musicians and non-musicians that persisted to the end of the performance. The role of *facial overgeneralisation* was suggested as a cause, in which the strong social effect of facial interpretation may have induced the raters to overinterpret the severity of the error and the degree to which it reflected upon the performer's ability (see Chapter 4 for discussion). That the control video featuring the negative facial reaction without an accompanying error had no effect demonstrated that it was not the performer's behaviour alone that was penalised; rather, it altered the way in which the musical material was interpreted, further highlighting the complex interrelationship between musical and extra-musical variables in performance assessment.

8.2.4 RQ4: The environment

Study 4 focussed on environmental factors relating to a choral performance in a live setting, asking concertgoers to report factors relating to the physical environment of the performance venue and the social environment relating to surrounding audience members. Regarding the physical environment, audience members' seat location (measured by row number or general section) showed no correlation with their enjoyment or perceived quality of the performance despite a wide variance in distance from the stage and a medium correlation with the perceived quality of their seat. However, regression analyses showed that the perceived acoustic quality and appropriateness of the venue were significant predictors of perceived quality of the performance. As the study was correlational one cannot assume a causal effect, although the variable nature of the acoustic in the stone cathedral leaves the possibility for concertgoers having varied aural experiences and adjusting their ratings accordingly.

Regarding the social environment, raters were asked to hypothesise how their immediate neighbours and the concertgoers as a whole would have judged the performance's quality. Raters tended to assume their own ratings were significantly higher than their peers by approximately one third of one point, thus underestimating the true evaluations of their fellow concertgoers. The degree to which the evaluators were directly affected by their peers' reactions to the performance could not be inferred from the data due to the complex and uncontrolled nature of studying physical, situational, and social factors in live settings. Thus, the *Evaluation Simulator* was described in Chapter 7 as a new approach in taking forward this line of research.

8.2.5 RQ5: The evaluator

Each of the four studies examined qualities of the people conducting the evaluations (i.e. the evaluator) and their relation to their assessments. Study 2 provided an explicit examination of the evaluators' musical abilities by comparing a sample of experienced musicians with non-musicians in their evaluations of the stage entrance, performance error, and corresponding facial reaction. While the experienced musicians differed from those without musical training in that they showed a brief but temporary negative reaction to the stage entrance and the performance error without any lasting effect, the lack of difference in the process or outcomes of the ratings between the two groups in every other feature was striking. Correlation analyses in Studies 1, 3, and 4 also found no relationship between musical experience and the ratings given.

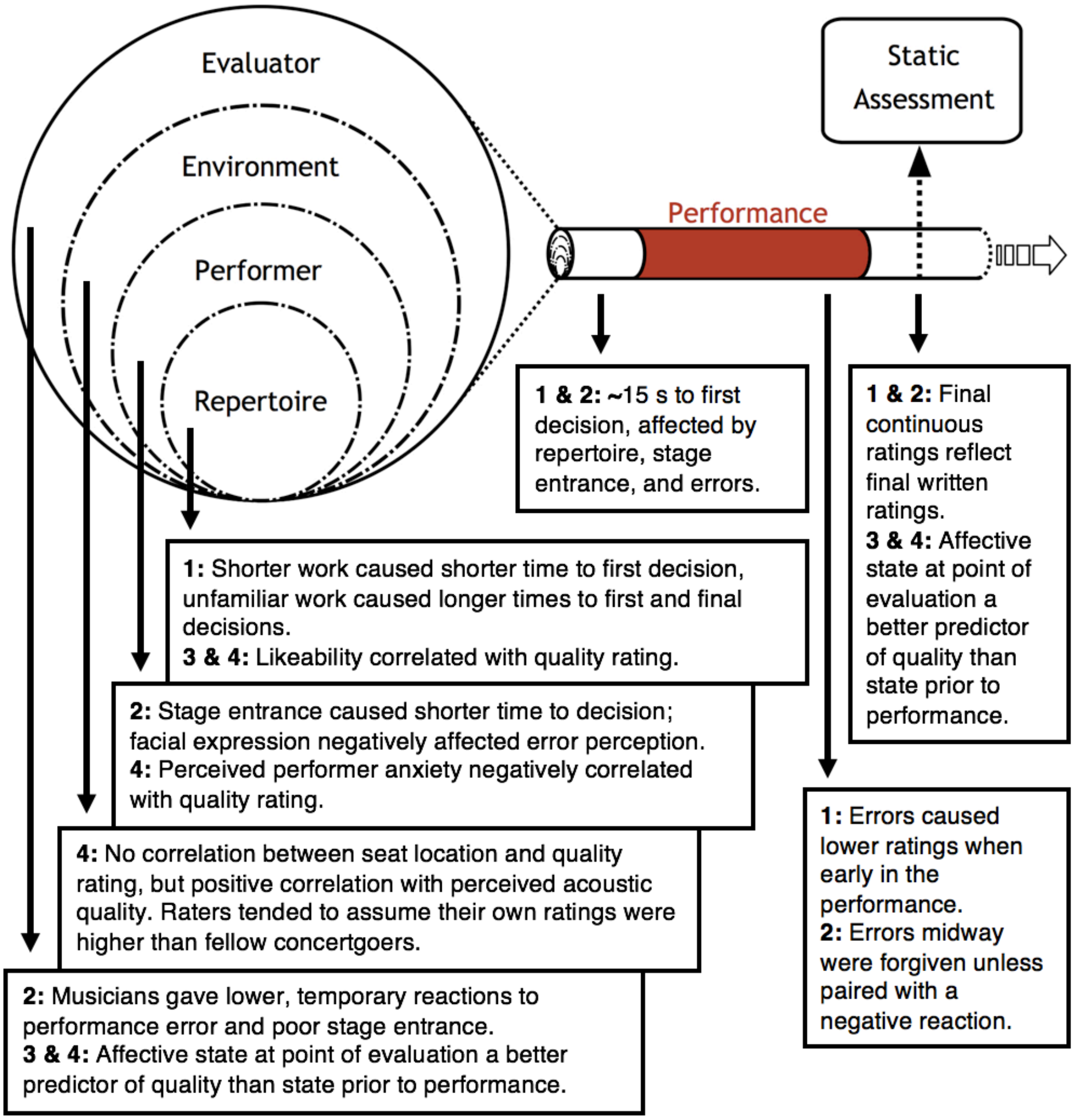
Studies 3 and 4 considered the mental and affective state of the evaluator before the performance and at the point of completing the evaluation, finding that self-reported mood, but not arousal, at the point of completing the evaluation after the performance was a better indicator of performance quality ratings than that at the outset of the performance or reported changes throughout. 'Anticipation' for the concert in Study 4 was also not predictive of quality rating or enjoyment, suggesting that the state of the rater at the point of completing the evaluation was more predictive than that they brought into the situation.

Studies 3 and 4 also examined the degree to which a preference response for the performance, indicated by the degree to which the evaluator enjoyed the performance, affected the outcome. Both studies found strong correlations (τs = .70 - .80) indicating an intrinsic link between the two constructs while maintaining some independence between them and in line with previous research (Thompson, 2007).

8.2.6 Summary

Looking across the four empirical studies, causal effects were found relating the evaluation process (RQ1), repertoire (RQ2), performer (RQ3), and evaluator (RQ5), and correlational relationships found relating to the evaluative environment

(RQ4). Figure 8.1 summarises the main findings with respect to the process model posited in Chapter 1.



1 & 2: ~15 s to first decision, affected by repertoire, stage entrance, and errors.

1 & 2: Final continuous ratings reflect final written ratings.
3 & 4: Affective state at point of evaluation a better predictor of quality than state prior to performance.

1: Shorter work caused shorter time to first decision, unfamiliar work caused longer times to first and final decisions.
3 & 4: Likeability correlated with quality rating.

2: Stage entrance caused shorter time to decision; facial expression negatively affected error perception.
4: Perceived performer anxiety negatively correlated with quality rating.

4: No correlation between seat location and quality rating, but positive correlation with perceived acoustic quality. Raters tended to assume their own ratings were higher than fellow concertgoers.

1: Errors caused lower ratings when early in the performance.
2: Errors midway were forgiven unless paired with a negative reaction.

2: Musicians gave lower, temporary reactions to performance error and poor stage entrance.
3 & 4: Affective state at point of evaluation a better predictor of quality than state prior to performance.

Figure 8.1. Main findings of the four empirical studies 1, 2, 3, and 4 with respect to the process model of music performance evaluation. Each finding results from a significant statistical test with effect sizes meeting the minimum accepted standard for a small or larger effect appropriate for the particular test.

8.3 IMPLICATIONS FOR PRACTICE

The results of this thesis offer a number of implications for musical practice across a range of activities. In Chapter 1, four roles of assessment were described; placement, summative, diagnostic, and formative (Goolsby, 1999). Implications for the present research are now presented in this order, beginning with placement evaluations done for their own sake (i.e. the music competition) followed by those done for organisational reasons (i.e. auditions) and then by implications for pedagogy and teaching, for audiences, for performers, and for domains beyond music.

8.3.1 Reconsidering music competitions

Music competitions represent an epitomal example of placement evaluation in a rarefied public setting, where the public assessment of the performers is as much a defining feature of the event as the performances themselves. This focus puts the objectivity and validity of the evaluative process and resulting decisions, which have been central to this thesis, to their greatest test. Sociological examinations of the music competition have highlighted their central and ever-growing role in the career trajectories of aspiring performers, (McCormick, 2008, 2009, 2015) in addition to tools to foster international interaction and national pride (McCormick, 2014). Other investigations have gone so far as to question the need for their very existence. In 1981-2, the European String Teachers Association (ESTA) hosted a debate and public discussion on the nature and roles of music competitions, leading to the establishment of a working party of expert music professionals comprising professors, teachers, producers, administrators, and critics, the majority of whom had direct experience evaluating musicians. In 1984 they published their report. It is worth considering the particular language they used to end the document, which is notable for both its strength and scope:

> Competitions are closely identified with some of the principal threats – in particular, the ‘star’ system and the exploitation of young musicians – and, until such time as they fade from the scene, they are best confined to the outer reaches of the profession where their influence may be negligible (ESTA, 1984; p. 27).

This call for the end of the music competition was based on several key observations. First, that competitions can cause great stress and ill-health for the performers involved, particularly in their developmental and early career stages, resulting from heightened expectations placed upon the winners and a loss of motivation and opportunity for the many more who were unsuccessful. While a full discussion of this is beyond the scope of the findings of the present thesis, considerable research supports the first point that musicians' health and wellbeing reveal an epidemic of stress, burnout, and injury exacerbated by heightened expectations, a competitively charged atmosphere, and a lack of appropriate coping mechanisms (McPherson & McCormick, 2000; Atlas et al., 2004; Spahn et al., 2004; Williamon & Thompson, 2006; Clark & Lisboa, 2013; Clark et al., 2014; Araújo et al., 2017; Bonneville-Roussy et al., 2017b).

Second, and relevant to this thesis, was the working group's key observation that objectively proclaiming differences in performance quality at top levels, among subjective artistic interpretations, and often between divergent repertoire and instrument families, can become meaningless. In their words, "If…we wish to judge relative grandeur, no form of measurement is conceivable since too many intangible qualities are involved. In musical performance, the only measurable attributes are aesthetically insignificant" (p. 17). The findings of this thesis support the notion that extra-musical factors relating to the repertoire, performer, environment, and evaluator have causal effects on music assessment, which add weight to the ever-expanding literature questioning the subjectivity of expert judgement (Williamon & Thompson, 2003; McPherson & Schubert, 2004; see Chapter 1 for a full review). Where competitions allow for varied repertoire, for example, can it be assumed that the judges are considering each performance using the same cognitive process, or will the differences found in Study 1 manifest as a different assessment processes across competitors?

This issue of performance discrimination leads to the third issue raised in the ESTA report that "because competitions have to produce a winner, even when there is no outstanding performer amongst many good ones, an arbitrary choice has to be made, thus creating an apparent rarity out of what is, in fact, an abundance" (p. 20). They describe how this harms the musical community by establishing and propagating a 'star' mentality in audiences in which a great deal of attention is given to individual winners while thousands of aspiring performers are offered limited opportunities of this type. Research has demonstrated evaluators' tendency to assume and invent differences in identical performances (Duerksen, 1972; Anglada-Tort & Müllensiefen, 2017) or struggle to differentiate between conductors (Madsen et al., 2007, 2009) or instrumental players (Mitchell & MacDonald, 2012, 2016) without paired visual information. Study 2 of this thesis demonstrated how a slight change in facial expression significantly altered the assessment of the pianist's performance despite an acoustically identical outcome and the relative experience of the musical evaluators. In a close competition this could have had a drastic effect on the outcome.

Of course, music competitions are not without benefits to the performer, including experience, a motivator to excel, contact with fellow performers, and feedback from other teachers and professionals (ESTA, 1984; McCormick, 2015). However, as the ESTA report suggested, these are all benefits that could be had through public showcases and festivals. Thus, competition organisers could remove the final placement evaluation and distribute prizes among a group of excellent performances, selected through privately-held auditions kept out of the public spotlight, and "above all, television coverage of competitions should be avoided" (p. 28). The continued proliferation of regional, national, international, and televised public competitions (McCormick, 2008, 2015) suggests that the recommendations of the ESTA report were never heeded.

Based on the uncertainty surrounding music performance evaluation highlighted throughout this thesis, the most drastic form of evaluative improvement could perhaps be doing away with the evaluation altogether when its value is not clear. In many cases, however, assessment is necessary. If the ESTA recommendations are to be followed, for example, private auditions are still required to select those to be showcased in public platforms where minimum standards are to be maintained and where practicalities prevent giving equal time to hundreds of applicants. What is needed is a reliable and scientifically-informed audition process.

8.3.2 Improving auditions

Three factors can be examined in improving the audition; the judges' training, the assessment process, and the procedures by which the audition occurs. In training the next generation of expert evaluators, the literature discussed in Chapter 7 highlighted the lack of explicit training 'expert' music evaluators receive and a demand for increased opportunities to practise the act of performance assessment. The *Evaluation Simulator* presented there provides one opportunity to engage further the processes of experiential and self-regulated learning and provide the next step in the evolution of assessment training as it grows in prominence in conservatoires (Hunter & Russ, 1996; Searby & Ewers, 1997; Bergee 1993, 1997; Bergee & Cecconi-Roberts, 2002; Daniel, 2004; Blom & Pool, 2004; Lebler, 2007; Latukefu, 2010; Hanken, 2016; Mitchel & Benedict, 2017; Dotger et al., 2018), although still primarily as a method to improve musical performance. McCormick (2015) identified the stigma among regulator competitions evaluators of the 'professional judges', or that small subset of jurors who sit upon a disproportionately large number of juries but are not perceived to have the performance careers to 'earn' them that position. Here, experience doing the task at hand is considered less salient than the tangential but elevated skill of performance. By treating evaluation as a skill to be trained with equal importance to performance itself, this stigma may be relaxed and the value of the 'professional judge' recognised. So long as performance ability is presumed to be the only prerequisite to effective evaluation, the commonly assumed abilities of the expert judge, as described by Thompson and Williamon (2003), will continue to stand in the way of proper scrutiny.

For those musicians already in their careers, one can turn to the forms of professional development and workshops used across organisational practices. Sessions educating evaluators on the biases of the repertoire, performer, environment, and evaluator discussed in Chapter 1 and demonstrated across the four empirical studies in this thesis would suit this function. However, it is not enough simply to educate judges of the effects of stereotyping and implicit bias; recent research has warned of a ‘backfire’ effect where simply stating the problem exists can actually increase prevalence of the issue by creating a social norm for the practice (Duguid & Thomas-Hunt, 2014). Instead, Duguid and Thomas-Hunt (2014) suggested that education programmes should focus on explicit practices that can be adopted in their decision-making processes and how their colleagues (or competitors) are already implementing them to improve outcomes.

Thus, new methods to improve the assessment process within auditions may be of value. The continuous measures methodologies used in Studies 1 and 2 can provide such an opportunity and may be particularly suited to combat the implicit biases endemic to performance evaluations. Initial negative impressions provoked by prior knowledge can be mediated when a specific outcome is asked to be measured (Neuberg, 1988). This may account for the lack of lasting effect in the stage entrance discussed in Study 2 or the gradual increase of ratings following the initial performance error in Study 1, where a true continuous peak-recency effect would have predicted a partial recovery to a stable plateau below but parallel to those who rated the error-free condition. As the listeners in each case were given instructions to focus on the quality of the performance and use the continuous measures to bring that task squarely to mind in the moment-to-moment evaluations, they may have been encouraged to focus on the higher quality of the performance at that given time. From another perspective, Studies 3 and 4 showed that mood and relaxation following the performance and at the time of a single evaluation was more predictive of the scores than mood at the outset of the performance or changes resulting from the listening, suggesting susceptibility to affective state when making an intuitive decision. Research has also found that delaying the point of decision-making can increase decision accuracy in simple stimulus tasks (Teichert et al., 2014). However, too much undirected introspection in choosing consumer products can lead to less agreement with experts’ assessments (Wilson et al., 1991) and lowered satisfaction with the decision (Wilson & Schooler, 1993). This reinforces the notion that good decisions should be made based on salient criteria, and reflection should be guided by a fixed process. Continuous measurement methodologies may again offer this opportunity.

An instructive parallel to this approach of providing a systematic, guided process of evaluation, as well as guidance on what should constitute the aforementioned ‘salient criteria’, can be found in Daniel Kahneman’s (2011) early work with the Israeli Army in developing their process of selection for officer training. The original method was to have recruits perform a ‘leaderless challenge’ in which teams of eight had to move a large log over a wall without it touching. Judges would observe and single out recruits who demonstrated leadership as predictive of success as future officers. The predictive value of the test was extremely poor, although use of the assessment continued. A similar phenomenon has been found in music, where Wolf and Kopiez (2014) found that entrance theory tests in a German music conservatoire served as poor predictors for final grades three years later. Kahneman’s solution was to have judges determine five key traits they felt were needed in the recruits, and then conduct structured interviews focussing on experiences and hypothetical situations targeting those traits. Following the interviews, not only were scores on the five items stronger predictors of success, but overall scores provided after the process were themselves better predictors than the original test. Thus, the specific criteria used in musical evaluation may not be so relevant as the process of determining and using them before forming a holistic judgement, forcing the evaluator to slow down thinking and avoid implicit bias. As a standardised set of segmented scores has not been identified despite decades of research and centuries of musical practice (e.g. Gutsch, 1964; Schmalstieg, 1972; Mills, 1991; Zdzinski & Barnes, 2002; Thompson & Williamon, 2003; Wesolowski, 2016, 2017; see Section 2.2) allowing judges to form their own criteria to then apply to a crucial overall score, perhaps aided by the use of a continuous measures methodology, could be the more productive approach.

Finally, in considering the procedures surrounding auditioning, several points can be addressed. The serial order effects discussed in Chapter 1 must be considered. Ideally, the same randomisation procedure as used in Study 1 and across most experimental research in the field would be employed in which the order is counterbalanced between judges. However, this is only possible when multiple judges separately evaluate recordings, thus the process is not suitable for live auditions or situations with only one judge. Here, randomising performance order, as is common in many competition and audition settings, does not solve the problem, and in the case of competitions there can be pressure to put the higher-ranked performers later in final rounds to make for a more compelling viewing experience (Bruine de Bruin, 2006). These factors again make a strong case for audition-via-recording and the use of separate panels.

Removing information relating to the performer's history, demographics, and nationality is also encouraged. This not only affects the process of serial judgement, but the degree to which judgements affect subsequent decisions. In studying gymnasts, Damisch and colleagues (2006) found that by informing participants that two competitors were of the same nationality they were more likely to adjust evaluations of the second-viewed performance to the first than if the gymnasts were not perceived to be from the same country. One must also consider use of the blind audition. Since its introduction in the 1970s it has been linked to a rebalancing of such biases, including a marked increase in the hiring of female performers (Goldin & Rouse, 2000). While one of course loses the added expressiveness that the visual element of performance can bring (see Section 1.6), this would also eliminate such effects of performance behaviour as the negative facial expressions examined in Study 3.

In considering how many judges are necessary, Bergee (2007) suggested that 17 hypothetical raters could overcome the measurement error demonstrated using Rasch modelling and reach a benchmark reliability index of .80. While this may not be a practical amount, it seems true that more judges can counteract individual bias. The data in Studies 1 and 2 demonstrated how, even with the large degree of inter-rater variance as found in Thompson and colleagues' (2007) study of continuous measures, the aggregate continuous rating showed very stable agreement across the error-free performances (and in the case of the non-musicians rating the aural-only error in Study 1, in performances containing errors as well).

After judges have made their decisions, should they be publicised? The studies of the Queen Elisabeth Music competition demonstrate how this opens the competition to scrutiny and criticism, while providing a rich data source for research (Flôres &

Ginsburgh, 1996; Glejser & Heyndels, 2001). A further step would be to identify the individual judges with their scores, as is done in professional figure skating, where a temporary removal of this practice in the hope of reducing corruption (i.e. removing influential parties' ability to check whether their puppet judge voted 'correctly') increased nationalistically-favoured voting and suspected vote-trading (Zitzewitz, 2014).

Summarising the research discussed across this thesis, the ideal audition procedure would comprise a panel of as many judges as possible (although more than 17 may be unnecessary), each independently and in a counterbalanced order reviewing audio-video recordings collected with the same equipment in the same venue without any influence from a confounding audience and no prior information about the performer. They would mutually agree upon a short set of basic criteria for evaluation, considering them using a continuous response methodology before settling upon a final, holistic written score to be used for the final comparison. Those scores might then be aggregated using one of the several statistical procedures being commissioned and adapted by numerous international competitions to remove extreme scores (McCormick, 2015); a simple mean might also be collected. The degree to which this approach is possible will be up to the individual institutions, especially in countering what Kahneman (2011) termed the 'Illusion of Validity' in people's overreliance on evaluative methods they feel should be accurate, but it cannot be said that guidance for improvement is not available.

8.3.3 Guiding educators and performers

The implications for improved training for audition jurors also applies to educators and their students. Educators trained in the art and science of assessment will be better suited to diagnose and help their students, and students will be better equipped to take on portfolio careers incorporating teaching and evaluation. Furthermore, continuous measures methodologies can provide new avenues for teachers to communicate performance feedback to their students. The RCM software used in Study 1 and bespoke software created for Study 2 could be adapted and used to provide real-time, intuitive feedback to students in masterclasses and lessons, as in other domains such as collecting information regarding the efficacy of participants in American presidential debates (Kirk & Schill, 2014). The use of Immersive Virtual Environments and distributed simulations described in Chapter 7 also provide opportunities for new forms of training. The value of self-assessment as a teaching tool across educational domains is well documented (Ross, 2006), so providing opportunities for students to assess themselves and others should be placed at the core of teaching provision in music and beyond.

Performers can take very specific advice from the findings of this thesis.

- Walk on stage with confidence.
- Do not pull faces when you make a mistake and you may be judged as if they never happened.
- Consider the familiarity of your repertoire to your audience or judging panel, especially if it is tonally or structurally unfamiliar.
- While mistakes are inevitable, do everything you can to avoid them in the opening moments of your performance.
- First impressions count.
- Concert audiences are more likely to base their impressions on their mood after the concert, not before.
- Your audience's distance from the stage may not correlate with their enjoyment or performance quality rating of your performance, but the acoustic quality of the venue might.
- No audience member is processing the musical content of your performance in isolation; their impressions are being informed, shaped, and co-opted by what they think, know, see, and hear.

In general, performers might benefit from increased awareness of their audience's cognitive processes and states, lest inflated or misplaced perceptions of the crowd's attention and expectations lead to unnecessary worry. It has been shown that the presence of an audience may result in an improved performance, possibly driven by the effects of social facilitation (Shoda & Adachi, 2014). Research in sport, for example, has found that the celebrated ‘home’ advantage can reverse in critical competition settings, where the increased pressure felt from the inferred expectations of a supportive crowd can cause increased arousal leading to impaired performance (Voyer et al., 2006).

8.3.4 Understanding audiences

The research in this thesis can also inform how performers and organisations accommodate their audiences. As Studies 3 and 4 linked concert enjoyment and perception of quality to mood at the time of assessment, venues can focus on maintaining the concertgoer’s experience following the performance to maximise the recency effect demonstrated in behavioural psychology (Redelmeir & Kahneman, 1996; Kahneman et al., 1993) and ensure that any memories are not tainted by poor experiences while leaving the hall. Organisers and ticket sellers may be encouraged by the lack of correlation between seat location and evaluation or enjoyment, although it may also call into the question the premium paid for seats at the front. Alternatively, the satisfaction of a bargain might be countering the enjoyment of a privileged seat.

The findings of Study 1 highlight the importance of familiarity with unconventional repertoire, while also stressing the potential of audiences to acclimatise within a performance. While the Caprice received initially late and low performance quality ratings that took longer to reach their final outcome than the comparable Chopin work, the final result was descriptively higher than any other performance in the study. As providing audiences with structural programme notes has been found to be counterproductive (Margulis, 2010; Bennett & Ginsborg, 2018), it may be the case that those programming concert material must simply trust their audience to discover and understand the work at it progresses, and that they simply need to be given time to come to their own conclusions. Pitts (2016) described the complexity of personal and practical factors driving audience’s engagement with new artistic experiences, highlighting the great deal of variability between concertgoers in why they choose to attend. The reasons cited went far beyond a search for performance quality and repertoire familiarity. Study 2 also highlighted the savviness of audiences with little-to-no musical training in the degree to which their responses mirrored those of the experienced musicians.

The method of continuous measures may again provide an avenue to better understand the real-time reactions of audience members. The relative simplicity for users offers possibilities for the mass collection of continuous data through mobile devices, requiring a browser-based platform in line with popular data collection platforms (e.g. Kahoot.com; Surveymonkey.com; Mentimeter.com). The use of tools such as the *Evaluation Simulator* might also provide a new method with which to interact with audiences. The employment of the simulator at the Cheltenham Music Festival (see Section 7.6.6) allowed the researchers and members of the public to discuss the processes and pitfalls of choosing the artists they eventually hear using a shared experience via the simulator. An audience better informed of the inner workings of musical selection might reconsider the ways in which they choose to patronise a particular concert or purchase a recording.

8.3.5 Looking beyond music

As theories and practices in other performance domains have been drawn upon to inform the findings of this thesis and their implications for music practice, so too can the current findings apply to domains beyond music. As stated at the outset of this thesis, music performance evaluation is ultimately a lens through which the larger issues of decision-making, assessment, skills training, and expertise can be considered in a socially and aesthetically complex domain. Other performative artistic acts form a clear first step, especially those that rely on the evaluation of a temporal stream of information such as dance, theatre, film, and acting. One can then move to other areas where human performance is assessed by their peers. The sports literature provides significant insight into the role of assessment procedures and the effects of audience response. While the role of the competition may serve a more utilitarian function in that atmosphere, the implications for auditions described above could apply to the panel assessing the aesthetics of a dive, gymnastic routine, or figure skating performance; approximately one-third of Olympic sports rely heavily or entirely on human assessment (Stefani, 1988). In return, musicians and those studying within the domain would benefit from examining carefully the rigour with which those in sport treat the quantification of their intended outcomes, or combine automatic quantification and human judgement (for example, the addition of computerised ball tracking in tennis and baseball to indicate an 'in' or a 'strike', respectively, which was once solely the domain of the umpire). While a risk is carried of overquantifying music, there are scenarios wherein following a written score can lead to objectively correct and incorrect outcomes. Automated systems could be used to consistently measure these outcomes, reducing cognitive load on and risk of bias by the human judge who might then be better able to focus on the more intangible, subjective qualities of the interpretation.

Businesses too are heavily dependent on human assessment, from the job interviews from which the original continuous measures research sprang (Springbett, 1958) to meeting presentations to investor pitches. Each is a performance relying on a balance of content, presentation, and assessment, and each has a counterpart judge who is expected to make good predictive decisions for the benefit of their organisation and their own job security. Reducing bias in selection also offers the opportunity to increase workplace diversity, which can increase productivity (Saxena, 2014), although it requires extra effort managing intra-office relations and personal clashes that can, if poorly handled, lead to lowered output (Ellison & Mullin, 2014). Of course, the worlds of business and music are closely intertwined, and musical organisations such as orchestras and schools should be mindful of the benefits such diversity can bring to their ensemble as well as their administration. They should also ensure that their approach to filling the concertmaster's seat is as carefully considered as the one used to choose their Director, and that similar issues can arise in both scenarios.

The medical domain provides a more lateral but still salient opportunity for applications of evaluative theory and practice. Healthcare professionals must perform a variety of clinical and interpersonal tasks, many of which still rely on human assessment in development and outcome measurement. The American Board of Medical Specialties has included self-assessment and lifelong learning as one of its four components of maintaining clinical licensing through Continuing Medical

Education, and research has found poor self-assessment abilities in healthcare providers across international systems, as well as a lack of provision to help develop these skills (Gordon, 1991; Davis, 2006). While the medical profession has been a leader in the use of simulation and Immersive Virtual Environments for teaching clinical skills, as described in Chapter 7, the *Evaluation Simulator* described in this thesis offers a model for learning through assessment. A similar approach could be taken in medicine, where students judge and give feedback on the execution of medical skills. Practitioner-patient interactions, from taking patient histories to delivering bad news, could a be a particular area of focus due to the complex interaction of social and environmental factors similar to what is seen in music (Sustersic et al., 2018). In these situations, patients become the audiences and jurors for their practitioner's performance, and the degree to which they judge their clinician or carer to be competent, empathetic, and sufficiently motivating can significantly affect the efficiency and efficacy of the treatment they receive within the healthcare system (Edwards et al., 2018; Santana et al., 2018). A recent report estimates that poor interaction costs the UK healthcare system over one billion pounds per year in litigation, poor treatment compliance, and reduced mental health for both doctors and patients (McDonald, 2016). There is much to gain in improving these abilities through better training via self-assessment and understanding what the patient expects from a medical performance.

Musicians, then, might learn from the high degree of risk medical practitioners find themselves under and for which they prepare themselves and their procedures. For musicians, a poor outcome might be a lacklustre performance. For those in healthcare, a bad decision or placement can lead to pain, trauma, and death. To combat this, the medical profession has embraced the checklist in ensuring that the confidence brought by expertise does not diminish their attention to or memory of the basic tenets of assessing a patient or critiquing the surgery of another (Clay-Williams & Colligan, 2015). The basic rubrics of the segmented music assessment (see Chapter 2) provide a rudimentary example of this, though more could be done in considering the true range of qualities to be considered, not to mention the order or points in time at which they might be the most salient. Such an approach would be strongly compatible with the principles of Immersive Virtual Environments discussed in Chapter 7 and use of the *Evaluation Simulator*. A series of corresponding checklists to its various use cases would allow musicians to practice the process of evaluation, and ensure that the heightened environment does not cause them to miss out on the fundamentals of the performance they are assessing.

8.4 LIMITATIONS OF THIS WORK

A key aim of this thesis was to engage with evaluative decision-making in musical performance as it related to the actual acts that occur as part of daily practice, acknowledging all of the complexity of the musical material and the extra-musical factors that influence it. For this reason, the topic was approached from both sides of the spectrum: the highly-controlled, laboratory-based experimental studies employed in Chapters 3 and 4, and the in-situ, minimally-invasive survey studies of concertgoers in genuine performance settings in Chapters 5 and 6. Each of these approaches bring with them specific limitations to the interpretation of the material based on issues of ecological validity and control of extraneous variables, the specifics of which are discussed in their respective chapters. Chapter 7 also engaged with this disparity of approaches directly by proposing a new methodological approach that bridges the gap between these two extremes and provides a new way forward in evaluative research.

Nonetheless, four methodological limitations that apply across the thesis must be acknowledged. First and foremost, the 880 participants across the four empirical studies represent convenience samples comprising a range of musical and evaluative experience. While this scope provides a wide range of insight into the evaluative act, it must be said that no attempt was made to target the specific population of 'professional' evaluators with significant experience in serving on audition panels, juries, exam panels, etc. As Chapter 4 demonstrated more similarities than differences in the rating processes between those with and without significant musical training, and as Chapter 7 highlighted that musical training remains the leading if not only qualifier for evaluative experience, conclusions may still be drawn about the evaluative process. However, a gap remains in the collective knowledge regarding differences in this elite group of musical evaluators, and whether experience alone can drive improvements in consistency and reliability.

Second, while the research questions and proposed process model (see Figure 8.1) targeted the broad categories of repertoire, performer, environment, and evaluator, sample properties from each condition had to be identified to allow for focussed examination. The relative weighting and interaction of each of these four categories can therefore not be determined from the present data, and thus the model can only serve as a categorisation tool until broader studies can be performed that take a wider set of criteria into consideration.

Third, while this thesis examined the process as well as the product of evaluation, it relied on self-reports via written and continuous ratings from which select indicators (e.g. time to first decision, time to final decision, relationship between time-sensitive variables and final decision) could be extracted. These are small windows into the complex cognitive processes underlying these evaluations, and while they are representative of the relatively reductive real-world act of converting a nuanced musical performance into a single, holistic score, much more remains to be learned of what drives evaluative decision-making.

Finally, this thesis has taken an almost exclusively post-positivist, quantitative approach to the generation of knowledge. As the aims were to determine generalisable, quantifiable measures of correlation and causality, the experimental and survey-based procedures used in the four empirical studies were deemed appropriate. However, these approaches do not give insight into the subjective experiences of the evaluators, or the degree to which the participants were consciously aware of the factors driving their decisions. It is generally recognised that the field of performance studies would benefit from a wider use of qualitative methodologies (Holmes & Holmes, 2013), and the subdiscipline of performance evaluation is no exception. The written open comments in Study 2 provided a hint here of what is possible, where it could be ascertained that several non-musicians were aware of the significant performance error (without negative facial reaction) despite not penalising it in the continuous ratings. Further work employing qualitative and mixed-method approaches (e.g. Davidson and

Coimbra, 2001; Kokotsaki et al., 2001; Bonshor, 2017) will provide a wider breadth of knowledge to the field.

8.5 AREAS FOR FUTURE RESEARCH

The findings of this work offer numerous avenues for further research. Where continuous methodologies were used to measure the effects of a select few factors in Studies 1 and 2, they could provide insight into every extra-musical feature previously studied through single ratings alone. Continuous approaches could also be extended to live-audience and panel settings to better understand rating processes, with opportunities to integrate data from new technologies used to collect physiological measurements of arousal in theatre productions (Wang et al., 2016) and art galleries (Tschacher et al., 2015). The ever-expanding field of neurological study could also be brought to the field, as it has in the domain of affective musical response as measured using electroencephalography (e.g. Khalfa et al., 2002; Palmer et al., 2009; Chapin et al., 2010). Such study could give greater insight into the continuous processes of performance evaluation, the examination of which was limited to self-report in the present thesis. In particular, the examination of temporally specific performance points (e.g. a stage entrance, a performance error) would provide a salient stimulus against which neurological perception and processing times could be examined through the measurement of EEG-based action potentials (e.g. Palmer et al., 2009), pupil dilation (e.g. Preuschoff et al., 2011; de Gee et al., 2014), and, in the case of visual stimuli, eye gaze tracking (Christoforou et al., 2015). Further qualitative work is also required to provide insight into the expectations, experiences, and subjective reasonings of not only the evaluators but the organisations who employ and train them, the performers who are judged by them, and the audiences whose own decisions are guided by those of the experts.

Regarding the *Evaluation Simulator* in Chapter 7, further research must begin with both quantitative and qualitative efficacy studies of the technology's ability to mimic the heightened pressures of real-world evaluation and serve as an effective training tool for future evaluators. However, current knowledge limits this research. While a great deal is now understood concerning performers' physiological arousal and anxiety responses in music (e.g. Kenny, 2011; Aufegger et al., 2017; Chanwimalueang et al., 2017), no research has examined the physiological experiences of expert evaluators. As has been the theme throughout this thesis, the skill of evaluation requires the same attention given to the skill of performance. Approaches balancing ecological validity with experimental control, such as the simulator, will allow the elusive factors comprising the social and physical environments of performance evaluation to be examined.

An exciting new path for the field of evaluation research is that of artificial intelligence and machine learning. While the prospect of the automatic evaluation of musical ability has interested musicians and researchers for decades (Welch, 1994), new data-centric approaches are bringing such concepts to reality. Computers have been long used to collect and analyse music performance evaluation data (Nakamura, 1987; Zdzinsky, 1991). Machine-learning techniques now offer the opportunity to train digital systems against holistic human judgements so that they may recognise and deconstruct the salient features of a ‘good’ performance, such as systems that can assess doctors’ skills with human-level accuracy (Gibbons et al., 2017). This could address the challenges discussed in Chapter 2 of developing standardised rubrics of performance assessment; perhaps the criteria comprising music performance quality, their interactions, and their relative weightings are so inherently complex that only advanced machine learning techniques may be able to unravel them. Researchers are already making great strides in systems to identify the markers of high expressivity and quality in music performance (Wu & Lerch, 2018), and progress on this front will undoubtedly move quickly.

Finally, quantification of the skills of the assessor and the efficacy of the assessment method must be pursued so that marked improvements in the delivery of competitions, auditions, and education can be tracked and best practices identified. Until then, musical practice will continue to suffer from a lack of standardised approaches and ambiguity in the goals and validity of the assessment practices employed within. Such work will not only have benefits for musical practice, but any domain that relies on human assessment to achieve its aims.

8.6 CONTRIBUTIONS TO KNOWLEDGE

This thesis has generated new insights into performance evaluation by examining not only evaluative products, but also the processes that lead to them. It has demonstrated the effects of a range of factors on the act of forming a music performance assessment. Using both experimental and naturalistic means, it has contributed to an ever-growing body of research that has questioned the validity and subjectivity of expert and amateur judgement.

Chapters 3 and 4 of this thesis expanded early use of continuous measures methodologies in the study of music performance evaluation (Thompson et al., 2007; Himonides, 2011) with the first known use of continuous measures methodologies to examine time-specific effects of experimental variables on the process of forming performance evaluations. This method provided novel insights into the role of repertoire length and features, stage entrance behaviour, and the location and nature of performance errors and accompanying facial reactions on evaluative processes. Through the experimental methods employed and, in particular, the novel data collection tool developed in Chapter 4, it also provides a template by which future research can examine this work using continuous approaches.

Chapters 5 and 6 collected data from large audiences at professional concerts to demonstrate in situ relationships between self-reported mood and anxiety states before and after performance with perceived quality and enjoyment of the music. These studies confirmed Thompson's (2007) correlations between familiarity, enjoyment, and perceived quality of the music in a live setting, and expanded knowledge of the relationship between social and physical environmental factors on audience responses. Chapter 7 then demonstrated the need for and novel application of Immersive Virtual Environments and distributed simulation in the *Evaluation Simulator*, opening new avenues for research and training in performance assessment.

This work sets out an agenda to reconceptualise quality evaluation as a skill equal in importance and complexity to the performance it seeks to capture. Only when the judge behind the desk is considered with the same attention given to the performer on stage will the processes and products of musical evaluation, and their implications for decision-making in general, be truly understood.

REFERENCES

Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, *21*(3), 246-255.

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.

Alessandri, E., Williamson, V. J., Eiholzer, H., & Williamon, A. (2015). Beethoven recordings reviewed: A systematic method for mapping the content of music performance criticism. *Frontiers in Psychology*, *6*(57), 1-14.

Álvarez-Morales, L., Zamarreño, T., Girón, S., & Galindo, M. (2014). A methodology for the study of the acoustic environment of Catholic cathedrals: Application to the Cathedral of Malaga. *Building and Environment*, *72*, 102-115.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431.

Anglada-Tort, M., & Müllensiefen, D. (2017). The Repeated Recording Illusion: The effects of extrinsic and individual difference factors on musical judgments. *Music Perception: An Interdisciplinary Journal*, *35*(1), 94-117.

Araújo, L. S., Wasley, D., Perkins, R., Atkins, L., Redding, E., Ginsborg, J., & Williamon, A. (2017). Fit to perform: An investigation of higher education music students' perceptions, attitudes, and behaviors toward health. *Frontiers in Psychology*, *8*(1558), 1-19.

Aronson, E., Wilson, T. D., Akert, R. M., & Fehr, B. F. (2007). *Social Psychology* (3rd Canadian Edition). Pearson Education Canada.

Arruda, J. E., Stern, R. A., Hooper, C. R., Wolfner, G. D., Somerville, J. A., & Bishop, D. S. (1996). Visual analogue mood scales to measure internal mood state in aphasic patients: Description and initial validity evidence with normal and neurologically impaired subjects. *Archives of Clinical Neuropsychology*, *5*(11), 364.

Asch, S. E. (1956). Studies of independence and conformity. A minority of one against a unanimous majority. *Psychological Monographs*, *70*, 1-70.

Atlas, G., D., Taggart, T., & Goodell, D., J. (2004). The effects of sensitivity to criticism on motivation and performance in music students. *British Journal of Music Education*, *21*(01), 81-87.

Aufegger, L., Perkins, R., Wasley, D., & Williamon, A. (2017). Musicians' perceptions and experiences of using simulation training to develop performance skills. *Psychology of Music*, *45*(3), 417-431.

Ballantyne, J., Ballantyne, R., & Packer, J. (2014). Designing and managing music festival experiences to enhance attendees' psychological and social benefits. *Musicae Scientiae*, *18*(1), 65-83.

Balteș, F. R., & Miu, A. C. (2014). Emotions during live music performance: Links with individual differences in empathy, visual imagery, and mood. *Psychomusicology: Music, Mind, and Brain*, *24*(1), 58-65.

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. W. H. Freeman & Co: New York.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182.

Barratt, E., & Moore, H. (2005). Researching group assessment: Jazz in the conservatoire. *British Journal of Music Education*, *22*(03), 299-314.

Behne, K.-E., & Wollner, C. (2011). Seeing or hearing the pianists? A synopsis of an early audiovisual perception experiment and a replication. *Musicae Scientiae*, *15*(3), 324-342.

Bennett, D. (2008). A gendered study of the working patterns of classical musicians: Implications for practice. *International Journal of Music Education*, *26*(1), 89-100.

Bennett, D., & Ginsborg, J. (2018). Audience reactions to the program notes of unfamiliar music. *Psychology of Music*, *46*(4), 588-605.

Bergee, M. J. (1993). A comparison of faculty, peer, and self-evaluation of applied brass jury performances. *Journal of Research in Music Education*, *41*(1), 19-27.

Bergee, M. J. (1997). Relationships among faculty, peer, and self-evaluations of applied performances. *Journal of Research in Music Education*, *45*(4), 601-612.

Bergee, M. J. (2006). Validation of a model of extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education*, *54*(3), 244-256.

Bergee, M. J. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education*, *55*(4), 344-358.

Bergee, M. J. (2015). A theoretical structure of high school concert band performance. *Journal of Research in Music Education*, *63*(2), 145-161.

Bergee, M. J., & Cecconi-Roberts, L. (2002). Effects of small-group peer interaction on self-evaluation of music performance. *Journal of Research in Music Education*, *50*(3), 256-268.

Bergee, M. J., & McWhirter, J. L. (2005). Selected influences on solo and small-ensemble festival ratings: Replication and extension. *Journal of Research in Music Education*, *53*(2), 177-190.

Bergee, M. J., & Westfall, C. R. (2005). Stability of a model explaining selected extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education*, *53*(4), 358-374.

Berlo, D. K., Lemert, J. B., & Mertz, R. J. (1969). Dimensions for evaluating the acceptability of message sources. *The Public Opinion Quarterly*, *33*(4), 563-576.

Bishop, L., & Goebl, W. (2014). Context-specific effects of musical expertise on audiovisual integration. *Frontiers in Psychology*, *5*(1123), 1-14.

Bissonnette, J., Dubé, F., Provencher, M. D., & Moreno Sala, M. T. (2015). Virtual reality exposure training for musicians: Its effect on performance anxiety and quality. *Medical Problems of Performing Artists*, *30*(3), 169-177.

Bissonnette, J., Dubé, F., Provencher, M. D., & Sala, M. T. M. (2011). The effect of virtual training on music performance anxiety. In A. Williamon, D. Edwards, & L. Bartel (Eds.), *Proceedings of the International Symposium on Performance Science 2011* (pp. 585-590), Ultrecht, The Netherlands: European Association of Conservatoires.

Bissonnette, J., Dubé, F., Provencher, M. D., & Moreno Sala, M. T. (2016). Evolution of music performance anxiety and quality of performance during virtual reality exposure training. *Virtual Reality*, *20*(1), 71-81.

Blanchette, I., & Richards, A. (2010). The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning. In J. D. Houwer & D. Hermans (Eds.), *Cognition & Emotion: Reviews of Current Research and Theories* (pp. 561-595). Taylor & Francis.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002a). Immersive Virtual Environment technology as a methodological tool for social psychology. *Psychological Inquiry*, *13*(2), 103-124.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002b). Immersive Virtual Environment technology: Just another methodological tool for social psychology? *Psychological Inquiry*, *13*(2), 146-149.

Bliss, J. P., Tidwell, P. D., & Guest, M. A. (1997). The effectiveness of virtual reality for administering spatial navigation training to firefighters. *Presence: Teleoperators & Virtual Environments*, *6*(1), 73-86.

Blom, D., & Encarnacao, J. (2012). Student-chosen criteria for peer assessment of tertiary rock groups in rehearsal and performance: What's important? *British Journal of Music Education*, *29*(01), 25-43.

Blom, D., & Poole, K. (2004). Peer assessment of tertiary music performance: Opportunities for understanding performance assessment and performing through experience and self-reflection. *British Journal of Music Education*, *21*(1), 111-125.

Bodenhausen, G. V., Kramer, G. P., & Süsser, K. (1994). Happiness and stereotypic thinking in social judgment. *Journal of Personality and Social Psychology*, *66*(4), 621-631.

Bonneville-Roussy, A., & Bouffard, T. (2015). When quantity is not enough: Disentangling the roles of practice time, self-regulation and deliberate practice in musical achievement. *Psychology of Music*, *43*(5), 686-704.

Bonneville-Roussy, A., Bouffard, T., & Vezeau, C. (2017a). Trajectories of self-evaluation bias in primary and secondary school: Parental antecedents and academic consequences. *Journal of School Psychology*, *63*, 1-12.

Bonneville-Roussy, A., Evans, P., Verner-Filion, J., Vallerand, R. J., & Bouffard, T. (2017b). Motivation and coping with the stress of assessment: Gender differences in outcomes for university students. *Contemporary Educational Psychology*, *48*, 28-42.

Bonshor, M. (2017). Conductor feedback and the amateur singer: The role of criticism and praise in building choral confidence. *Research Studies in Music Education*, *39*(2), 139-160.

Bouchard, S., Côté, S., St-Jacques, J., Robillard, G., & Renaud, P. (2006). Effectiveness of virtual reality exposure in the treatment of arachnophobia using 3D games. *Technology and Health Care*, *14*(1), 19-27.

Brattico, E., Bogert, B., & Jacobsen, T. (2013). Toward a neural chronometry for the aesthetic experience of music. *Frontiers in Psychology*, *4*(206), 1-21.

Brittin, R. V. (2002). Instrumentalists' assessment of solo performances with compact disc, piano, or no accompaniment. *Journal of Research in Music Education*, *50*(1), 63-74.

Brittin, R. V., & Sheldon, D. A. (1995). Comparing continuous versus static measurements in music listeners' preferences. *Journal of Research in Music Education*, *43*(1), 36-46.

Brittin, R. V., Sheldon, D., & Lee, T. T. (2002). Instrumentalists in Singapore: Assessment of solo performances with compact disc, piano, or no accompaniment. *Bulletin of the Council for Research in Music Education*, *153*(4), 1-7.

Brittin, R. V., & Duke, R. A. (1997). Continuous versus summative evaluations of musical intensity: A comparison of two methods for measuring overall effect. *Journal of Research in Music Education*, *45*(2), 245-258.

Broughton, M., & Stevens, C. (2009). Music, movement and marimba: An investigation of the role of movement and gesture in communicating musical expression to an audience. *Psychology of Music*, *37*(2), 137-153.

Bruine de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, *118*(3), 245-260.

Bruine de Bruin, W. (2006). Save the last dance II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica*, *123*(3), 299-311.

Bruine de Bruin, W., & Keren, G. (2003). Order effects in sequentially judged options due to the direction of comparison. *Organizational Behavior and Human Decision Processes*, *92*(1-2), 91-101.

Buckley, M. R., & Eder, R. W. (1988). BM. Springbett and the notion of the "Snap Decision" in the interview. *Journal of Management*, *14*(1), 59-67.

Bugos, J. A., Heller, J., & Batcheller, D. (2014). Musical nuance task shows reliable differences between musicians and nonmusicians. *Psychomusicology: Music, Mind, and Brain*, *24*(3), 207-213.

Byo, J. L. (1993). The influence of textural and timbral factors on the ability of music majors to detect performance errors. *Journal of Research in Music Education*, *41*(2), 156-167.

Byo, J. L. (1997). The effects of texture and number of parts on the ability of music majors to detect performance errors. *Journal of Research in Music Education*, *45*(1), 51-66.

Ceaser, D. K., Thompson, W. F., & Russo, F. (2009). Expressing tonal closure in music performance: auditory and visual cues. *Canadian Acoustics*, *37*(1), 29-34.

Chaffin, R., Lisboa, T., Logan, T., & Begosh, K. T. (2010). Preparing for memorized cello performance: the role of performance cues. *Psychology of Music*, *38*(1), 3-30.

Chanwimalueang, T., Aufegger, L., Adjei, T., Wasley, D., Cruder, C., Mandic, D. P., & Williamon, A. (2017). Stage call: Cardiovascular reactivity to audition stress in musicians. *PLoS One*, *12*(4), e0176023.

Chapados, C., & Levitin, D. J. (2008). Cross-modal interactions in the experience of musical performances: Physiological correlates. *Cognition*, *108*(3), 639-651.

Chapin, H., Jantzen, K., Kelso, J. A., Steinberg, F., & Large, E. (2010). Dynamic emotional and neural responses to music depend on performance expression and listener experience. *PLoS One*, *5*(12), e13812.

Charleston, S. (2008). Determinants of home atmosphere in English football: A committed supporter perspective. *Journal of Sport Behavior*, *31*(4), 312-328.

Christoforou, C., Christou-Champi, S., Constantinidou, F., & Theodorou, M. (2015). From the eyes and the heart: A novel eye-gaze metric that predicts video preferences of a large audience. *Frontiers in Psychology*, *6*(579), 1-11.

Ciorba, C. R., & Smith, N. Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research in Music Education*, *57*(1), 5-15.

Clark, T., Lisboa, T., & Williamon, A. (2014). An investigation into musicians' thoughts and perceptions during performance. *Research Studies in Music Education*, *36*(1), 19-37.

Clark, T., & Lisboa, T. (2013). Training for sustained performance: Moving toward long-term musician development. *Medical Problems of Performing Artists*, *28*(3), 159-168.

Clay-Williams, R., & Colligan, L. (2015). Back to basics: checklists in aviation and healthcare. *BMJ Quality & Safety*, *24*(7), 428-431.

Clarke, S. R., & Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *44*(4), 509-521.

Clerides, S., & Stengos, T. (2006). Love thy neighbor, love thy kin: Voting biases in the Eurovision Song Contest. *University of Cyprus Working Papers in Economics*, *2006-01*, 1-25.

Clynes, M. (1995). Microstructural musical linguistics: composers' pulses are liked most by the best musicians. *Cognition*, *55*(3), 269-310.

Clynes, M. (1989). Methodology in sentographic measurement of motor expression of emotion: Two-dimensional freedom of gesture essential. *Perceptual and Motor Skills*, *68*(3), 779-783.

Colley, A., North, A., & Hargreaves, D. J. (2003). Gender bias in the evaluation of New Age music. *Scandinavian Journal of Psychology*, *44*(2), 125-131.

Conway, C., & Jeffers, T. (2004). Parent, student, and teacher perceptions of assessment procedures in beginning instrumental music. *Bulletin of the Council for Research in Music Education*, *160*, 16-25.

Coutinho, E., & Scherer, K. R. (2017). The effect of context and audio-visual modality on emotions elicited by a musical performance. *Psychology of Music*, *45*(4), 550-569.

Cross, I. (2010). Listening as covert performance. *Journal of the Royal Musical Association*, *135*(sup1), 67-77.

Dahl, S., & Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception: An Interdisciplinary Journal*, *24*(5), 433-454.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of compari son? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, *12*(3), 166-178.

Daniel, R. (2001). Self assessment in performance. *British Journal of Music Education*, *18*(3), 215-226.

Daniel, R. (2004). Peer assessment in musical performance: The development, trial and evaluation of a methodology for the Australian tertiary environment. *British Journal of Music Education*, *21*(1), 89-110.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, USA*, *108*(17), 6889-6892.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, *21*(2), 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*(1), 33-53.

Davidson, J. W., & Edgar, R. (2003). Gender and race bias in the Judgement of Western art music performance. *Music Education Research*, *5*(2), 169-181.

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence. *The Journal of the American Medical Association*, *296*(9), 1094-1102.

de Borst, A. W., & de Gelder, B. (2015). Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective. *Frontiers in Psychology*, *6*(576), 1-12.

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences*, *111*(5), E618-E625.

Dienes, Z., & Longuet-Higgins, C. (2004). Can musical transformations be implicitly learned? *Cognitive Science*, *28*(4), 531-558.

Difede, J., Cukor, J., Jayasinghe, N., Patt, I., Jedel, S., Spielman, L., Giosan, C., & Hoffman, H. G. (2002). Virtual reality exposure therapy for World Trade Center post-traumatic stress disorder: A case report. *Cyberpsychology & Behavior*, *5*(6), 529-535.

Dohmen, T. J. (2008). The influence of social forces: evidence from the behavior of football referees. *Economic Inquiry*, *46*(3), 411-424.

Dotger, B., Dekaney, E., & Coggiola, J. (2018). In the limelight: Utilizing clinical simulations to enhance music teacher education. *Research Studies in Music Education*, doi.org/10.1177/1321103X18773102.

Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education*, *20*(2), 268-272.

Duguid, M. M., & Thomas-Hunt, M. C. (2015). Condoning stereotyping?: How awareness of stereotyping prevalence impacts expression of stereotypes. *Journal of Applied Psycholology*, *100*, 343-359.

Duke, R. A. (1999). Measures of instructional effectiveness in music research. *Bulletin of the Council for Research in Music Education*, *143*, 1-48.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*(1), 109-128.

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology*, *4*(487), 1-12.

Egermann, H., Grewe, O., Kopiez, R., & Altenmuller, E. (2009a). Social feedback influences musically induced emotions. *Annals of the New York Academy of Sciences*, *1169*, 346-350.

Egermann, H., Nagel, F., Altenmüller, E., & Kopiez, R. (2009b). Continuous measurement of musically-induced emotion: A web experiment. *International Journal of Internet Science*, *4*(1), 4-20.

Egermann, H., Sutherland, M. E., Grewe, O., Nagel, F., Kopiez, R., & Altenmuller, E. (2011). Does music listening in a social context alter experience? A physiological and psychological perspective on emotion. *Musicae Scientiae*, *15*(3), 307-323.

Eggleston, J. (1991). Teaching teachers to assess. *European Journal of Education*, *26*(3), 231-237.

Elliott, C. A. (1995). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education*, *127*, 50-56.

Elliott, C. A., Schneider, M. C., & Zembower, C. M. (2000). Influence of the audition hour on selection to an all-state band. *Journal of Band Research*, *35*(2), 20.

Ellison, S. F., & Mullin, W. P. (2014). Diversity, social goods provision, and performance in the firm. *Journal of Economics & Management Strategy*, *23*(2), 465-481.

Emerson, J. W., Seltzer, M., & Lin, D. (2009). Assessing judging bias: An example from the 2000 Olympic Games. *The American Statistician*, *63*(2), 124-131.

Endo, S., Juhlberg, K., Bradbury, A., & Wing, A. M. (2014). Interaction between physiological and subjective states predicts the effect of a judging panel on the postures of cellists in performance. *Frontiers in Psychology*, *5*(773), 1-11.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363-406.

ESTA. (1984). *Music Competitions: A Report*. London: Alfred Russell.

Evans, P., & Schubert, E. (2008). Relationships between expressed and felt emotions in music. *Musicae Scientiae*, *12*(1), 75-99.

Everett, R. S., & Wojtkiewicz, R. A. (2002). Difference, disparity, and race/ethnic bias in federal sentencing. *Journal of Quantitative Criminology*, *18*(2), 189-211.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287-322.

Fancourt, D., Aufegger, L., & Williamon, A. (2015). Low-stress and high-stress singing have contrasting effects on glucocorticoid response. *Frontiers in Psychology*, *6*(1242), 1-5.

Fancourt, D., & Williamon, A. (2016). Attending a concert reductes glucocorticoids, progesterone and the coritsol/DHEA ratio. *Public Health*, *132*, 101-104.

Fancourt, D., Williamon, A., Carvalho, L. A., Steptoe, A., Dow, R., & Lewis, I. (2016). Singing modulates mood, stress, cortisol, cytokine and neuropeptide activity in cancer patients and carers. *Ecancermedicalscience*, *10*(631), 1-13.

Ferguson, S., Schubert, E., & Dean, R. T. (2011). Continuous subjective loudness responses to reversals and inversions of a sound recording of an orchestral excerpt. *Musicae Scientiae*, *15*(3), 387-401.

Ferm Almqvist, C., Vinge, J., Väkevä, L., & Zandén, O. (2016). Assessment learning in music education: The risk of "criteria compliance" replacing "learning" in the Scandinavian countries. *Research Studies in Music Education*, *39*(1), 3-18.

Fiske, H. E. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education*, *23*(3), 186-196.

Fiske, H. E. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education*, *25*(4), 256-263.

Fiske, H. E. (1979). Musical performance evaluation ability: Toward a model of specificity. *Bulletin of the Council for Research in Music Education*, *59*, 27-31.

Flôres, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking? *The Statistician*, *45*(1), 97-104.

Flossmann, S., & Widmer, G. (2011). Toward a model of performance errors: A qualitative review of Magaloff's Chopin. In A. Williamon, D. Edwards, & L. Bartel (Eds.), *Proceedings of the International Symposium on Performance Science 2011* (pp. 63-68), Ultrecht, The Netherlands: European Association of Conservatoires.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*(1), 45-55.

Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, *5*(1 suppl), 123-147.

Gabrielsson, A. (2003). Musim performance research at the millennium. *Psychology of Music*, *31*(3), 221-272.

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

Gabrielsson, A., & Wik, S. L. (2003). Strong experiences related to music: A descriptive system. *Musicae Scientiae*, *7*(2), 157-217.

Galiana, M., Llinares, C., & Page, Á. (2012). Subjective evaluation of music hall acoustics: Response of expert and non-expert users. *Building and Environment*, *58*, 1-13.

Garicano, L., Palacios-Huerta, I., & Prendergast, C. (2005). Favoritism under social pressure. *Review of Economics and Statistics*, *87*(2), 208-216.

Garrido, S., & Macritchie, J. (2018). Audience engagement with community music performances: Emotional contagion in audiences of a 'pro-am'orchestra in suburban Sydney. *Musicae Scientiae*, doi.org/10.1177/1029864918783027.

Gaunt, H. (2017). Apprenticeship and empowerment: The role of one-to-one lessons. In J. Rink, H. Gaunt, & A. Williamon (Eds.), *Musicians in the Making: Pathways to Creative Performance* (pp. 28-56), Oxford University Press.

Geringer, J. M. (1995). Continuous loudness judgments of dynamics in recorded music excerpts. *Journal of Research in Music Education*, *43*(1), 22-35.

Geringer, J. M., Allen, M. L., MacLeod, R. B., & Scott, L. (2009). Using a prescreening rubric for all-state violin selection: Influences of performance and teaching experience. *Update: Applications of Research in Music Education*, *28*(1), 41-46.

Geringer, J. M., Cassidy, J. W., & Byo, J. L. (1997). Nonmusic majors' cognitive and affective responses to performance and programmatic music videos. *Journal of Research in Music Education*, *45*(2), 221-233.

Geringer, J. M., & Madsen, C. K. (1995). Focus of attention to elements: Listening patterns of musicians and nonmusicians. *Bulletin of the Council for Research in Music Education*, *127*, 80-87.

Geringer, J. M., & Madsen, C. K. (1998). Musicians' ratings of good versus bad vocal and string performances. *Journal of Research in Music Education*, *46*(4), 522-534.

Geringer, J. M., & Madsen, C. K. (2003). Gradual tempo change and aesthetic responses of music majors. *International Journal of Music Education*, *40*(1), 3-15.

Geringer, J. M., Madsen, C. K., & Gregory, D. (2004). A fifteen-year history of the Continuous Response Digital Interface: Issues relating to validity and reliability. *Bulletin of the Council for Research in Music Education*, *160*, 1-15.

Geringer, J. M., & Sasanfar, J. K. (2013). Listener perception of expressivity in collaborative performances containing expressive and unexpressive playing by the pianist. *Journal of Research in Music Education*, *61*(2), 160-174.

Gibbons, C., Richards, S., Valderas, J. M., & Campbell, J. (2017). Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. *Journal of Medical Internet Research*, *19*(3), e65.

Gillespie, R. (1997). Ratings of violin and viola vibrato performance in audio-only and audiovisual presentations. *Journal of Research in Music Education*, *45*(2), 212-220.

Gilman, B. I. (1892a). Report of an experimental test of musical expressiveness (continued). *The American Journal of Psychology*, *5*(1), 42-73.

Gilman, B. I. (1892b). Report on an experimental test of musical expressiveness. *The American Journal of Psychology*, *4*(4), 558-576.

Gingras, B. (2017). Commentary on Kopiez, Wolf, and Platz: The impact of playing from memory on performance evaluation. *Empirical Musicology Review*, *12*(1-2), 15-18.

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*(2), 109-129.

Glowinski, D., Baron, N., Shirole, K., Coll, S. Y., Chaabi, L., Ott, T., Rappaz, M.-A., & Grandjean, D. M. (2015). Evaluating music performance and context-sensitivity with Immersive Virtual Environments. *EAI Endorsed Transactions on Creative Technologies*, *2*, e3.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of “blind” auditions on female musicians. *The American Economic Review*, *90*(4), 715-741.

Goldstein, A. (1980). Thrills in response to music and other stimuli. *Physiological Psychology*, *8*(1), 126-129.

Goolsby, T. W. (1999). Assessment in instrumental music. *Music Educators Journal*, *86*(2), 31-50.

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine*, *66*(12), 762-769.

Graves, C. L. (1903). *The Life & Letters of Sir George Grove*. London: Macmillan.

Gregory, D. (1989). Using computers to measure continuous music responses. *Psychomusicology: A Journal of Research in Music Cognition*, *8*(2), 127-134.

Gregory, D. (1995). Research note: The Continuous Response Digital Interface: An analysis of reliability measures. *Psychomusicology: A Journal of Research in Music Cognition*, *14*(1-2), 197-208.

Grewe, O., Nagel, F., Altenmüller, E., & Kopiez, R. (2009). Individual emotional reactions towards music: Evolutionary-based universals? *Musicae Scientiae*, *13*(2 suppl), 261-287.

Griffiths, N. K. (2008). The effects of concert dress and physical appearance on perceptions of female solo performers. *Musicae Scientiae*, *12*(2), 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*(2), 159-177.

Griffiths, N. K. (2011). The fabric of performance: Values and social practices of classical music expressed through concert dress choice. *Music Performance Research*, *4*, 30-48.

Griffiths, N. K., & Reay, J. L. (2018). The relative importance of aural and visual information in the evaluation of Western canon music performance by musicians and nonmusicians. *Music Perception: An Interdisciplinary Journal*, *35*(3), 364-375.

Grüne-Yanoff, T. (2017). Reflections on the 2017 Nobel Memorial Prize awarded to Richard Thaler. *Erasmus Journal for Philosophy and Economics*, *10*(2), 61-75.

Gutsch, K. U. (1964). One approach toward the development of an individual test for assessing one aspect of instrumental music achievement. *Bulletin of the Council for Research in Music Education*, *2*, 1-5.

Gutsch, K. U. (1965). Evaluation in instrumental music performance: An individual approach. *Bulletin of the Council for Research in Music Education*, *4*, 21-29.

Gynnild, V. (2016). Assessing vocal performances using analytical assessment: A case study. *Music Education Research*, *18*(2), 224-238.

Haddon, E. (2014). Observational learning in the music masterclass. *British Journal of Music Education*, *31*(01), 55-68.

Hales, L. W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement*, *12*(2), 115-117.

Hamman, W. R. (2004). The complexity of team training: what we have learned from aviation and its applications to medicine. *Quality and Safety in Health Care*, *13*(suppl_1), i72-i79.

Hanken, I. M. (2008). Teaching and learning music performance: The master class. *Finnish Journal of Music Education*, *11*(1-2), 26-36.

Hanken, I. M. (2010). The benefits of the master class. The masters' perspective. *Nordic Research in Music Education*, *12*, 149-160.

Hanken, I. M. (2016). Peer learning in specialist higher music education. *Arts and Humanities in Higher Education*, *15*(3-4), 364-375.

Hanoch, Y., & Vitouch, O. (2004). When less is more: Information, emotional arousal and the ecological reframing of the Yerkes-Dodson law. *Theory & Psychology*, *14*(4), 427-452.

Hargreaves, D. J., Messerschmidt, P., & Rubert, C. (1980). Musical preference and evaluation. *Psychology of Music*, *8*(1), 13-18.

Harrison, S. D., Lebler, D., Carey, G., Hitchcock, M., & O'Bryan, J. (2013). Making music or gaining grades? Assessment practices in tertiary music ensembles. *British Journal of Music Education*, *30*(1), 27-42.

Hash, P. M. (2012). An analysis of the ratings and interrater reliability of high school band contests. *Journal of Research in Music Education*, *60*(1), 81-100.

Hatfield, J. L., Halvari, H., & Lemyre, P.-N. (2016). Instrumental practice in the contemporary music academy: A three-phase cycle of Self-Regulated Learning in music students. *Musicae Scientiae*, 316-337.

Herr, P. M. (1989). Priming price: Prior knowledge and context effects. *Journal of Consumer Research*, *16*, 67-75.

Hevner, B. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, *48*, 246-268.

Hewitt, M. P. (2002). Self-evaluation tendencies of junior high instrumentalists. *Journal of Research in Music Education*, *50*(3), 215-226.

Hewitt, M. P. (2005). Self-evaluation accuracy among high school and middle school instrumentalists. *Journal of Research in Music Education*, *53*(2), 148-161.

Hewitt, M. P. (2015). Self-efficacy, self-evaluation, and music performance of secondary-level band students. *Journal of Research in Music Education*, *63*(3), 298-313.

Highben, Z., & Palmer, C. (2004). Effects of auditory and motor mental practice in memorized piano performance. *Bulletin of the Council for Research in Music Education*, 58-65.

Himonides, E. (2011). Mapping a beautiful voice: The continuous response measurement apparatus (CReMA). *Journal of Music, Technology and Education*, *4*(1), 5-25.

Himonides, E. (2017). Music technology and response measurement. In A. King, E. Himonides, & A. Ruthmann (Eds.), *The Routledge Companion to Music, Technology, and Education* (pp. 427-436). Routledge.

Himonides, E., & Welch, G. F. (2005). Building a bridge between aesthetics and acoustics with new technology: A proposed framework for recording emotional response to sung performance quality. *Research Studies in Music Education*, *24*(1), 58-73.

Holmes, P., & Holmes, C. (2013). The performer's experience: A case for using qualitative (phenomenological) methodologies in music performance research. *Musicae Scientiae*, *17*(1), 72-85.

Howard, S. A. (2012). The effect of selected nonmusical factors on adjudicators' ratings of high school solo vocal performances. *Journal of Research in Music Education*, *60*(2), 166-185.

Huang, J., & Krumhansl, C. L. (2011). What does seeing the performer add? It depends on musical style, amount of stage behavior, and audience expertise. *Musicae Scientiae*, *15*(3), 343-364.

Humphreys, J. T. (1998). Musical aptitude testing: From James McKeen Cattell to Carl Emil Seashore. *Research Studies in Music Education*, *10*(1), 42-53.

Hunter, D., & Russ, M. (1996). Peer assessment in performance studies. *British Journal of Music Education*, *13*(01), 67.

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Lee Gordon, D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher*, *27*(1), 10-28.

Jesse, A., & Massaro, D. W. (2010). Seeing a singer helps comprehension of the song's lyrics. *Psychonomic Bulletin & Review*, *17*(3), 323-328.

Johnson, C. M., Madsen, C. K., & Geringer, J. M. (2012). A study of music students' tempo changes of a soloist's performance of Mozart's 1st Horn Concerto. *Journal of Research in Music Education*, *60*(2), 217-231.

Johnson, P. (1997). Performance as experience: the problem of assessment criteria. *British Journal of Music Education*, *14*(03), 271-282.

Johnston, H. (1993). The use of video self-assessment, peer-assessment, and instructor feedback in evaluating conducting skills in music student teachers. *British Journal of Music Education*, *10*(01), 57.

Jørgensen, H. (2004). Strategies for individual practice. In A. Williamon (Ed.), *Musical Excellence: Strategies and Techniques to Enhance Performance* (pp. 85-104). Oxford University Press.

Jørgensen, H. (2008). Instrumental practice: quality and quantity. *Finnish Journal of Music Education*, *11*, 8-18.

Juchniewicz, J. (2008). The influence of physical movement on the perception of musical performance. *Psychology of Music*, *36*(4), 417-427.

Juslin, P. N. (2009). Emotional responses to music. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford Handbook of Music Psychology* (pp. 131-140). Oxford University Press.

Juslin, P. N., Friberg, A., & Bresin, R. (2002). Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae*, *5*(1 suppl), 63-122.

Juslin, P. N., & Lindström, E. (2010). Musical expression of emotions: Modelling listeners' judgements of composed and performed features. *Music Analysis*, *29*(1-3), 334-364.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449-1475.

Kahneman, D. (2011). *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business*, *59*(4.2), S285-S300.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*(6), 1325-1348.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, *5*(1), 193-206.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., & Stone, A. A. (2006). Would you be happier if you were richer? A focusing illusion. *Science*, *312*(5782), 1908-1910.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251.

Kahneman, D., & Tversky, A. (1977). *Intuitive Prediction: Biases and Corrective Procedures*. Arlington, Virgina: Cybernetics Technology Office, DARPA.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341-350.

Kallinen, K., & Ravaja, N. (2006). Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, *10*(2), 191-213.

Kassab, E., Tun, J. K., Arora, S., King, D., Ahmed, K., Miskovic, D., Cope, A., Vadhwana, B., Bello, F., Sevdalis, N., & Kneebone, R. (2011). “Blowing up the barriers” in surgical training: Exploring and validating the concept of distributed simulation. *Annals of Surgery*, *254*(6), 1059-1065.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*(390), 1-16.

Kawakami, A., Furukawa, K., Katahira, K., & Okanoya, K. (2013). Sad music induces pleasant emotion. *Frontiers in Psychology*, *4*(311), 1-15.

Kenny, D. T. (2011). *The Psychology of Music Performance Anxiety*. Oxford University Press.

Khalfa, S., Isabelle, P., Jean-Pierre, B., & Manon, R. (2002). Event-related skin conductance responses to musical emotions in humans. *Neuroscience Letters*, *328*(2), 145-149.

Kim, Y. (2008). The effect of improvisation-assisted desensitization, and music-assisted progressive muscle relaxation and imagery on reducing pianists' music performance anxiety. *Journal of Music Therapy*, *45*(2), 165-191.

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*(4), 322-337.

Klee, D. A. (1999). The effect of computer-generated accompaniment on the preparation of solo flute literature. *Southeastern Journal of Music Education*, *11*, 59-70.

Kneebone, R., Arora, S., King, D., Bello, F., Sevdalis, N., Kassab, E., Aggarwal, R., Darzi, A., & Nestel, D. (2010). Distributed simulation--accessible immersive training. *Medical Teacher*, *32*(1), 65-70.

Kokotsaki, D., Davidson, J., & Coimbra, D. (2001). Investigating the assessment of singers in a music college setting: The students' perspective. *Research Studies in Music Education*, *16*(1), 15-32.

Kolb, A. Y., & Kolb, D. A. (2005). Learning styles and learning spaces: Enhancing experiential learning in Higher Education. *Academy of Management Learning & Education*, *4*(2), 193-212.

Kopiez, R., Wolf, A., & Platz, F. (2017). Small influence of performing from memory on audience evaluation. *Empirical Musicology Review*, *12*(1-2), 2-14.

Kopiez, R. (2003). Intonation of harmonic intervals: Adaptability of expert musicians to equal temperament and just intonation. *Music Perception: An Interdisciplinary Journal*, *20*(4), 383-410.

Krahe, C., Hahn, U., & Whitney, K. (2015). Is seeing (musical) believing? The eye versus the ear in emotional responses to music. *Psychology of Music*, *43*(1), 140-148.

Kroger, C., & Margulis, E. H. (2017). "But they told me it was professional": Extrinsic factors in the evaluation of musical performance. *Psychology of Music*, *45*(1), 49-64.

Kuusi, T. (2015). Musical training and musical ability: Effects on chord discrimination. *Psychology of Music*, *43*(2), 291-301.

Kuwano, S., & Namba, S. (1985). Continuous judgment of level-fluctuating sounds and the relationship between overall loudness and instantaneous loudness. *Psychological Research*, *47*(1), 27-37.

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology*, *29*(3), 299-304.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72-107.

Latukefu, L. (2010). Peer assessment in tertiary level singing: Changing and shaping culture through social interaction. *Research Studies in Music Education*, *32*(1), 61-73.

Lebler, D. (2007). Student-as-master? Reflections on a learning innovation in popular music pedagogy. *International Journal of Music Education*, *25*(3), 205-221.

Lehmann, M., & Kopiez, R. (2013). The influence of on-stage behavior on the subjective evaluation of rock guitar performances. *Musicae Scientiae*, *17*(4), 472-494.

Lerman, L., & Borstel, J. (2003). *Critical Response Process: A Method for Getting Useful Feedback on Anything you Make, from Dance to Dessert*. Liz Lerman Dance Exchange.

Levinson, J. (1987). Evaluating musical performance. *Journal of Aesthetic Education*, *21*(1), 75-88.

Lisboa, T., Chaffin, R., & Demos, A. P. (2014). Recording thoughts while memorizing music: a case study. *Frontiers in Psychology*, *5*(1561), 1-13.

Livingstone, S. R., Thompson, W. F., Wanderley, M. M., & Palmer, C. (2015). Common cues to emotion in the dynamic facial expressions of speech and song. *The Quarterly Journal of Experimental Psychology*, *68*(5), 952-970.

Long, M., Creech, A., Gaunt, H., Hallam, S., & Robertson, L. (2012). Blast from the past: Conservatoire students' experiences and perceptions of public master classes. *Musicae Scientiae*, *16*(3), 286-306.

Madsen, C. K. (1990). Measuring musical response. *Music Educators Journal*, *77*(3), 26-28.

Madsen, C. K. (1997). Focus of attention and aesthetic response. *Journal of Research in Music Education*, *45*(1), 80-89.

Madsen, C. K. (1998). Emotion versus tension in Haydn's Symphony no. 104 as measured by the two-dimensional Continuous Response Digital Interface. *Journal of Research in Music Education*, *46*(4), 546-554.

Madsen, C. K. (2011). From research to the general music classroom. *Music Educators Journal*, *98*(2), 78-82.

Madsen, C. K., & Coggiola, J. C. (2001). The effect of manipulating a CRDI dial on the focus of attention of musicians/nonmusicians and perceived aesthetic response. *Bulletin of the Council for Research in Music Education*, *149*, 13-22.

Madsen, C. K., & Geringer, J. M. (1999). Comparison of good versus bad tone quality/intonation of vocal and string performances: Issues concerning measurement and reliability of the Continuous Response Digital Interface. *Bulletin of the Council for Research in Music Education*, *141*, 86-92.

Madsen, C. K., Geringer, J. M., & Madsen, K. (2009). Adolescent musicians' perceptions of conductors within musical context. *Journal of Research in Music Education*, *57*(1), 16-25.

Madsen, C. K., Geringer, J. M., & Wagner, M. J. (2007). Context specificity in music perception of musicians. *Psychology of Music*, *35*(3), 441-451.

Madura, P. D. (1995). An exploratory investigation of the assessment of vocal jazz improvisation. *Psychology of Music*, *23*(1), 48-62.

Maidhof, C., Pitkaniemi, A., & Tervaniemi, M. (2013). Predictive error detection in pianists: a combined ERP and motion capture study. *Frontiers in Human Neuroscience*, *7*(587), 1-14.

Maidhof, C., Rieger, M., Prinz, W., & Koelsch, S. (2009). Nobody is perfect: ERP effects prior to performance errors in musicians indicate fast monitoring processes. *PLoS One*, *4*(4), e5032.

Maidhof, C., Vavatzanidis, N., Prinz, W., Rieger, M., & Koelsch, S. (2010). Processing expectancy violations during music performance and perception: an ERP study. *Journal of Cognitive Neuroscience*, *22*(10), 2401-2413.

Mann, R. P., Faria, J., Sumpter, D. J., & Krause, J. (2013). The dynamics of audience applause. *Journal of the Royal Society Interface*, *10*(85), 1-7.

Margulis, E. H., Kisida, B., & Greene, J. P. (2015). A knowing ear: The effect of explicit information on children's experience of a musical performance. *Psychology of Music*, *43*(4), 596-605.

Margulis, E. H. (2010). When program notes don't help: Music descriptions and enjoyment. *Psychology of Music*, *38*(3), 285-302.

McAdams, S., Vines, B. W., Vieillard, S., Smith, B. K., & Reynolds, R. (2004). Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Perception: An Interdisciplinary Journal*, *22*(2), 297-350.

McCormick, J., & McPherson, G. (2003). The role of self-efficacy in a musical performance examination: An exploratory structural equation analysis. *Psychology of Music*, *31*(1), 37-51.

McCormick, L. (2008). *Playing to win: A cultural sociology of the international music competition.* Unpublished doctoral thesis, Yale University.

McCormick, L. (2009). Higher, faster, louder: Representations of the international music competition. *Cultural Sociology*, *3*(1), 5-30.

McCormick, L. (2014). Tuning in or turning off: Performing emotion and building cosmopolitan solidarity in international music competitions. *Ethnic and Racial Studies*, *37*(12), 2261-2280.

McCormick, L. (2015). *Performing Civility: International Competitions in Classical Music*. Cambridge University Press.

McDonald, A. (2016). *A Long and Winding Road: Improving Communication with Patients in the NHS*. Marie Curie.

McPherson, G. E. (1995). The assessment of musical performance: Development and validation of five new measures. *Psychology of Music*, *23*(2), 142-161.

McPherson, G. E., & McCormick, J. (2000). The contribution of motivational factors to instrumental performance in a music examination. *Research Studies in Music Education*, *15*(1), 31-39.

McPherson, G. E., & McCormick, J. (2006). Self-efficacy and music performance. *Psychology of Music*, *34*(3), 322-336.

McPherson, G. E., & Schubert, E. (2004). Measuring performance enhancement in music. In A. Williamon (Ed.), *Musical Excellence: Strategies and Techniques to Enhance Performance* (pp. 61-82). Oxford University Press.

McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: Issues and influences. *Research Studies in Music Education*, *10*(1), 12-24.

Medland, E. (2015). Examining the assessment literacy of external examiners. *London Review of Education*, *13*(3), 21-33.

Miles, H. C., Pop, S. R., Watt, S. J., Lawrence, G. P., & John, N. W. (2012). A review of virtual environments for training in ball sports. *Computers & Graphics*, *36*(6), 714-726.

Mills, J. (1987). Assessment of solo musical performance-a preliminary study. *Bulletin of the Council for Research in Music Education*, *91*, 119-125.

Mills, J. (1991). Assessing musical performance musically. *Educational Studies*, *17*(2), 173-181.

Mishra, J. (2010). Effects of structure and serial position on memory errors in musical performance. *Psychology of Music*, *38*(4), 447-461.

Mitchell, H. F., & Benedict, R. (2017). The moot audition: Preparing music performers as expert listeners. *Research Studies in Music Education*, *39*(2), 195-208.

Mitchell, H. F., Kenny, D. T., & Ryan, M. (2010). Perceived improvement in vocal performance following tertiary-level classical vocal training: Do listeners hear systematic progress? *Musicae Scientiae*, *14*(1), 73-93.

Mitchell, H. F., & MacDonald, R. A. R. (2012). Listeners as spectators? Audio-visual integration improves music performer identification. *Psychology of Music*, *42*(1), 112-127.

Mitchell, H. F., & MacDonald, R. A. R. (2016). What you see is what you hear: The importance of visual priming in music performer identification. *Psychology of Music*, *44*(6), 1361-1371.

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior*, *27*(4), 237-254.

Moore, D. A. (1999). Order effects in preference judgments: Evidence for context dependence in the generation of preferences. *Organizational Behavior and Human Decision Processes*, *78*(2), 146-165.

Moore, M. R. (1972). A consideration of the perceptual process in the evaluation of musical performance. *Journal of Research in Music Education*, *20*(2), 273-279.

Morrison, S. J., Price, H. E., Smedley, E. M., & Meals, C. D. (2014). Conductor gestures influence evaluations of ensemble performance. *Frontiers in Psychology*, *5*(806), 1-8.

Morrison, S. J., & Selvey, J. D. (2012). The effect of conductor expressivity on choral ensemble evaluation. In E. Cambouropoulos, C. Tsougras, P. Mavromatis, & K. Pastiadis (Eds.), *Proceedings of the 12th International Conference on Music Perception and Cognition* (p. 700), ESCOM.

Nagel, F., Kopiez, R., Grewe, O., & Altenmüller, E. (2007). EMuJoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, *39*(2), 283-290.

Nakamura, T. (1987). The communication of dynamics between musicians and listeners through musical performance. *Perception & Psychophysics*, *41*(6), 525-533.

Nakra, T. M., & BuSha, B. F. (2014). Synchronous sympathy at the symphony: Conductor and audience accord. *Music Perception: An Interdisciplinary Journal*, *32*(2), 109-124.

Namba, S., & Kuwano, S. (1980). The relation between overall noisiness and instantaneous judgment of noise and the effect of background noise level on noisiness. *Journal of the Acoustical Society of Japan*, *1*(2), 99-106.

Namba, S., & Kuwano, S. (1990). Continuous multi-dimensional assessment of musical performance. *Journal of the Acoustical Society of Japan*, *11*, 43-52.

Namba, S., Kuwano, S., Hatoh, T., & Kato, M. (1991). Assessment of musical performance by using the method of continuous judgment by selected description. *Music Perception: An Interdisciplinary Journal*, *8*(3), 251-275.

Napoles, J., & Madsen, C. K. (2008). Measuring emotional responses to music within a classroom setting. *International Journal of Music Education*, *26*(1), 63-71.

Negut, A., & Sârbescu, P. (2014). Problem music or problem stereotypes? The dynamics of stereotype activation in rock and hip-hop music. *Musicae Scientiae*, *18*(1), 3-16.

Neuberg, S. L. (1989). The goal of forming accurate impressions during social interactions: Attenuating the impact of negative expectancies. *Journal of Personality and Social Psychology*, *56*(3), 374.

Nevill, A. M., & Holder, R. L. (1999). Home advantage in sport. *Sports Medicine*, *28*(4), 221-236.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218.

Nieuwenhuys, A., & Oudejans, R. (2012). Anxiety and perceptual-motor performance: Toward an integrated model of concepts, mechanisms, and processes. *Psychological Research*, *76*(6), 747-759.

North, A. C., Colley, A. M., & Hargreaves, D. J. (2003). Adolescents' perceptions of the music of male and female composers. *Psychology of Music*, *31*(2), 139-154.

North, A. C., & Hargreaves, D. J. (1998). Complexity, prototypicality, familiarity, and the perception of musical quality. *Psychomusicology: A Journal of Research in Music Cognition*, *17*(1-2), 77.

Ockelford, A., & Sergeant, D. (2013). Musical expectancy in atonal contexts: Musicians' perception of "antistructure". *Psychology of Music*, *41*(2), 139-174.

Orman, E. K. (2003). Effect of virtual reality graded exposure on heart rate and self-reported anxiety levels of performing saxophonists. *Journal of Research in Music Education*, *51*(4), 302-315.

Orman, E. K. (2004). Effect of virtual reality graded exposure on anxiety levels of performing musicians: A case study. *Journal of Music Therapy*, *41*(1), 70-78.

Page, L., & Page, K. (2010). Last shall be first: A field study of biases in sequential performance evaluation on the Idol series. *Journal of Economic Behavior & Organization*, *73*(2), 186-198.

Palmer, C., Jewett, L. R., & Steinhauer, K. (2009). Effects of context on electrophysiological response to musical accents. *Annals of the New York Academy of Sciences*, *1169*, 470-480.

Papageorgi, I., Creech, A., Haddon, E., Morton, F., De, B., C., Himonides, E., Potter, J., Duffy, C., Whyton, T., & Welch, G. (2010). Perceptions and predictions of expertise in advanced musical learners. *Psychology of Music*, *38*(1), 31-66.

Paris, S. G., & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In B. F. Jones & L. Idol (Eds.), *Dimensions of Tinking and Cognitive Instruction* (pp. 15-51). Hillsdale, NJ, US: Lawurence Erlbaum Associates, Inc.

Park, J., & Banaji, M. R. (2000). Mood and heuristics: The influence of happy and sad states on sensitivity and bias in stereotyping. *Journal of Personality and Social Psychology*, *78*(6), 1005-1023.

Perkins, R. (2013). Learning cultures and the conservatoire: An ethnographically-informed case study. *Music Education Research*, *15*(2), 196-213.

Pitts, S. E. (2004). ‘Everybody wants to be Pavarotti’: The experience of music for performers and audience at a Gilbert and Sullivan festival. *Journal of the Royal Musical Association*, 143-160.

Pitts, S. E. (2005). What makes an audience? Investigating the roles and experiences of listeners at a chamber music festival. *Music and Letters*, *86*(2), 257-269.

Pitts, S. E. (2016). On the edge of their seats: Comparing first impressions and regular attendance in arts audiences. *Psychology of Music*, *44*(5), 1175-1192.

Plack, D. S. (2006). *The effect of performance medium on the emotional response of the listener as measured by the Continuous Response Digital Interface.* Unpublished doctoral thesis, The Florida State University.

Platz, F., & Kopiez, R. (2013). When the first impression counts: Music performers, audience and the evaluation of stage entrance behaviour. *Musicae Scientiae*, *17*(2), 167-197.

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception: An Interdisciplinary Journal*, *30*(1), 71-83.

Plazak, J., & Huron, D. (2011). The first three seconds: Listener knowledge gained from brief musical excerpts. *Musicae Scientiae*, *15*(1), 29-44.

Preuschoff, K. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*(115), 1-12.

Price, H. E., Mann, A., & Morrison, S. J. (2016). Effect of conductor expressivity on ensemble evaluations by nonmusic majors. *International Journal of Music Education*, *34*(2), 135-142.

Quinto, L. R., Thompson, W. F., & Taylor, A. (2014a). The contributions of compositional structure and performance expression to the communication of emotion in music. *Psychology of Music*, *42*(4), 503-524.

Quinto, L. R., Thompson, W. F., Kroos, C., & Palmer, C. (2014b). Singing emotionally: A study of pre-production, production, and post-production facial expressions. *Frontiers in Psychology*, *5*(262), 1-15.

Radbourne, J., Glow, H., & Johanson, K. (2010). Measuring the intrinsic benefits of arts attendance. *Cultural Trends*, *19*(4), 307-324.

Radbourne, J., Johanson, K., & Glow, H. (2010). Empowering audiences to measure quality. *Participations: Journal of Audience & Reception Studies*, *7*(2), 360-379.

Radbourne, J., Johanson, K., & Glow, H. (2014). The value of 'being there': How live experience measures quality for the audience. In K. Burland & S. Pitts (Eds.), *Coughing and Clapping: Investigating Audience Experience* (pp. 55-68). Farnham, UK: Ashgate Publishing, Ltd.

Radbourne, J., Johanson, K., Glow, H., & White, T. (2009). The audience experience: Measuring quality in the performing arts. *International Journal of Arts Management*, *11*(3), 16-29.

Radocy, R. E. (1976). Effects of authority figure biases on changing judgments of musical events. *Journal of Research in Music Education*, *24*(3), 119-128.

Radocy, R. E. (1986). On quantifying the uncountable in musical behavior. *Bulletin of the Council for Research in Music Education*, *88*, 22-31.

Rakowski, A. (1990). Intonation variants of musical intervals in isolation and in musical contexts. *Psychology of Music*, *18*(1), 60-72.

Ravel, M. (1919). *Le tombeau de Couperin: Suite d'orchestre (Orchestral Score: 1st & 2nd Oboe)*. Paris: Durands & Fils.

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*(1), 3-8.

Repp, B. H. (1996). The art of inaccuracy: Why pianists' errors are difficult to hear. *Music Perception: An Interdisciplinary Journal*, *14*(2), 161-183.

Repp, B. H., & Knoblich, G. (2004). Perceiving action identity: How pianists recognize their own performances. *Psychological Science*, *15*(9), 604-609.

Ringle, C. M., Sarstedt, M., & Straub, D. (2012). A critical look at the use of PLS-SEM in MIS Quarterly. *MIS Quarterly*, *36*(1), iii-xiv.

Ritchie, L., & Williamon, A. (2011). Measuring distinct types of musical self-efficacy. *Psychology of Music*, *39*(3), 328-344.

Ritchie, L., & Williamon, A. (2012). Self-efficacy as a predictor of musical performance quality. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 334-340.

Robinson, C. R. (1993). Singers' self-assessment of choral performance: Next-day recollections versus concert tape evaluation. *Southeastern Journal of Music Education*, *4*, 224-233.

Rodenberg, R. (2011). Perception ≠ reality: Analyzing specific allegations of NBA referee bias. *Journal of Quantitative Analysis in Sports*, *7*(2), 1-11.

Rodger, M. W., Craig, C. M., & O'Modhrain, S. (2012). Expertise is perceived from both sound and body movement in musical performance. *Human Movement Science*, *31*(5), 1137-1150.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation*, *11*(10), 1-13.

Rothbaum, B. O., Hodges, L., Smith, S., Lee, J. H., & Price, L. (2000). A controlled study of virtual reality exposure therapy for the fear of flying. *Journal of Consulting and Clinical Psychology*, *68*(6), 1020-1031.

Ruiz, M. H., Jabusch, H.-C., & Altenmüller, E. (2009). Fast feedforward error-detection mechanisms in highly skilled music performance. In A. Williamon, S. Pretty, & R. Buck (Eds.), *Proceedings of the International Symposium on Performance Science 2009* (pp. 187-197), Ultrecht, The Netherlands: European Association of Conservatoires.

Russell, B. E. (2015). An empirical study of a solo performance assessment model. *International Journal of Music Education*, *33*(3), 359-371.

Russell, P. A. (1987). Effects of repetition on the familiarity and likeability of popular music recordings. *Psychology of Music*, *15*(2), 187-197.

Ryan, C., & Costa-Giomi, E. (2004). Attractiveness bias in the evaluation of young pianists' performances. *Journal of Research in Music Education*, *52*(2), 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*(4), 559-572.

Samplaski, A. (2004). The relative perceptual salience of Tn and TnI. *Music Perception: An Interdisciplinary Journal*, *21*(4), 545-559.

Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews: Neuroscience*, *6*(4), 332-339.

Sandstrom, G. M., & Russo, F. A. (2013). Absorption in music: Development of a scale to identify individuals with strong emotional responses to music. *Psychology of Music*, *41*(2), 216-228.

Saunders, T. C., & Holahan, J. M. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. *Journal of Research in Music Education*, *45*(2), 259-272.

Saxena, A. (2014). Workforce diversity: A key to improve productivity. *Procedia Economics and Finance*, *11*, 76-85.

Schmalstieg, E. (1972). The construction of a reliable test for assessing musical performance. *Journal of Research in Music Education*, *20*(2), 280-282.

Schubert, E. (1996). Enjoyment of negative emotions in music: An associative network explanation. *Psychology of Music*, *24*(1), 18-28.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, *51*(3), 154-165.

Schubert, E. (2001). Continuous measurement of self-report emotional response to music. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and Emotion: Theory and Research* (pp. 393-414). Oxford University Press.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception: An Interdisciplinary Journal*, *21*(4), 561-585.

Schubert, E. (2007). The influence of emotion, locus of emotion and familiarity upon preference in music. *Psychology of Music*, *35*(3), 499-515.

Schubert, E. (2013). Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music. *Psychology of Music*, *41*(3), 350-371.

Schutz, M. (2017). Acoustic constraints and musical consequences: Exploring composers' use of cues for musical emotion. *Frontiers in Psychology*, *8*(1402), 1-10.

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception*, *36*(6), 888-897.

Searby, M., & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the School of Music, Kingston University. *Assessment & Evaluation in Higher Education*, *22*(4), 371-383.

Seashore, C. E. (1939). *Measures of Musical Talent and Manual of Instructions*. Camden, New Jersey: RCA Victor Divison.

Shahani, C., Dipboye, R. L., & Gehrlein, T. M. (1993). Attractiveness bias in the interview: Exploring the boundaries of an effect. *Basic and Applied Social Psychology*, *14*(3), 317-328.

Shoda, H., & Adachi, M. (2014). Why live recording sounds better: a case study of Schumann's Traumerei. *Frontiers in Psychology*, *5*(1564), 1-15.

Silva, K. M., & Silva, F. J. (2009). What radio can do to increase a song's appeal: a study of Canadian music presented to American college students. *Psychology of Music*, *37*(2), 181-194.

Silveira, J. M. (2014). The effect of body movement on listeners' perceptions of musicality in trombone quartet performance. *International Journal of Music Education*, *32*(3), 311-323.

Silveira, J. M., & Diaz, F. M. (2014). The effect of subtitles on listeners' perceptions of expressivity. *Psychology of Music*, *42*(2), 233-250.

Silvey, B. A. (2009). The effects of band labels on evaluators' judgments of musical performance. *Update: Applications of Research in Music Education*, *28*(1), 47-52.

Singhal, A., Tien, Y.-Y., & Hsia, R. Y. (2016). Racial-ethnic disparities in opioid prescriptions at emergency department visits for conditions commonly associated with prescription drug abuse. *PLoS One*, *11*(8), e0159224.

Slater, M., Pertaub, D.-P., & Steed, A. (1999). Public speaking in virtual reality: Facing an audience of avatars. *IEEE Computer Graphics and Applications*, *19*(2), 6-9.

Sloboda, J. A., & Lehmann, A. C. (2001). Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude. *Music Perception: An Interdisciplinary Journal*, *19*(1), 87-120.

Smith, B. R. (2004). Five judges' evaluation of audiotaped string performance in international competition. *Bulletin of the Council for Research in Music Education*, *160*, 61-69.

Spahn, C., Strukely, S., & Lehmann, A. (2004). Health conditions, attitudes toward study, and attitudes toward health at the beginning of university study: music students in comparison with other student populations. *Medical Problems of Performing Artists*, *19*(1), 26-33.

Springbett, B. M. (1958). Factors affecting the final decision in the employment interview. *Canadian Journal of Psychology*, *12*(1), 13-22.

Springer, D. G., & Schlegel, A. L. (2016). Effects of applause magnitude and musical style on listeners' evaluations of wind band performances. *Psychology of Music*, *44*(4), 742-756.

Springer, D. G., & Silvey, B. A. (2018). The role of accompaniment quality in the evaluation of solo instrumental performance. *Journal of Research in Music Education*, *66*(1), 82-110.

Stacho, L., Saarikallio, S., Van, Z., A., Huotilainen, M., & Toiviainen, P. (2013). Perception of emotional content in musical performances by 3-7-year-old children. *Musicae Scientiae*, *17*(4), 495-512.

Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education*, *18*(1), 46-56.

Stefani, R. (1998). Predicting outcomes. In J. Bennett (Ed.), *Statistics in Sport* (pp. 249-275). London: Arnold.

Stern, R. A., Arruda, J. E., Hooper, C. R., Wolfner, G. D., & Morey, C. E. (1997). Visual analogue mood scales to measure internal mood state in neurologically impaired patients: Description and initial validity evidence. *Aphasiology*, *11*(1), 59-71.

Stewart, A. (2011). Examiner training: The full story. Retrieved June, 2018 from https://us.abrsm.org/en/exam-support/exam-support-articles/article/examiner-training-the-full-story/175/.

Sustersic, M., Gauchet, A., Kernou, A., Gibert, C., Foote, A., Vermorel, C., & Bosson, J. L. (2018). A scale assessing doctor-patient communication in a context of acute conditions based on a systematic review. *PLoS One*, *13*(2), e0192306.

Sutherland, M. E., Grewe, O., Egermann, H., Nagel, F., Kopiez, R., & Altenmuller, E. (2009). The influence of social situations on music listening. *Annals of the New York Academy of Sciences*, *1169*, 363-367.

Szigeti, J. (1947). *With Strings Attached: Reminiscences and Reflections*. New York: Alfred A. Knopf.

Taylor, A. (2010). Participation in a master class: Experiences of older amateur pianists. *Music Education Research*, *12*(2), 199-217.

Tchaikovsky, P. I. (1946). *Symphony No. 4 in F minor, Op. 36 (Orchestral Score: 1st Oboe)*. Leipzig: Bruckner-Verlag.

Teichert, T., Ferrera, V. P., & Grinband, J. (2014). Humans optimize decision-making by delaying decision onset. *PloS One*, *9*(3), e89638.

Thaler, R. H., Tversky, A., Kahneman, D., & Schwartz, A. (1997). The effect of myopia and loss aversion on risk taking: An experimental test. *The Quarterly Journal of Economics*, *112*(2), 647-661.

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review*, *106*(7), 1577-1600.

Thompson, M. R., & Luck, G. (2012). Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. *Musicae Scientiae*, *16*(1), 19-40.

Thompson, S. (2005). *Evaluating evaluation: An investigation of quality judgements in musical performance.* Unpublished doctoral thesis, University of London.

Thompson, S. (2006). Audience responses to a live orchestral concert. *Musicae Scientiae*, *10*(2), 215-244.

Thompson, S. (2007). Determinants of listeners' enjoyment of a performance. *Psychology of Music*, *35*(1), 20-36.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception: An Interdisciplinary Journal*, *21*(1), 21-41.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Perception: An Interdisciplinary Journal*, *25*(1), 13-29.

Thompson, W. F., Diamond, C. T. P., & Balkwill, L.-L. (1998). The adjudication of six performances of a Chopin Etude: A study of expert knowledge. *Psychology of Music*, *26*(2), 154-174.

Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica*, *2005*(156), 203-227.

Thompson, W. F., & Robitaille, B. (1992). Can composers express emotions through music? *Empirical Studies of the Arts*, *10*(1), 79-89.

Thompson, W. F., Russo, F. A., & Livingstone, S. R. (2010). Facial expressions of singers influence perceived pitch relations. *Psychonomic Bulletin & Review*, *17*(3), 317-322.

Thompson, W. F., Russo, F. A., & Quinto, L. (2008). Audio-visual integration of emotional cues in song. *Cognition & Emotion*, *22*(8), 1457-1470.

Timmers, R. (2007). Perception of music performance on historical and modern commercial recordings. *Journal of the Acoustical Society of America*, *122*(5), 2872-2880.

Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences, USA*, *110*(36), 14580-14585.

Tsay, C.-J. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes*, *124*(1), 24-33.

Tschacher, W., Greenwood, S., Kirchberg, V., Wintzerith, S., van, D. B., Karen, & Tröndle, M. (2012). Physiological correlates of aesthetic perception of artworks in a museum. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(1), 96-103.

Tucker, D. H., & Rowe, P. M. (1977). Consulting the application form prior to the interview: An essential step in the selection process. *Journal of Applied Psychology*, *62*(3), 283-287.

Tullar, W. L., Mullins, T. W., & Caldwell, S. A. (1979). Effects of interview length and applicant quality on interview decision time. *Journal of Applied Psychology*, *64*(6), 669-674.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207-232.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039-1061.

Ueno, K., & Tachibana, H. (2005). Cognitive modeling of musician's perception in concert halls. *Acoustical Science and Technology*, *26*(2), 156-161.

Upham, F. (2011). *Quantifying the temporal dynamics of music listening: A critical investigation of analysis techniques for collections of continuous responses to music.* Unpublished MA thesis, McGill University.

VanWeelden, K. (2004). Racially stereotyped music and conductor race: Perceptions of performance. *Bulletin of the Council for Research in Music Education*, *160*, 38-48.

Varey, C., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making*, *5*(3), 169-185.

Vasil, T. (1973). *The effects of systematically varying selected factors on music performance adjudication.* Unpublished doctoral thesis, University of Connecticut.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Dalca, I. M., & Levitin, D. J. (2011). Music to my eyes: cross-modal interactions in the perception of emotions in musical performance. *Cognition*, *118*(2), 157-170.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition*, *101*(1), 80-113.

Voyer, D., Kinch, S., & Wright, E. F. (2006). The home disadvantage: Examination of the self-image redefinition hypothesis. *Journal of Sport Behavior*, *29*(3), 270-279.

Vurma, A. (2014). Timbre-induced pitch shift from the perspective of Signal Detection Theory: the impact of musical expertise, silence interval, and pitch region. *Frontiers in Psychology*, *5*(44), 1-13.

Wallerstedt, C., Pramling, N., & Saljo, R. (2014). Learning to discern and account: The trajectory of a listening skill in an institutional setting. *Psychology of Music*, *42*(3), 366-385.

Wang, C., Wong, J., Zhu, X., Röggla, T., Jansen, J., & Cesar, P. (2016). Quantifying audience experience in the wild: Heuristics for developing and deploying a biosensor infrastructure in theaters. In the *Proceedings of the Eighth International Conference on Quality of Multimedia Experience (QoMEX), 2016* (pp. 1-6), IEEE.

Wapnick, J., Campbell, L., Siddell-Strebel, J., & Darrow, A.-A. (2009). Effects of non-musical attributes and excerpt duration on ratings of high-level piano performances. *Musicae Scientiae*, *13*(1), 35-54.

Wapnick, J., Darrow, A. A., Kovacs, J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education*, *45*(3), 470-479.

Wapnick, J., Flowers, P., Alegant, M., & Jasinskas, L. (1993). Consistency in piano performance evaluation. *Journal of Research in Music Education*, *41*(4), 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*(4), 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.

Wapnick, J., & Rosenquist, M.-J. (1991). Preferences of undergraduate music majors for sequenced versus performed piano music. *Journal of Research in Music Education*, *39*(2), 152-160.

Wapnick, J., Ryan, C., Campbell, L., Deek, P., Lemire, R., & Darrow, A.-A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performances. *Journal of Research in Music Education*, *53*(2), 162-176.

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: Current state of the art. *Advances in Health Sciences Education*, *7*(1), 63-80.

Warren, R. A., & Curtis, M. E. (2015). The actual vs. predicted effects of intonation accuracy on vocal performance quality. *Music Perception: An Interdisciplinary Journal*, *33*(2), 135-146.

Watson, K. B. (1942). The nature and measurement of musical meanings. *Psychological Monographs: General and Applied*, *54*(2), i-43.

Welch, G. (1994). The assessment of singing. *Psychology of Music*, *22*(1), 3-19.

Wesolowski, B. C. (2016). Assessing jazz big band performance: The development, validation, and application of a facet-factorial rating scale. *Psychology of Music*, *44*(3), 324-339.

Wesolowski, B. C. (2017). A facet-factorial approach towards the development and validation of a jazz rhythm section performance rating scale. *International Journal of Music Education*, *35*(1), 17-30.

Wesolowski, B. C., Amend, R. M., Barnstead, T. S., Edwards, A. S., Everhart, M., Goins, Q. R., Grogan III, R. J., Herceg, A. M., Jenkins, S. I., & Johns, P. M. (2017). The development of a secondary-level solo wind instrument performance rubric using the Multifaceted Rasch Partial Credit Measurement Model. *Journal of Research in Music Education*, *65*(1), 95-119.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147-170.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch Partial Credit Model. *Music Perception: An Interdisciplinary Journal*, *33*(5), 662-678.

Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of social psychology*, *26*(4), 557-580.

Wilkinson, T., & Pollard, R. (2006). A temporary decline in home advantage when moving to a new stadium. *Journal of Sport Behavior*, *29*(2), 190-197.

Williamon, A. (1999). The value of performing from memory. *Psychology of Music*, *27*(1), 84-95.

Williamon, A., Aufegger, L., & Eiholzer, H. (2014). Simulating and stimulating performance: introducing distributed simulation to enhance musical learning and performance. *Frontiers in Psychology*, *5*(25), 1-9.

Williamon, A., & Thompson, S. (2006). Awareness and incidence of health problems among conservatoire students. *Psychology of Music*, *34*(4), 411-430.

Williamon, A., & Valentine, E. (2000). Quantity and quality of musical practice as predictors of performance quality. *British Journal of Psychology*, *91*(3), 353-376.

Williams, L. R., Fredrickson, W. E., & Atkinson, S. (2011). Focus of attention to melody or harmony and perception of music tension: An exploratory study. *International Journal of Music Education*, *29*(1), 72-81.

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*(3), 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*(2), 181-192.

Wilson, V. E. (1977). Objectivity and effect of order of appearance in judging of synchronized swimming meets. *Perceptual and Motor Skills*, *44*(1), 295-298.

Wing, H. D. (1947). *Standardized Tests of Musical Intelligence*. Sheffield, UK: City Training College.

Winter, N. (1993). Music performance assessment: A study of the effects of training and experience on the criteria used by music examiners. *International Journal of Music Education*, *22*(1), 34-39.

Wolf, A., & Kopiez, R. (2014). Do grades reflect the development of excellence in music students? The prognostic validity of entrance exams at universities of music. *Musicae Scientiae*, *18*(2), 232-248.

Woody, R. H. (2002). The relationship between musicians' expectations and their perception of expressive features in an aural model. *Research Studies in Music Education*, *18*(1), 57-65.

Wrigley, W. J., & Emmerson, S. B. (2013). Ecological development and validation of a music performance rating scale for five instrument families. *Psychology of Music*, *41*(1), 97-118.

Wu, C.-W., & Lerch, A. (2018). Learned features for the assessment of percussive music performances. In D. Bulterman, A. Kitazawa, D. Ostrowski, & P. Sheu (Eds.), *Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC)* (pp. 93-99), Laguna Hills, USA: IEEE.

Ybarra, O. (2001). When first impressions don't last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition*, *19*(5), 491-520.

Yuen, K. S. L., & Lee, T. M. C. (2003). Could mood state affect risk-taking decisions? *Journal of Affective Disorders*, *75*(1), 11-18.

Zdzinski, S. F. (1991). Measurement of solo instrumental music performance: A review of literature. *Bulletin of the Council for Research in Music Education*, *109*, 47-58.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, *50*(3), 245-255.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personal Psychology Compass*, *2*(3), 1497-1517.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3-17.

Zitzewitz, E. (2006). Nationalism in winter sports judging and its lessons for organizational decision making. *Journal of Economics & Management Strategy*, *15*(1), 67-99.

Zitzewitz, E. (2014). Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging. *Journal of Sports Economics*, *15*(1), 3-30.

APPENDIX 1: STUDY 1 AUDIO RECORDINGS

The audio recordings listed below were used as experimental stimuli for Study 1 (see Section 3.2.2). Recordings A - C were manipulated by the author to contain performance errors, allowing for causal effects of the error on performance quality ratings to be determined.

These recordings can be downloaded as Supplementary Files with the following publication:

Waddell, G., Perkins, R., & Williamon, A. (2018). Making an impression: Error location and repertoire features affect performance quality rating processes. *Music Perception*, *36*(1), 60-76.

A1. Chopin Etude (no error)
A2. Chopin Etude (error-start)
A3. Chopin Etude (error-recap)
B1. Chopin Waltz (no error)
B2. Chopin Waltz (error-start)
B3. Chopin Waltz (error-recap)
C1. Chopin Prelude (no error)
C2. Chopin Prelude (error-start)
D. Chopin Tarantelle
E. Eckhardt-Gramatté Caprice

APPENDIX 2: STUDY 1 FORMS

The custom questionnaires on the following two pages were developed and used for data collection within Study 1 (see Section 3.2.4). Copies of the first form (Questionnaire A) were given to the participant immediately following their rating of each of the stimuli using the RCM continuous measurement software. The second (Questionnaire B) was given at the end of the session (see Section 3.2.5 for the full procedure).

QUESTIONNAIRE A

(Please circle your answer)

How would you rate the *overall quality* of this performance?

Poor						Excellent
1	2	3	4	5	6	7

How *familiar* were you with this work?

Never heard it						Extremely Familiar
1	2	3	4	5	6	7

Have you performed this work? **Yes / No**

Do you *like* this composition?

Not at all						Very Much
1	2	3	4	5	6	7

How typical was the performance of this work compared to others you have heard (if applicable)?

Very Differing						Highly Typical
1	2	3	4	5	6	7

How *difficult* do think this work is to perform?

Not Difficult						Very Difficult
1	2	3	4	5	6	7

Any comments about this performance?

____

____

____

QUESTIONNAIRE B
Personal Information

Name: ____

Date of Birth: _______________________(DD/MM/YY) **Gender**: M / F

How old were you when you began your musical training? ___________________

Principal Instrument: ____

Other Instrument(s): ____

____

Programme: ____________________________ **Year in Programme**: ___________

To what extent do you enjoy listening to the following types of music (please circle):

	Not at All						Very Much
Baroque	1	2	3	4	5	6	7
Classical	1	2	3	4	5	6	7
Romantic	1	2	3	4	5	6	7
20th Century	1	2	3	4	5	6	7

Any comments about the study?

____

____

____

____

____

APPENDIX 3: STUDY 2 VIDEOS

The video recordings listed below were used as experimental stimuli for Study 2 (see Section 4.2.2). A single, staged performance of a Chopin Etude was manipulated to create five versions varying in their inclusion of an 'appropriate' versus 'inappropriate' stage entrance, as well as the presence of a major aural performance error and/or negative facial reaction.

These recordings can be downloaded as Supplementary Files with the following publication:

Waddell, G. & Williamon, A. (2017). Eye of the beholder: Stage entrance behaviour and facial expression affect continuous quality ratings in music performance. *Frontiers in Psychology, 8*(513), 1-14.

Video 1. Standard

Video 2. Inappropriate stage entrance (Entrance)

Video 3. Aural error with facial reaction (Aural/facial)

Video 4. Aural error only (Aural)

Video 5. Facial reaction only (Facial)

APPENDIX 4: STUDY 2 SOFTWARE

The code below was written by the author to collect continuous responses of perceived performance quality synchronised with an audiovisual stimulus, as required by the experimental design in Chapter 4 (see Section 4.2.3). With the explanatory text in italics removed, it can be used in conjunction with the software package *Presentation* (Neurobehavioral Software) to replicate the experiment.

Defines the scenario within the Presentation software framework. Each experimental condition has its own scenario, in this case the performance with aural performance error and corresponding negative facial reaction (Facial_Aural).

```
scenario = "Facial_Aural";
response_matching = simple_matching;
max_y = 200;
active_buttons = 3;

begin;
```

Defines the grey bar at the bottom of the text screen, on top of which sits a red bar whose width will be later defined by horizontal movement of the mouse. Includes text for rating scale of 'Poor-2-3-4-5-6-Excellent', and instructions for the user.

```
picture {
      box {color = 100, 100, 100; height = 50; width = 490; } box2;
      left_x = -250; y = -150;

      box {color = 255, 0, 0; height = 50; width = 490; } box1;
      left_x = -250; y = -150;

      text {caption = "Poor"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = -250; y = -120;
```

```
      text {caption = "2"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = -168; y = -120;
      text {caption = "3"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = -87; y = -120;
      text {caption = "4"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = -5; y = -120;
      text {caption = "5"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = 76; y = -120;
      text {caption = "6"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = 158; y = -120;
      text {caption = "Excellent"; font_size = 6; font =
"Helvetica"; font_color = 255, 255, 255; } ; x = 240; y = -120;

      text {caption = "Click to begin recording"; font_size = 10;
font = "Helvetica"; font_color = 255, 255, 255; } ; x = 0; y = -190;

} pic;
```

Defines an identical setup, except the overlaid bar of varying width is blue (to be used later to indicate that the user has clicked the mouse and started recording) and the instruction to click and begin has been removed.

```
picture {
      box {color = 100, 100, 100; height = 50; width = 490; } box2b;
      left_x = -250; y = -150;

      box {color = 50, 150, 255; height = 50; width = 490; } box1b;
      left_x = -250; y = -150;

      text {caption = "Poor"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = -250; y = -120;
      text {caption = "2"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = -168; y = -120;
      text {caption = "3"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = -87; y = -120;
      text {caption = "4"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = -5; y = -120;
      text {caption = "5"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = 76; y = -120;
      text {caption = "6"; font_size = 6; font = "Helvetica";
font_color = 255, 255, 255; } ; x = 158; y = -120;
      text {caption = "Excellent"; font_size = 6; font =
"Helvetica"; font_color = 255, 255, 255; } ; x = 240; y = -120;

} pic2;
```

Defines the video player and relevant file for the experimental condition/scenario.

```
video {
      filename = "Facial_Aural.avi";
      x = 0; y = 50;
      height = 300; width = 533;
} vid1;
```

Defines the opening screen of the experiment with instructions for the participant.

```
picture {
      text {caption = "Rate the following performance's quality from
'Poor' to 'Excellent'. Move the mouse left and right and CLICK to
record your first judgement as soon as you are able to make it. Keep
moving the mouse as your judgement changes."; font_size = 20; font =
"Helvetica"; font_color = 225, 225, 225; max_text_width = 400;} ;
      x = 0; y = 50;
      text {caption = "Press 'Enter' when you are ready to begin";
font_size = 20; font = "Helvetica"; font_color = 225, 225, 225;
max_text_width = 400;} ;
      x = 0; y = -130;
} firstscreen;
```

Defines the final screen of the experiment when the video has completed.

```
picture {
      text {caption = "The video is complete. Please ask the
experimenter for further instructions."; font_size = 20; font =
"Helvetica"; font_color = 225, 225, 225; max_text_width = 400;} ;
      x = 0; y = 0;
} finalscreen;
```

Sequential operating instructions for the experiment. First, the program creates a .csv file in a locally defined folder with the participant number (entered earlier in the Presentation software package) in the filename. Column headings for subject number, date and time, time since the start of the video playback, horizontal mouse position, and whether or not the mouse had been clicked are written to the file.

```
begin_pcl;

output_file file = new output_file;
file.open(logfile.subject() + "-PositionResults" + ".csv");
file.print(logfile.subject());
file.print("\n");
file.print(date_time());
file.print("\n");
file.print("Time (ms),");
file.print("Position,");
file.print("Clicked?,");
file.print("\n");
```

The program calls up the opening experiment screen and holds it until the participant presses "Enter", moving to the next section.

```
response_manager.set_button_active(1, false);

loop until response_manager.total_response_count(3) == 1 begin
      firstscreen.present();
end;
```

The program calls and begins the video player. An embedded series of conditional loops are started that first display the measurement scale with the red bar at the bottom of the screen. With every new video frame the width of the red box is defined by the horizontal mouse position, thus allowing the user to move the scale back and forth. Also, with every frame the program checks the time since the video started playing, and if it has passed a 500 ms milestone (leading to some variance based on asynchrony with the framerate, although never exceeding 67 ms) writes the time since the video has been playing, horizontal position of the mouse on a 1 - 70 scale, and whether the mouse has been clicked to the .csv file created above. When the mouse is clicked the program enters an embedded loop that continues the same pattern but displays the blue bar instead of the read and writes to the file that the mouse has now been clicked and the mouse position can now be considered active evaluation.

```
response_manager.set_button_active(1, true);

vid1.present();

mouse mse = response_manager.get_mouse(1);

double i = 1.0;

loop until vid1.frame_position() == vid1.frame_duration() begin

      mse.poll();
      box1.set_width(mse.x() );
      pic.present();

      if (vid1.position() >= 500.0*i) then

            file.print(vid1.position());
            file.print(", ");
            file.print(mse.x() / 7);
            file.print(", ");
            file.print(response_manager.total_response_count(1));
            file.print("\n");

            i = i + 1.0;

      end;

      if ((response_manager.total_response_count(1)) >= 1) then

            loop until vid1.frame_position() ==
vid1.frame_duration() begin

                  mse.poll();
                  box1b.set_width(mse.x() );
                  pic2.present();

                  if (vid1.position() >= 500.0*i) then

                        file.print(vid1.position());
                        file.print(", ");
                        file.print(mse.x() / 7);
                        file.print(", ");

      file.print(response_manager.total_response_count(1));
                        file.print("\n");

                        i = i + 1.0;

                  end;

            end;

      end;

end;
```

The end of the video triggers the closing of the .csv file, the final screen of text for the participant, and allows a final button press to exit the program.

```
file.close();

loop until response_manager.total_response_count(2) == 1 begin
      finalscreen.present();
end;
```

APPENDIX 5: STUDY 2 MEANS AND STANDARD DEVIATIONS

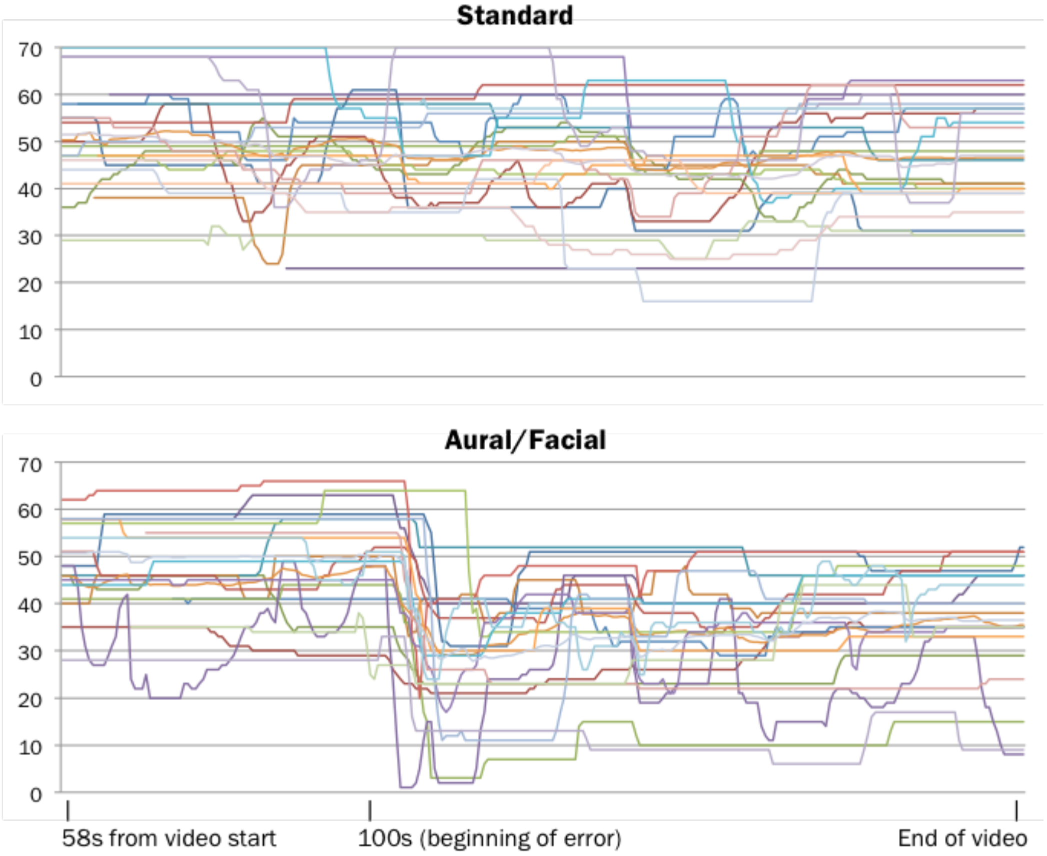
Supplementary Table 4.1 on the following page accompanies Section 4.3 of Chapter 4 (Study 2). It displays means, medians, and standard deviations drawn from the continuous measures and written scores, from which time to first and final decision, and first, final, and overall written ratings were extracted (see Section 4.2.5).

Supplementary Table 4.1. Means, medians, and standard deviations of T$_1$ (time to first rating in seconds from first note), T$_2$ (time to final rating in seconds from first note), R$_1$ (first continuous rating score on 70-point scale), R$_2$ (final continuous rating score on a 70-point scale), and R$_3$ (overall score on a 7-point scale) for the musician and non-musician groups and 5 conditions (1 = *standard*, 2 = *entrance*, 3 = *aural/facial*, 4 = *aural*, and 5 = *facial*).

		Overall (N=105)		*Musicians (n=53)*		*Non-musicians (n=52)*	
		M (Median)	*SD*	*M (Median)*	*SD*	*M (Median)*	*SD*
1	T_1	21.00 (14.75)	20.26	24.25 (18.25)	17.99	18.29 (14.75)	22.39
	T_2	123.80 (134.25)	30.44	114.00 (132.75)	38.52	131.96 (140.00)	19.87
	R_1	49.05 (48.00)	10.84	49.10 (50.00)	12.00	49.00 (46.50)	10.33
	R_2	46.82 (47.00)	11.55	46.60 (47.00)	12.90	47.00 (46.50)	10.88
	R_3	4.86 (5.00)	1.32	4.80 (5.00)	1.23	4.92 (5.00)	1.44
2	T_1	8.00 (7.00)	17.00	7.50 (4.50)	21.03	8.55 (7.25)	12.26
	T_2	123.36 (127.00)	24.00	126.86 (130.50)	23.93	119.50 (124.25)	24.74
	R_1	40.81 (43.00)	15.16	34.91 (36.00)	17.18	47.30 (46.50)	9.66
	R_2	49.86 (46.00)	8.40	49.09 (44.00)	8.14	50.70 (51.00)	9.04
	R_3	4.95 (5.00)	0.92	4.82 (5.00)	1.08	5.10 (5.00)	0.74
3	T_1	15.53 (13.00)	16.22	13.90 (12.25)	13.76	17.15 (17.50)	18.98
	T_2	139.78 (140.0)	11.79	140.25 (139.75)	12.15	139.30 (141.25)	12.06
	R_1	44.75 (42.50)	8.31	44.00 (42.00)	7.23	45.50 (45.50)	9.61
	R_2	36.00 (36.50)	13.37	35.50 (36.50)	14.75	36.50 37.50)	12.61
	R_3	3.90 (4.00)	0.97	3.80 (4.00)	1.03	4.00 (4.00)	0.94
4	T_1	18.64 (10.50)	23.45	16.46 (10.50)	25.72	21.05 (10.50)	21.78
	T_2	131.88 (141.50)	27.68	127.59 (137.00)	25.76	136.60 (147.00)	30.29
	R_1	46.86 (48.00)	8.49	48.55 (48.00)	7.54	45.00 (46.00)	9.48
	R_2	47.33 (48.00)	7.73	48.55 (48.00)	5.26	46.00 (47.00)	9.91
	R_3	5.00 (5.00)	0.55	4.91 (5.00)	0.30	5.10 (5.00)	0.74
5	T_1	18.68 (12.50)	23.06	26.85 (13.50)	29.76	9.61 (9.50)	4.95
	T_2	123.48 (126.50)	21.44	123.04 (126.50)	24.39	123.95 (122.50)	18.96
	R_1	55.00 (56.00)	9.60	55.55 (52.00)	11.07	54.40 (57.00)	8.25
	R_2	49.33 (51.00)	10.83	49.27 (51.00)	9.22	49.40 (53.00)	12.89
	R_3	5.10 (5.00)	0.89	5.00 (5.00)	0.63	5.20 (5.50)	1.14

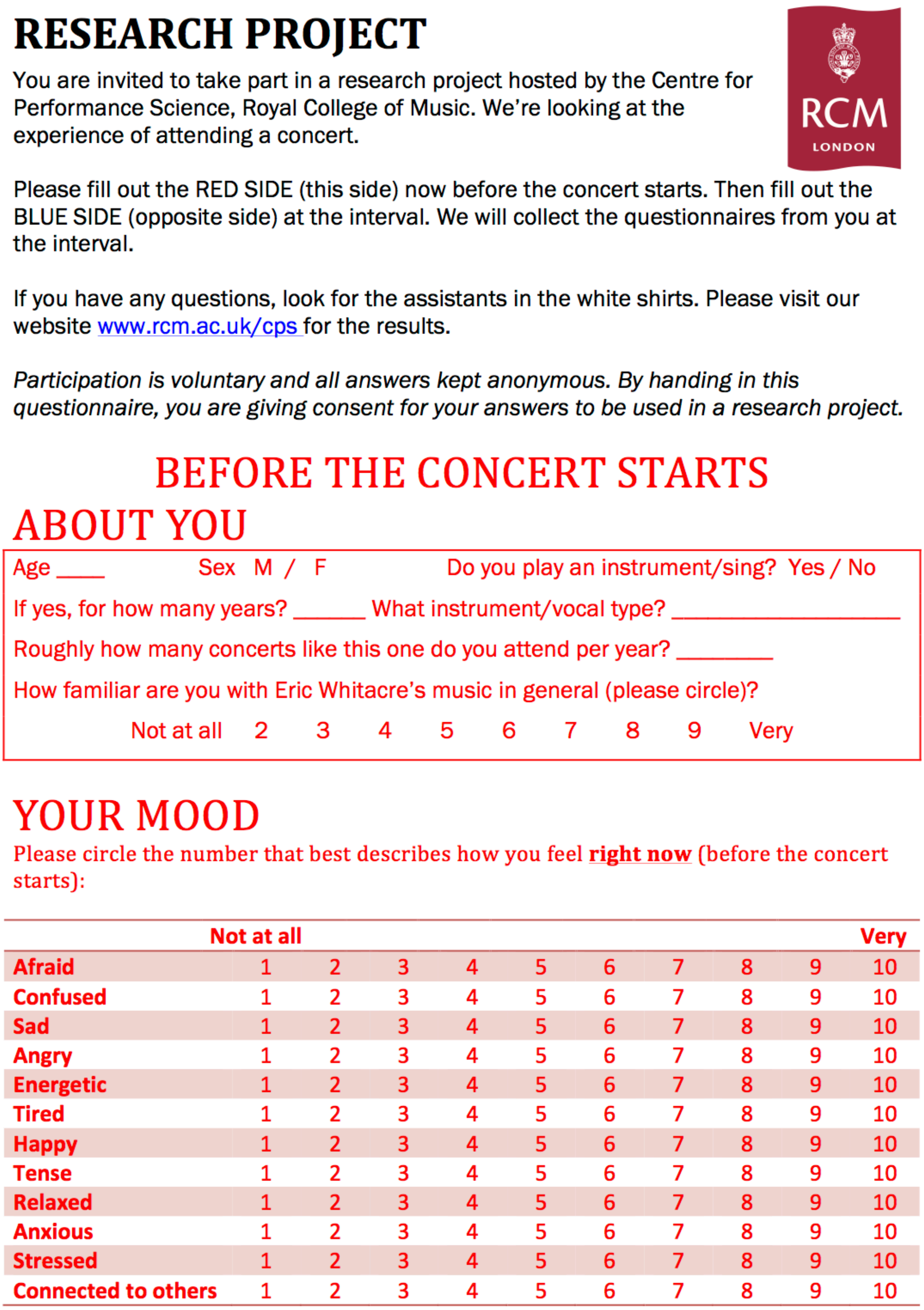
APPENDIX 6: STUDY 2 RAW CONTINUOUS DATA

The Figure below displays a sample of the raw continuous data from Study 2 (see Section 4.3.4). It shows the immediate drop in the aural/facial condition resulting from the performance error with negative facial reaction, as well as the general variability in the continuous data that was seen across conditions.


Standard
70
60
50
40
30
20
10
0
Aural/Facial
70
60
50
40
30
20
10
0
58s from video start
100s (beginning of error)
End of video

APPENDIX 7: STUDY 3 SURVEY

The custom surveys on the following two pages were developed and used for data collection within Study 3 (see Section 5.2.2). Printed on two sides of A4 paper and placed on the chapel seats prior to the concert, participants completed the first side (red) before the concert began and the second side (blue) at the interval.



RESEARCH PROJECT

You are invited to take part in a research project hosted by the Centre for Performance Science, Royal College of Music. We're looking at the experience of attending a concert.

Please fill out the RED SIDE (this side) now before the concert starts. Then fill out the BLUE SIDE (opposite side) at the interval. We will collect the questionnaires from you at the interval.

If you have any questions, look for the assistants in the white shirts. Please visit our website www.rcm.ac.uk/cps for the results.

Participation is voluntary and all answers kept anonymous. By handing in this questionnaire, you are giving consent for your answers to be used in a research project.

BEFORE THE CONCERT STARTS

ABOUT YOU

Age ____ Sex M / F Do you play an instrument/sing? Yes / No

If yes, for how many years? ______ What instrument/vocal type? ____________________

Roughly how many concerts like this one do you attend per year? _________

How familiar are you with Eric Whitacre's music in general (please circle)?

Not at all 2 3 4 5 6 7 8 9 Very

YOUR MOOD

Please circle the number that best describes how you feel **right now** (before the concert starts):

	Not at all									Very
Afraid	1	2	3	4	5	6	7	8	9	10
Confused	1	2	3	4	5	6	7	8	9	10
Sad	1	2	3	4	5	6	7	8	9	10
Angry	1	2	3	4	5	6	7	8	9	10
Energetic	1	2	3	4	5	6	7	8	9	10
Tired	1	2	3	4	5	6	7	8	9	10
Happy	1	2	3	4	5	6	7	8	9	10
Tense	1	2	3	4	5	6	7	8	9	10
Relaxed	1	2	3	4	5	6	7	8	9	10
Anxious	1	2	3	4	5	6	7	8	9	10
Stressed	1	2	3	4	5	6	7	8	9	10
Connected to others	1	2	3	4	5	6	7	8	9	10

AT THE INTERVAL

YOUR MOOD

Please circle the number that best describes how you feel **right now** (at the start of the interval):

	Not at all									Very
Afraid	1	2	3	4	5	6	7	8	9	10
Confused	1	2	3	4	5	6	7	8	9	10
Sad	1	2	3	4	5	6	7	8	9	10
Angry	1	2	3	4	5	6	7	8	9	10
Energetic	1	2	3	4	5	6	7	8	9	10
Tired	1	2	3	4	5	6	7	8	9	10
Happy	1	2	3	4	5	6	7	8	9	10
Tense	1	2	3	4	5	6	7	8	9	10
Relaxed	1	2	3	4	5	6	7	8	9	10
Anxious	1	2	3	4	5	6	7	8	9	10
Stressed	1	2	3	4	5	6	7	8	9	10
Connected to others	1	2	3	4	5	6	7	8	9	10

THE CONCERT SO FAR

How have you found the concert so far? Please circle the number that best applies:

	Not at all									Very
Stimulating	1	2	3	4	5	6	7	8	9	10
Meaningful	1	2	3	4	5	6	7	8	9	10
Enjoyable	1	2	3	4	5	6	7	8	9	10

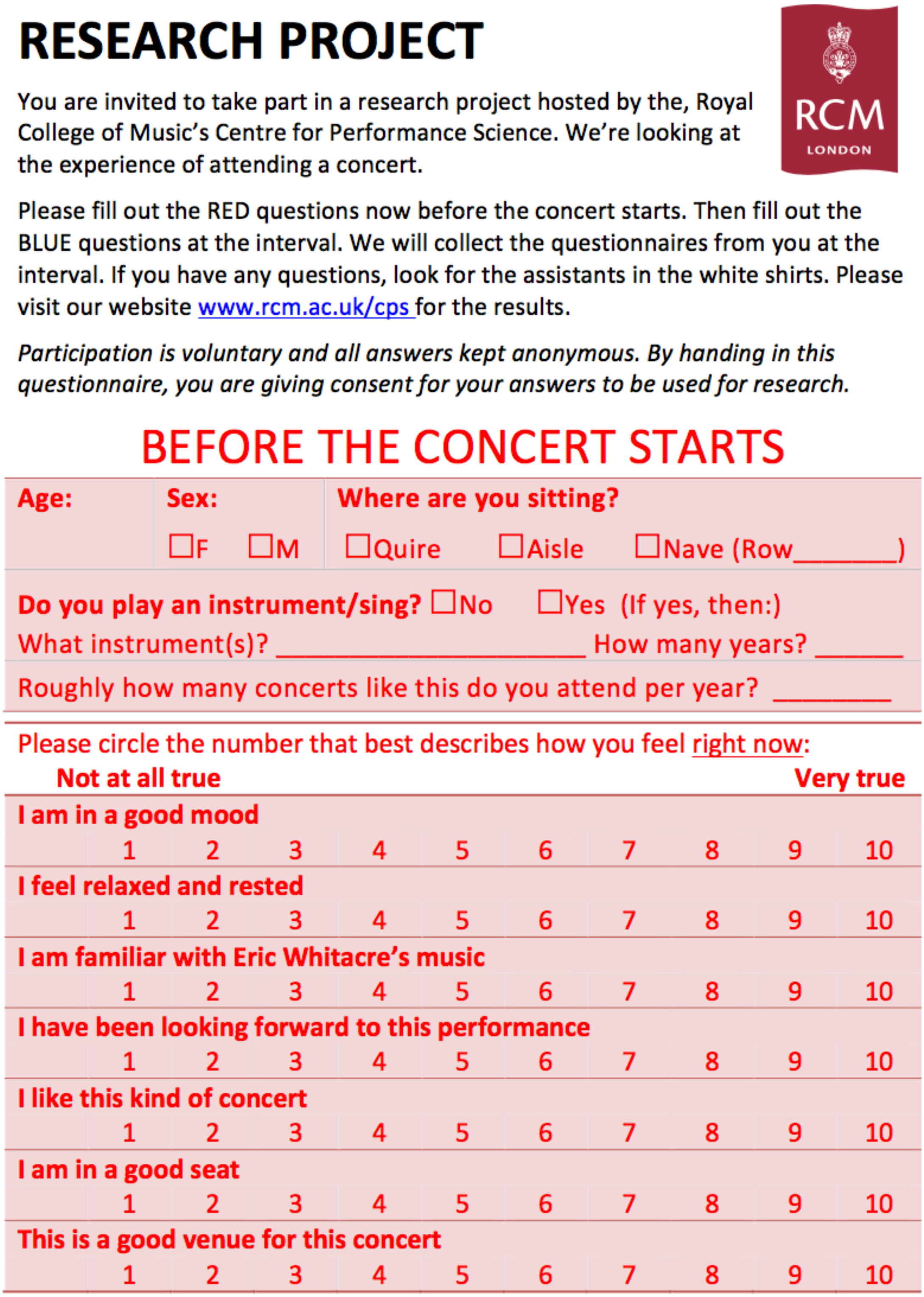
THE FIRST PIECE

Think now of the **first piece** in the programme. Please circle the number that best applies:

	Low									High
Quality of the performance	1	2	3	4	5	6	7	8	9	10
How much you enjoyed the performance	1	2	3	4	5	6	7	8	9	10
Your familiarity with the piece	1	2	3	4	5	6	7	8	9	10
How much you like this piece	1	2	3	4	5	6	7	8	9	10

APPENDIX 8: STUDY 4 SURVEY

The custom surveys on the following two pages were developed and used for data collection within Study 4 (see Section 6.2.2). Printed on two sides of A5 paper and placed on the cathedral seats prior to the concert, participants completed the first side (red) before the concert began and the second side (blue) at the interval.



RESEARCH PROJECT

You are invited to take part in a research project hosted by the, Royal College of Music's Centre for Performance Science. We're looking at the experience of attending a concert.

Please fill out the RED questions now before the concert starts. Then fill out the BLUE questions at the interval. We will collect the questionnaires from you at the interval. If you have any questions, look for the assistants in the white shirts. Please visit our website www.rcm.ac.uk/cps for the results.

Participation is voluntary and all answers kept anonymous. By handing in this questionnaire, you are giving consent for your answers to be used for research.

BEFORE THE CONCERT STARTS

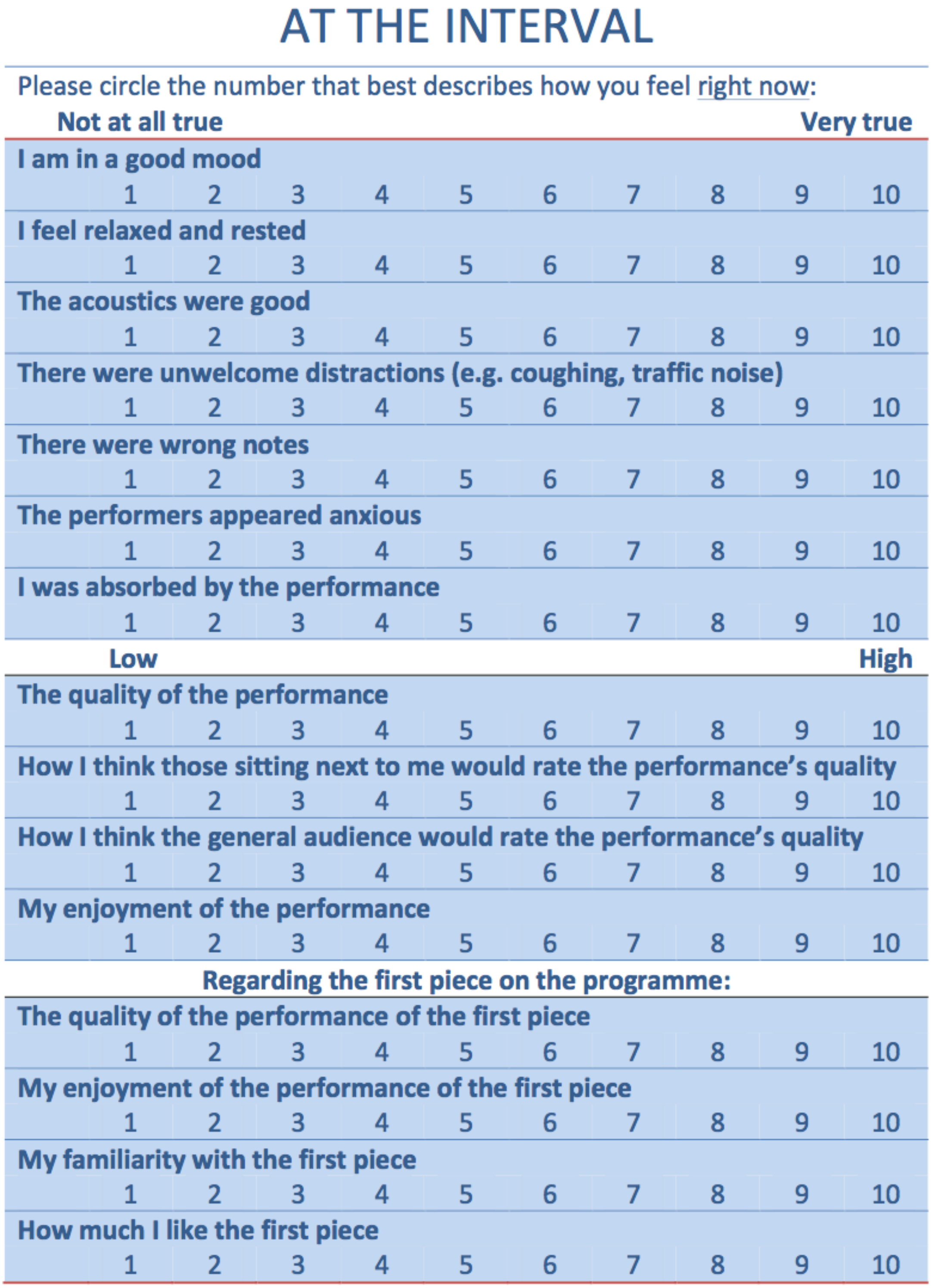
Age:	**Sex:** □F □M	**Where are you sitting?** □Quire □Aisle □Nave (Row________)

Do you play an instrument/sing? □No □Yes (If yes, then:)
What instrument(s)? ______________________ How many years? ______

Roughly how many concerts like this do you attend per year? _________

Please circle the number that best describes how you feel right now:

	Not at all true									**Very true**
I am in a good mood	1	2	3	4	5	6	7	8	9	10
I feel relaxed and rested	1	2	3	4	5	6	7	8	9	10
I am familiar with Eric Whitacre's music	1	2	3	4	5	6	7	8	9	10
I have been looking forward to this performance	1	2	3	4	5	6	7	8	9	10
I like this kind of concert	1	2	3	4	5	6	7	8	9	10
I am in a good seat	1	2	3	4	5	6	7	8	9	10
This is a good venue for this concert	1	2	3	4	5	6	7	8	9	10



AT THE INTERVAL

Please circle the number that best describes how you feel right now:

	Not at all true									Very true
I am in a good mood	1	2	3	4	5	6	7	8	9	10
I feel relaxed and rested	1	2	3	4	5	6	7	8	9	10
The acoustics were good	1	2	3	4	5	6	7	8	9	10
There were unwelcome distractions (e.g. coughing, traffic noise)	1	2	3	4	5	6	7	8	9	10
There were wrong notes	1	2	3	4	5	6	7	8	9	10
The performers appeared anxious	1	2	3	4	5	6	7	8	9	10
I was absorbed by the performance	1	2	3	4	5	6	7	8	9	10
	Low									High
The quality of the performance	1	2	3	4	5	6	7	8	9	10
How I think those sitting next to me would rate the performance's quality	1	2	3	4	5	6	7	8	9	10
How I think the general audience would rate the performance's quality	1	2	3	4	5	6	7	8	9	10
My enjoyment of the performance	1	2	3	4	5	6	7	8	9	10
Regarding the first piece on the programme:										
The quality of the performance of the first piece	1	2	3	4	5	6	7	8	9	10
My enjoyment of the performance of the first piece	1	2	3	4	5	6	7	8	9	10
My familiarity with the first piece	1	2	3	4	5	6	7	8	9	10
How much I like the first piece	1	2	3	4	5	6	7	8	9	10

APPENDIX 9: VIDEOS FROM THE EVALUATION SIMULATOR

The video recordings listed below were used to create the *Evaluation Simulator* described in Chapter 7 (see Section 7.6.2 and Table 7.1). They depict an oboe soloist taking the stage, playing one of two works (Ravel or Tchaikovsky) either well or poorly, and entering into one of three response modes in which the performer waits for feedback while appearing *confident*, *frustrated*, or *distraught*. In the simulator, these four performances can be combined with any of the three reactions for a total of 12 possible permutations. In the videos below they are presented in groupings by which they were originally captured.

These recordings can be downloaded as Supplementary Files with the following publication:

Waddell, G., Perkins, R., & Williamon, A. (2019). The Evaluation Simulator: A new approach to training music performance assessment. *Frontiers in Psychology, 10*(557), 1-17.

Sim Video A. Good quality Ravel with *confident* reaction and exit

Sim Video B. Poor quality Ravel with *frustrated* reaction and exit

Sim Video C. Poor quality Tchaikovsky with *distraught* reaction and exit

Sim Video D. Good quality Tchaikovsky