CHAPTER: MEASURING THE AUDIENCE

GEORGE WADDELL and AARON WILLIAMON
Centre for Performance Science, Royal College of Music, London UK

Introduction

Just as performance requires a performer, it also demands a listener. Research into music performance often turns its attention to those for whom music is played, how their physiological and psychological states are affected, and in some cases, how they make decisions based on what they have heard and seen. Such knowledge can inform the activities of a wide variety of music practitioners. Performers can anticipate their audiences' expectations to craft performances that are better received and assessed. Teachers can prepare their pupils for the realities of a career on the stage. Composers can gain insight into how well their programmatic or affective intents are communicated. Musicologists, psychologists, and other researchers can investigate how tastes and traditions of listening change over time and across cultures, and those concerned with social policy, whether in education, public health, or the creative and digital economies, can capture the benefits that exposure to music performance can have on people's lives. To understand the audience is to understand the impact performance has.

This chapter explores how one measures an audience through the lens of performance science. This emerging field seeks to understand and enhance the act of performance, focusing on such elements as creativity, skill, practice, teamwork, leadership, and motivation. In music, performance scientists study the realities of the performer, educator, and audience to unravel the inherent complexity of performance. Consequently, research approaches used within the field have been largely pragmatic, employing both quantitative and qualitative tools to address questions relevant to researchers and music practitioners alike. While these tools are often drawn from related scientific disciplines, researchers have begun to commandeer and create novel techniques especially adapted to the study of performance. These range from in-depth qualitative investigations of performance phenomena to the development of standardized self-report inventories and questionnaires and to the monitoring of performers' and audiences' psychological and physiological states.

Performance science embraces a range of methods discussed across the previous chapters, including naturalistic (Chapter 7) and experimental (Chapter 9) designs that may incorporate descriptive sampling (Chapter 10), the creation and use of surveys and questionnaires (Chapter 10), and the application of various statistical models (Chapters 11 and 12). These approaches must then be applied to highly complex subject matter: music performance, which involves an intricate combination of processes compounded by the historical, cultural, and theoretical richness of the repertoire being performed. Furthermore, performances often occur within rich environments comprising specific venues, audiences, tools, and team members. Transferring these activities to the laboratory can alter the very nature of performance; thus, researchers are constantly striving to improve the ecological validity of their research (see Chapter 7) and ensure that the task they are measuring is as authentic to the true performance as possible. As a result of these challenges, a significant catalogue of methods has been developed and employed. While an exhaustive summary of these is beyond the scope of a single chapter, a sample from across the field of music performance science is presented here as case studies into how audiences think and behave and how this can be measured.

Audiences' experiences of performance can fall within a wide spectrum of activities, ranging from passive listening to explicit evaluation of an experience. This chapter first provides examples of how affect and arousal can be altered by performances, both in the concert hall and when listening to recordings. These reactions may be measured through audience members' self-reports or by measuring their bodies' physiological responses. The chapter then considers audiences' evaluations of performance quality–where the public, evaluators, teachers, critics, and judges make critical decisions regarding the quality and value of a performance–and the role that the evaluation criteria, as well as factors beyond the musical material, can play in these decisions. The worlds of affective and evaluative response are then brought together in two unique areas of study that consider musical performance in a wider sense: the visual effect of the performer's appearance and behavior on stage, and how responses can be measured continuously as the music unfolds.

Measuring Affect and Arousal

That emotional and physiological reactions can be elicited from musical experiences is unquestioned. The performer is often considered a conduit of expressive intent from the composer to the audience, although the nature of this transfer of expressive information is not well understood. By studying this process, performers and educators can better understand how their actions translate to audiences, facilitating deeper engagement with and enjoyment of their performances (assuming that is their intent).

In terms of research on music and emotion, the majority of studies have been conducted using self-report tools employed during and following a performance. Methods have ranged from open-ended responses where listeners provide their own emotional descriptors, to scales where they rate the intensity or applicability of a specified dimension (e.g. happiness or tension), to strict forced choices where they must choose the best-suited descriptor from a list. Examples of open-ended responses and emotion scales can be found in a study by Evans and Schubert (2008) where participants listened to experimenter-selected music and imagined their own performances, then provided descriptions of what emotions they felt and why, and completed 11-point scales of valence (e.g. happy-sad), arousal (e.g. excited-calm), emotional strength (i.e. none-strong), and dominance (i.e. dominant-submissive). These data were used to compare whether a piece's perceived *expressed* emotion matched the listener's *felt* emotion, which was found to occur in 61% of cases and led to greater listener enjoyment. An example of a forced choice model is a study by Quinto and colleagues (2014) where listeners were asked to identify whether excerpts conveyed anger, fear, happiness, sadness, tenderness, or nothing (neutral) after listening to performances composed and performed with the specific intent of conveying those emotions. The composition was found to be the better medium to express fear, anger was better expressed through the performance, and happiness and sadness were both found to transfer equally well through both composition and the performance.

In terms of research on music and physiological arousal, the effects on the body and brain have been discussed at length in both research and popular writing. The relation of neural activity in response to music listening has been explored through, for example, electroencephalography (EEG) to measure brain activation as a result of musical accents (Palmer *et al.* 2009) or listening to compositions with different emotional intent (Khalfa *et al.* 2002), and functional magnetic resonance imaging (fMRI) to show different processing patterns when listening to emotionally expressive performances versus a comput-

er-controlled mechanical performance (Chapin *et al.* 2010). Researchers have also examined automatic responses of the body. For example, Egermann and colleagues (2011) measured skin conductance (an indicator of arousal), showing that listeners experienced more emotional arousal and musical "chills" when listening to music alone than when in a group. Research into the health implications of music listening also employs such physiological approaches, such as one study by Koelsch *et al.* (2011) that demonstrated how listening to instrumental music before surgery can result in lower cortisol levels (a stress hormone that can be measured in both blood and saliva) and lower requirements for sedatives.

In each of these cases, the effects of music were studied in the laboratory or a clinical setting. Investigating the influence of live music performance on the body in real-life performances is more difficult. Tight experimental control, as is so often required in applied physiological research, is not conducive to the concert hall, where precise timing and a distraction-free environment are usually key to the experience. However, attempts are being made to bring such analysis to live concert settings. A recent study by Fancourt and Williamon (2016) recruited 117 concertgoers across two professional choral concerts to provide saliva samples and complete a questionnaire before the concert began and at the intermission. This method allowed the researchers to investigate the effect of the live performance on hormones measured from saliva with minimal disruption to the participants' experience of the music. There was a significant drop in stress hormones across the performances, demonstrating a positive and relaxing impact of attending the event. These reductions were stable across participants despite differences in familiarity with the repertoire, musical ability, and age.

Measuring Quality

Often, musical performances are compared with one another in terms of quality, where one performance (and often by extension one performer) is determined to be better. Studying this act is important in performance research for three reasons: performance quality evaluation forms a major part of the career of practicing musicians, from teacher feedback to auditions into educational and professional institutions to career-boosting competitions; many performers will be called upon to conduct evaluations in their careers (i.e. as an "expert" evaluator) and thus should develop the skills necessary to make

such decisions; and much research focusing on the performer incorporates performance quality as a dependent variable, where interventions, behaviors, thoughts, traits, or states are examined for a relationship with the quality of the resultant performance (for a review, see McPherson and Schubert 2004).

With each use of such evaluations, Thompson and Williamon (2003, pp. 22-24) identify three assumptions that are typically made (1) performance quality is a dimension with a common psychological reality for experienced listeners; (2) experienced musicians are able to offer consistent judgments of music performance quality; and (3) experienced musicians are able to distinguish between aspects of a performance such as technique and interpretation. Recent research into performance quality calls into question these assumptions, testing the musical criteria themselves as well as what extra-musical factors may influence the evaluation.

Criteria

As in studies of emotional and enjoyment reactions to music, determining the criteria on which feedback is collected is key. Two general approaches are used in both musical and research settings (Mills 1991). *Holistic,* or global, assessments consist of a single, overall rating to encapsulate the quality of a given performance: i.e. the classic "eight out of ten" or "Grade-A" performance. In practical terms the advantages are clear: a single score allows for easy comparison between performances and performers, giving evaluators freedom to employ their own criteria and weighting of specific points. The strengths of the holistic rating in flexibility and adaptability weaken its reliability. The ratings of multiple performances by a single evaluator may be comparable; however, without 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 judgments.

Segmented assessments break the ratings into specific categories, often divided into thematic groupings and totaled 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 judgment. 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 coloring overall technique, creativity, and interpretation. Mills (1991) acknowledged that the trend in musical academia showed a shift from holistic to segmented assessments but emphasized the need for careful consideration of its makeup, warning that "introduction of a segmented system with arbitrary weighting does not remove the problem: it only hides it" (p. 174).

Thompson and Williamon (2003) examined the utility of a segmented measurement scheme using 13 criteria over three general categories (perceived instrumental competence, musicality, and communication) plus an overall quality mark, each assessed using a scale of one to ten. Three expert evaluators assessed 61 video recorded performances of varying instruments. Analyses showed that the three general categories were able to predict a high degree of variance in the final mark (approximately 90%). However, the correlations between the three general categories were strong, questioning the assumption that separately graded aspects of the performance can be differentiated.

The criteria that contribute to formal evaluations have also been examined in qualitative and survey-based examinations. Davidson and Coimbra (2001) observed panel evaluations of 21 singers undergoing mid-term performance assessments and observed the discussions that resulted with and without the musician present. They found that the examiner's grades reflected the points they discussed, and, interestingly, aspects of the singers' appearance were taken into consideration with their vocal control in assessing their expressive abilities (see "Measuring visual responses" below). Alessandri and colleagues (2014) examined the nature of and factors contributing to critical recording reviews through an exhaustive survey of Beethoven piano sonata reviews published in *Gramophone* and statistical analysis of the resulting metadata (i.e. length, structure, reviewer, pianist, and work in question). They found that reviews concentrated on a small number of performers by a select group of critics, and that comparisons to established performances were commonly used.

Extra-musical Influences

The examinations of performance criteria usually focus on musical parameters; i.e. elements of technique, expression, musicality, etc. with which most musicians are familiar. However, these criteria do not acknowledge the variety of possible "extra-musical" fac-

tors–variables that are not intended to play a role in a purely "musical" assessment–that may be influencing performance decisions. This might include the order in which performances are experienced or the experience of the judges themselves. Thus, numerous experimental methods have been used to isolate these variables in laboratory and competition settings.

In one study, Fiske (1975) examined the role of evaluator expertise in evaluations, distinguishing between *experts* who had a great deal of experience in music performance, evaluation, and teaching, and *specialists* 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 that they were in fact hearing 64 unique performances. This approach allowed for an examination of test-retest reliability for each judge. Ratings were collected on a five-point scale across five categories: intonation, rhythm, technique, interpretation, and overall. Judges were found to be moderately consistent (though far from perfect) when hearing the same work twice, with no difference between specialists or non-specialist experts in reliability or how high they ranked individual aspects of performance. However, when specialists were defined more widely as wind players (the recordings were of trumpeters) they provided higher technique scores than the non-wind players.

Flôres Jr. and Ginsburg (1996) examined whether the final ranking of performers in the Queen Elisabeth competition correlated with the day on which the candidate performed. The rankings of the 12 semi-finalists in 21 competitions (from 1951 to 1993; 120 violinists and 132 pianists) were aggregated. As the performance order of the 12 performers (two per day over six days) was randomly chosen, the null hypothesis stated that each permutation of rankings in 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 trough 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 formalizing 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).

Measuring Visual Responses

In both affective and evaluative music situations, the musical expression and extra-musical factors are often presented visually. This extra information can have a profound effect on performance reactions, and in the case of evaluative responses, it remains debatable whether these visual factors fall into the "musical" or the "extra-musical" categories described above. While the rating scales and methodologies employed in audio-only studies may be used in this research, the visual nature requires special consideration in preparing the stimuli, often in the form of manipulated audio and video recordings.

Drawing from research on human motion, gymnastics, dance, and acting, Davidson (1993) 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). 21 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 the participants could not only identify the differences in expressive intension by movement information alone but also rated a stronger difference between the most- and least-expressive performances when they were presented with visual-only information than with audio-only information. The audio-video condition ratings were in the middle, indicating that the audio information may have been tempering the reaction to the visual information.

Elliot (1995) examined whether visually presented racial differences would influence the evaluations of experienced musicians. Four trumpeters (an instrument carrying masculine associations) and four flautists (female) were video recorded. Each group consisted of 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. 88 music education majors evaluated the tapes, with the performance order randomized and the ability to delay evaluation until performances of each instrument were viewed.

A similar method was used in a series of studies by Griffiths (e.g. 2010) to examine the effects of concert dress. By dubbing the same audio track over the video recordings of a performer wearing different clothing (e.g. jeans vs. formal concert attire), she was able to show that clothing worn by female soloists significantly influenced the listener's perception of performance quality in the evaluation of dubbed video recordings.

Studies have looked at visually presented pre-performance rituals, such as one that 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 the viewers' wish to continue observing the performance (Platz and Kopiez 2013). Others have examined how visual information contributes to intuitive judgments of performance. Tsay (2013) found that, while novices or experts could not reliably predict the winner of an international piano competition based on six-second audio or audio-video recordings, silent video recordings demonstrating only the musician's physical behavior allowed the winner to be chosen at a rate greater than chance regardless of the evaluator's experience.

Measuring Continuous Responses

The majority of the studies on evaluation and experience employed methods for collecting data *post hoc*, self-reported assessments provided by the listener following the experience of a performance. While this technique allows for simple descriptions and comparisons of an event, it does not consider that performances take place over time. Listeners' opinions, attitudes, and emotions can shift dynamically with the music. Thus, continuous response measurements have been adapted and developed to allow researchers to track these changes over the course of a performance, revealing the cognitive processes behind decisions and linking specific reactions to particular musical events.

An early form of this involved a "method of continuous judgment by category" that was applied to the musical performance with the addition of a "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 (Namba *et al.*, 1991). The frequency of temporal use of each judgment correlated with the adjectives chosen to describe the overall impression of the work, with the authors hy-

pothesizing that overall impression is based on a weighted average of temporal impressions.

Continuous measurements methodologies have been aided by several technologies developed specifically for use in musical studies, in particular the Continuous Response Digital Interface (Geringer *et al.* 2004), where dimensions can be applied to a dial or physical slider, and the Continuous Response Measurement Apparatus (CReMA) (Himonides 2011), where participants can track a finger across a horizontal bar which simultaneously detects pressure. MIDI devices, normally used for the performance of and interaction with musical stimuli, have also been employed, as well as bespoke computer software. These tools have been used to examine listeners' preferences, perception of loudness and phrasing, focus of attention, perception of musical intensity, perceived tension, perceived expressivity and aesthetic and emotional responses in relation to musical stimuli as they change over time, often comparing them to overall ratings (for a review, see Geringer *et al.* 2004).

In a few cases, continuous measures methodologies have been applied to music quality evaluations. Himonides (2011) conducted a pilot study with the custom CReMA device and examined quality ratings of sung vocal performances, including criteria such as diction, dynamics, and vibrato, and compared their responses to physiological data (i.e. heart rate, and skin conductance). Another application was conducted by Thompson and colleagues (2007) in which a customized piece of software was created to allow for continuous data to be collected by moving a mouse pointer across a horizontal bar. Two pianists each audio-recorded contrasting performances (slow, natural, and fast) of two works, resulting in a total of ten performances (one pianist's fast recordings were discounted as unrealistic). 33 participants were then divided in to three experimental groups that evaluated either each performance's overall quality, technical proficiency, or musicality both continuously using the software and as an overall judgment with written scales following the performance. This methodology allowed the researchers to show that listeners took an average of 15 seconds to reach an initial evaluative decision, that their decisions changed approximately 2.6 times per minute, and that a final decision was reached by approximately 60 seconds into the performance. They also found systematic differences between how assessments of technicality, musicality, and general quality were produced over time.

Summary

The field of performance science draws upon a wide variety of methodologies to examine the nature, practices, and cognitions of performers' audiences. Without such knowledge our understanding of performance remains in a vacuum, without reference to the effect it has on the world. This chapter has provided a sample of techniques used to measure the audience, ranging from in-depth qualitative case studies to experimental designs involving customized tools, scales, and stimuli. Affective reactions and arousal can be measured using both self-reports and studies of physiological response, the latter of which are beginning to show the benefits of attending live performances. Studies involving evaluations of quality reveal the subjective nature of decisions that audiences and judges make with every performance, considering the relationship between holistic and segmented criteria as well as the effects of extra-musical variables. The influence of the visual aspects of performance on audience reactions can be studied through responses to carefully manipulated audio and video recordings. Due to the temporal nature of performance, continuous measures techniques can be used to map both evaluative and affective responses across an entire performance. With the knowledge gained from this research, performances can be crafted and presented that intentionally affect, move, and drive decisions in audiences, and the full impact of music performance on society can be understood.

References and Recommended Futher Reading

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), 131−49.

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.

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. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, *21*(2), 103−13.

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), 307−23.

Elliott, C. A. (1995). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education*, *127*, 50−6.

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

Fancourt, D. & Williamon, A. (2016). Attending a concert reduces glucocorticoids, progesterone and the coritsol/ DHEA ratio. *Public Health, 132,* 101−4.

Fiske, H. E. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education*, *23*(3), 186−96.

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

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.

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), 159−77.

Himonides, E. (2011). Mapping a beautiful voice: The continuous response measurement apparatus (CReMA). *Journal of Music, Technology and Education*, *4*(1), 5−25.

Khalfa, S., Isabelle, P., Jean-Pierre, B., & Manon, R. (2002). Event-related skin conductance responses to musical emotions in humans. *Neuroscience Letters*, *328*(2), 145−9.

Koelsch, S., Fuermetz, J., Sack, U., Bauer, K., Hohenadel, M., Wiegel, M., Kaisers, U. X., & Heinke, W. (2011). Effects of music listening on cortisol levels and propofol consumption during spinal anesthesia. *Frontiers in Psychology*, *2*, 58.

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, UK: Oxford University Press.

Mills, J. (1991). Assessing musical performance musically. *Educational Studies*, *17*(2), 173−81. 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, 8*(3), 251−75.

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*, 470−80.

Platz, F. & Kopiez, R. (2013). When the first impression counts: Music performers, audience and the evaluation of stage entrance behaviour. *Musicae Scientiae*, *17*(2), 167−97.

Quinto, L., Thompson, W. F., & Taylor, A. (2014). The contributions of compositional structure and performance expression to the communication of emotion in music. *Psychology of Music*, *42*(4), 503−24.

Thompson, S. & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception*, *21*(1), 21−41.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Perception*, *25*(1), 13−29.

Tsay, C. J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences USA*, *110*(36), 14580−5.