



Assessing Music Performance: Issues and Influences

Gary E. McPherson & William F. Thompson

Abstract

Assessing musical performance is common across many types of music education practice, yet research clarifying the range of factors which impact on a judge's assessment is relatively scarce. This article attempts to provide focus to the current literature, by proposing a process model of assessing musical performance that identifies some of the main elements that affect a judge's assessment in formal performance settings such as competitions, auditions, recitals, Eisteddfods and graded examinations. The article includes a review of the literature according to the categories defined in the model and suggestions which are intended to form the basis for further research in the area.

Music performance assessment is the process by which one individual attempts to balance and synthesise the various qualities of a performance by another individual, with the aim of providing a judgement, such as a ranking, grade or qualitative description. While we might like to think otherwise, the assessment of music performances by adjudicators and teachers is not without difficulties; reliability among assessors is sometimes low and significant biases often influence the results. Studies of formal music assessment have involved many different aims, contexts, and evaluation strategies. They demonstrate that the formal assessing of music performance may be conceptualised as a complex system comprising numerous interrelated influences.

With this idea in mind, our article will address some of the main issues concerned with assessing musical performances. We begin by proposing a process model (McPherson, 1996) that has been influenced by research in other areas (Landy & Farr, 1980). This model is used to group the available literature in a way that will allow us to discuss some of the most important factors which we believe impact on the assessment of musical performance in formal settings.

A Process Model of Assessing Musical Performance

Figure 1 provides a schematic representation of the primary issues surrounding performance assessment. The model illustrates a complex set of interacting factors that affect performance and assessment, including context, musical and nonmusical factors, evaluation instruments and/or criteria, performer and evaluator characteristics, and feedback to the performer. The latter influence underscores the dual role of assessment as both a description of performance and a guide for improvement.

Performance context represents a central influence on assessment and includes at least four factors. First, the *purpose of the assessment* - whether the musician is performing in a music competition, festival, end of semester recital/examination, auditioning for an ensemble, or even participating in a music research project - strongly influences the way a judge will listen to, and therefore, evaluate a musical performance.

Second, the *type of performance* that is being assessed will affect judgements. McPherson (1993, 1995a) proposed five distinct types of musical performance: sightreading, performing rehearsed repertoire, playing from memory, playing by ear and improvising. He asserts the importance of employing different strategies to assess each of these styles of performance (McPherson, 1995a, 1995b) because the evaluative criteria needed to make critical judgements vary so much between each of the five abilities. Other researchers have attempted to design specific measures for each type of skill. For example, Lidral (1997) described a contest in solo jazz improvisation, and the assessment

tool employed, called the Jazz Improvisers' Adjudication Form (JIAF). Welch (1994) reviewed theoretical issues in the assessment of singing.

Performances on different musical instruments may also be assessed differently because they involve different technical skills and are associated with different repertoire. Performance measures directed at specific types of instruments include the Watkins-Farnum Performance Scale for wind instruments (Watkins & Farnum, 1954), the Clarinet Performance Rating Scale (Abeles, 1973), and the Brass Performance Rating Scale (Bergee, 1988, 1989).

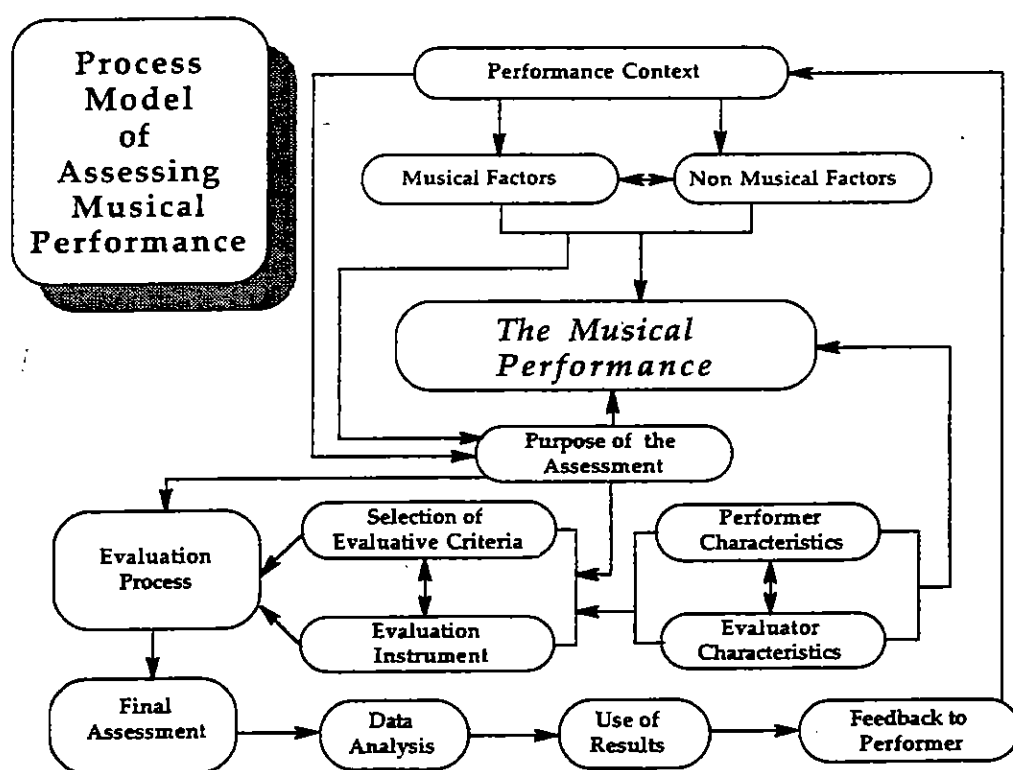


Figure 1: Process Model of Assessing Musical Performance (adapted from Landy & Farr, 1980)

Bias effects may be more or less evident depending on the type of performance assessed. For example, Radocy (1976) asked undergraduate music students to assess a number of performances. Students were assigned to one of five 'bias' conditions, ranging from no bias, to strong bias. Bias was manipulated by introducing false information about performers or composers prior to assessments. Students in the no-bias condition were told nothing about the performers. Students in moderate bias conditions were given misleading information about the performers. For example, one performer might be labelled 'former symphony player' while another might be labelled 'young graduate assistant.' Students in the strong bias condition were given misleading information about the performers, and were also given reasons why the performances by the professional performer were supposedly preferred by prior listeners. Radocy observed an overall effect of bias, but found that performances in some media (e.g., piano) were more susceptible to these bias effects than performances in other media (e.g., trumpet, orchestra).

Third, *performance proportions* (whether the performance involves a soloist or ensemble) affect assessment both by shaping the aesthetic goals of performance, and by constraining extramusical influences on assessment. For choral competitions, research suggests that large groups should restrict the degree of physical movement to avoid appearing too 'busy' on stage, whereas smaller groups can afford a larger range of movements during a performance (Morgan & Burrows, 1981). Such observations illustrate that extramusical influences on assessment vary with performance proportions.

Finally, both performance and assessment may be influenced by the *performance environment*, such as the size and acoustics of the performance space and the equipment available for the performer. In particular, balance problems may arise when voices or instruments are not positioned at equal distances from a microphone (Morgan & Burrows, 1981). These, together with a variety of other technical problems can have a marked effect on the outcome of any assessment.

The next two factors in the model concern the musical and non-musical factors which influence performance and assessment. **Musical factors** include choice of repertoire, the form and structure of the music, the size of ensemble, the skill of accompanying performers, and type of instrument. Different expressive devices, such as rubato and pedaling, are evaluated differently for each composer and historical period. Performance expression may also reflect a performer's interpretation (implicit or explicit) of formal and structural elements in the music (Berry, 1989). If adjudicators share (or are convinced by) this interpretation, their assessments may be more positive. Finally, as discussed above, performance assessment often depends on the instrument involved, such as brass (Bergee, 1988, 1989, 1992), strings (Reed, 1990), clarinet (e.g., Abeles, 1973), singing (Welch, 1994), and choral singing (e.g., Cooksey, 1977; Morgan & Burrows, 1981).

A variety of **nonmusical factors** can have a profound influence on the assessment. When nonmusical influences are unplanned, they may be considered bias effects. For example, Flores & Ginsburgh (1996) reported evidence of a significant bias in the Queen Elisabeth musical competition. They examined the results of ten violin competitions involving 120 performers, and eleven piano competitions involving 132 performers. The final rankings of performers were plotted against their order of appearance and an analysis was conducted to determine the relationship between these two variables. The rankings of performers were significantly related to the order in which they performed: those performing near the beginning had a lower chance of being ranked among the top, and those performing near the final day of the competition had a higher chance of being ranked among the top.

This bias was evident for both piano and violin competitions, but was most strongly evident for the piano competitions. The authors suggested that repeated hearings of the same piece may nurture a greater appreciation of that piece, which, in turn, might cast a more positive light on the performance itself. Another possibility is that adjudicators tend to start with "higher expectations and more strict rules, and then progressively adapt them to the reality of the actual performances." (Flores & Ginsburgh, 1996, p. 102).

Apart from the 'order effect' outlined above, there are a number of other non-musical factors that can effect a performer's ability to play the repertoire, but also a judge's perception of the worth or otherwise of their efforts. Among these, the interaction of a performer with other people, such as their peers or teacher, other musicians, or even the stage staff can help or hinder a performance, while all sorts of distractions before or during a performance can have an important impact on the quality of a musician's playing. As many performers will testify, one of the most frustrating things that can happen for a musician is to have equipment fail during the performance.

The most critical determinant of any assessment is the quality of the **musical performance**. This is shown at the heart of our model, and is used here to refer to the

actual act of performance. However, the model shows that there are many other influences that impact on a performance, and on its evaluation by assessors.

Apart from the skill and experience of the performer, it is important to consider the personal characteristics of the performers. Of critical importance is their actual technical and musical skill as a musician. However, one factor which is often overlooked concerns the performer's own thought processes immediately before and during their performance. This includes whether the musician feels nervous during their performance (i.e., level of anxiety), whether they feel they are capable of succeeding in their performance (i.e., self-efficacy), whether they believe their success or failure is a result of their own ability, good luck, sustained effort or the difficulty of the exam or repertoire (i.e., attributions), and also the extent to which they are able to maintain their concentration and approach the performance in a positive manner (i.e., self-regulation and motivation). While it is true that many young performers thrive in an environment which includes opportunities to perform for Eisteddfods and for graded music examinations, it is equally true that not all students are suited to or motivated by a music education that includes an emphasis on formal assessment. McPherson (1989) argued that cognitive mediational processes such as those outlined here affect a performer's perception of the task. Literature (McPherson, 1989; Austin & Vispoel, 1992; Brodsky, 1996) suggests that how musicians think about themselves, the task and their performance may be just as important as the time they spend practising and preparing for their performance.

Other important performer characteristics include the consistency with which the performers have prepared for their performance, how familiar they might be with the setting (i.e., equipment, room, acoustics), and whether they have previous experience in similar performance situations. Of the multitude of additional performer characteristics, dress and appearance, and stage presence and personality, as expressed in the performers' ability to take risks and/or perform in a confident and outgoing manner would also appear to have a profound effect on the perceived quality of a performance. At present, no research has been conducted that sheds light on these factors.

Interestingly, research has not established whether there are gender differences in performance style, although gender differences have been reported in studies of musical development and instrument choice (e.g., Abeles & Porter, 1978; Bruce & Kemp, 1993; Kemp, 1982; MacKenzie, 1991; Zervoudakes & Tanur, 1994). A finding, however, that provides some cause for concern was undertaken by Elliott (1995) who examined the possible effects of gender and race on judgements by 88 undergraduate and graduate music education majors. Four flautists (male/female, white/black) and four trumpeters (male/female, white/black) were videotaped performing the same *étude*, although the soundtracks for all four flautists and for each of the four trumpeters were identical. Results showed that black performers scored significantly lower than their white peers, with the female trumpeters scoring lower than the female flautists. For Elliott, these results:

support the body of research indicating that masculine/feminine associations for certain musical instruments are rather strong and that prior expectation can influence how even experienced musicians hear and judge musical performances. This should be of considerable concern to anyone devoted to bias-free music education, and it suggests that sensitivity to such racial and gender bias should become a part of the training required of all aspiring musician/educators. (p. 53)

Characteristics of the evaluator strongly influence the outcome of any assessment, and include personality, experience and musical ability, training in adjudication, familiarity with the performer, and familiarity with the repertoire. The gender of the assessor does not appear to be a strong predictor of musical preferences (Bradley, 1972; Johnson & Knapp, 1963) and there is no evidence that gender plays a large role in assessment. However, musical preferences may be strongly shaped by the

performer's or assessor's personality (e.g., Catell & Anderson, 1953), and emotional state (e.g., Cantor & Zillman, 1973).

If adjudicators are uncertain about their ability to provide an objective assessment, they may be more susceptible to bias effects. Duerksen (1972) found that judgements are generally more critical if the purpose of an assessment is to evaluate a student rather than if the purpose is to evaluate a professional musician. In his study, music majors tended to score recorded piano performances lower when they were told that the performance was a 'student', rather than when they were told the performer was a 'professional'. Similar evaluator bias effects have been reported by Radocy (1976).

Although teachers often view peer and self-evaluation as a valuable component of improving performance (Bergee, 1995), Hunter and Russ (1996) reported that, without training in assessment, students tend to form 'distorted' opinions of their peers' abilities (p. 67). That is, their assessments of their peers tend to be unrealistic. In this study, students were reluctant to award low marks to peers, and their marks frequently reflected their expectations about the performer's musical ability rather than the quality of the performance itself. With training, however, students were able to refine their ability and provide more objective assessments. Winter (1993) also reported positive effects of training in adjudication and suggests that training programs are even more important than experience in adjudication (see also, Bradley, 1972). These studies suggest that assessment strategies may be developed through training.

In contrast, Fiske (1978) reports that the performance ability of adjudicators, their musical knowledge, and their training in adjudication, do not guarantee above average rater reliability. His studies (Fiske, 1977, 1978) do not support the view that the ability to evaluate reliably is acquired through training and experience. In short, studies of the benefits of training for assessment have yielded mixed results.

Assessment strategies also depend on the adjudicator's familiarity with the repertoire. Different criteria may be used for familiar works than for unfamiliar works, such as new works commissioned specifically for a competition. Assessing performances of unfamiliar works is difficult, and, hence, subject to bias effects. For example, judges may be more severe for the first few performances of an unfamiliar work, with the consequence that musicians who perform early on are disadvantaged (Flores & Ginsburgh, 1996). One way of addressing this problem is to encourage adjudicators to familiarise themselves with new repertoire prior to the competition.

Other evaluator characteristics that have received little research attention include the aesthetic appeal of the repertoire being performed. Folklore would suggest that judges who prefer certain pieces over others, or believe that they are musically more challenging or worthy will tend to score performances of these works higher. However, no studies have been found that tackle this important component of assessment.

As noted earlier, the personality, mood and attitude of the evaluator is of utmost importance. Included here are factors such as personal biases (some of which were discussed previously) through to how familiar the evaluator is with the performer. One's knowledge of a musician from previous performances can leave a lasting impression on how s/he will be rated in subsequent performances. In a graded music examination, the personal attributes of the evaluator are of critical importance, because an examiner who appears to be friendly and encouraging is more likely to put a performer at ease compared with an examiner who acts impatient and aloof. Depending on the time of day and/or other personal and social distraction, an evaluator's mood might change markedly. Fatigue may also cause significant fluctuations to assessments during the course of a long day of examining, or adjudicating a competition or music festival. Research has yet to demonstrate the extent to which each of these can influence a judge's assessment.

In an environment of 'equal opportunity', most orchestras and professional ensembles now have 'blind' auditions for the first stage of selecting members of their ensemble, in which the selection committee is placed behind a screen and in a position

where they cannot see the performer. This procedure is also common in many band competitions and Eisteddfods internationally. But again, the advantages and disadvantages of applying such a system has received little research attention, and we cannot be sure that the benefits outweigh the negatives in all situations.

The purpose of the evaluation determines the choice of criteria and/or evaluative instruments. Appropriate criteria differ depending on whether the purpose of the evaluation is pedagogy, research, or ranking performers in a competition or audition. For example, a portfolio of taped performances may be appropriate for assessing students in a music program (Goolsby, 1995), but impractical for a performance competition. When music is performed in a non-competitive context, it is evaluated differently from when it is performed for competitive performance goals (Sheldon, 1994). Even within a given assessment context, judges may interpret the purposes of assessment differently. For example, students and teachers often view the purpose of pedagogical assessments differently, and therefore form different opinions about performances (Hunter & Russ, 1996).

The selection of evaluative criteria and evaluative instruments strongly shape the evaluation process. There are two primary aims in the development of standardised criteria and instruments: to improve the validity of assessments and to improve the reliability of assessments (Lehman, 1968; Whybrew, 1971). Validity is the degree to which an assessment tool measures the variables that it is intended to measure (e.g., performance excellence, sight-reading, improvisation ability). Reliability is the degree of correspondence between different judges (inter-judge reliability), or the consistency with which a given judge evaluates a set of performances on different occasions (intra-judge reliability). The aims of validity and reliability are somewhat independent: increased validity need not lead to increased reliability, and vice versa.

Statistical procedures may be adopted to emphasise validity (for a review, see Zdzinski, 1991). For example, Abeles (1973) used factor analysis to develop a rating scale for clarinet performance. Other techniques of assessment that emphasise validity include the method of continuous judgement, in which judges repeatedly provide judgements as a performance unfolds (Namba, Kuwano, Hatoh & Kato, 1991), and the repertory grid technique, which identifies and maps the unique set of personal constructs that each judge employs when assessing a performance (Thompson, Diamond & Balkwill, in press).

Estimates of judgement reliability are highly variable, making it difficult to evaluate the benefits of standardised evaluative tools. Reliability studies vary widely both in the range of performances and judges involved, and in the statistical procedures employed. Not surprisingly, when there is a large range in the quality of performances, interjudge reliability is high; when that range is restricted, as in a prestigious music competition, there is greater potential for disagreement.

It is generally acknowledged that in some assessment contexts, reliability is poor. Fiske (1978) presented a set of performances two times each to experienced musicians. These adjudicators were not aware that the performances were repeated, and therefore provided two judgements for each performance. Fiske reported alarmingly low correlations between the first and second set of judgements, with some judges actually showing a negative correlation between ratings of performances after the first and second hearing. He suggests that adjudicators may apply inconsistent criteria when judging performances (see also, Fiske, 1983; Wapnick, Flowers, Alegant & Jasinskas, 1993).

To help minimise the influence of this problem, researchers have employed predetermined evaluative criteria as an attempt to improve reliability (e.g., Fiske, 1978; Hunter & Russ, 1996; Winter, 1993). Harold Abeles (1973) asserts that "rating scales improve evaluation because adjudicators must use a common set of evaluative dimensions rather than develop their own subjective criticisms" (p. 246). To develop a clarinet performance adjudication scale, Abeles first asked 17 instrumental music teachers to write a short essay describing a music performance. Abeles then conducted a content

analysis of the essays which he used to identify 54 distinct statements about clarinet performance. These statements were combined with 40 statements obtained from other studies, and the 94 statements were converted to Likert scales for use in adjudication. Next, ratings for all items were obtained for 100 performances, and a factor analysis was conducted on the resultant data. A six-factor solution was identified. These factors were labelled interpretation, intonation, rhythmic continuity, tempo, articulation, and tone. Five Likert scales were selected to represent each of the six factors, yielding a total of 30 Likert scales. Each scale had a high factor loading on the factor it was selected to define, and a low correlation with the other five factors. The set of 30 Likert scales made up the Clarinet Performance Rating Scale (CPRS). Interjudge reliability estimates based on the CPRS were extremely high, although it is difficult to assess whether these reliability values were higher than they would have been if judges had employed another scale or used their own personal criteria.

As another example, Bergee (1988, 1989) developed a standardised assessment scale for brass performances, called the Brass Performance Rating Scale (BPRS). The BPRS consists of 27 rating scales, under four major headings. These headings were derived from a series of factor analyses, and were labelled: (1) interpretation/musical effect; (2) tone quality/intonation; (3) technique; and (4) rhythm/tempo. Again, estimates of interjudge reliability for this standardised scale were very high, but these estimates were not compared to estimates based on alternative assessment scales.

In addition to its potential for increasing reliability, assessment criteria can be used to justify and explain assessments to performers, parents, peers, or administrators. An innovative attempt to devise criteria-specific rating scales that allows judges to provide specific feedback according to the strengths and weaknesses of a performance has been promoted by Saunders (1993). His criteria-specific rating scales aim to provide clear indication of what is present in a student's performance and are therefore useful in situations in which specific information is gathered to provide feedback for the purpose of improving performance. Adjudicators using his woodwind/brass solo scale are asked to tick boxes according to aspects of solo evaluation (i.e., tone, intonation, technique/articulation, melodic accuracy, rhythmic accuracy, tempo, interpretation), scales (technique, note accuracy, musicianship), and sightreading (tone, note accuracy, rhythmic accuracy, technique/articulation, interpretation). For example, more than one tick can be placed in the boxes for the solo evaluation component dealing with technique/articulation, depending on whether the aspect has been demonstrated by the performer:

- appropriate and accurate tonguing
- appropriate slurs as marked
- appropriate accents as marked
- appropriate ornamentation as marked
- appropriate length of notes as marked (i.e., legato, staccato)

(Check all that apply. Each is worth 2 points)

More recently, Saunders & Holahan (1997) have attempted to validate the criteria-specific rating scales using 36 adult judges who auditioned 926 students for an All-State Band. Results demonstrate the reliability of the criteria-specific rating scales with different judges collectively demonstrating consistency in their performance evaluation results. The advantage of using these scales is that the judges were able to "provide specific information about (a) the areas and levels of performance accomplishment, and (b) the areas and levels of performance accomplishment not yet achieved". (p. 270) In addition, five instrumental performance dimensions (tone, technique/articulation, rhythmic accuracy, interpretation, sightreading-interpretation) were found to predict the students' total score.

In contrast to the above studies which advocate the use of specific, predetermined evaluative criteria, a variety of alternative views have become more pronounced in the literature. For instance, Mills (1991) expressed concern about the methods used to assess performances in the final two years of high school by some examination boards in Britain. According to one method, assessors are asked to assign marks of up to three for each of five *skill* categories (accuracy of notes, accuracy of rhythm, phrasing, control of medium, and technique adequate to the piece) and five *interpretation* categories (effective dynamics, appropriate tempo, suitable sense of styles, sense of involvement in the music, and sense of performance) (p. 174). This view rests on the assumption that a musical performance can be broken down into a number of meaningful components. However, according to Mills this premise does not make sense (p. 174), because musicians find it difficult to work this way. She provides the following comments to illustrate her point:

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

SKILLS + INTERPRETATION = PERFORMANCE

I recall performances which have overwhelmed me, despite there being a handful of wrong notes. I remember others in which the notes have been accurate, and the interpretation has been legitimate, and yet the overall effect has been sterile. A performance is much more than a sum of skills and interpretation. (p. 175)

Swanwick (1996) extends this view by stressing the importance of acknowledging the complexity of musical experience. In his opinion:

Such a rich activity cannot be reduced to a single dimension, say that of 'technique'. On the other hand, it does not make sense to identify several different dimensions and assess them giving a separate mark for each - say for technique, expressiveness and stylistic awareness - adding them up to get a single figure. When we conflate several observations we lose a lot of important information on the way. For instance, in competitive ice skating one performer might be given six out of ten for technique and nine for artistry, while another contender gets nine for technique and six for artistry. The sum of each set of marks happens to be the same - 15 - but the actual performances will be quite different. The fudge of adding a category called 'overall' only makes things worse. (p. 6)

For Swanwick, assessment of performances in formal music settings should involve criterion statements which are clear, qualitatively different from each other, brief enough to be memorable but substantial enough to be meaningful, able to be hierarchically ordered, and useful in a range of settings. Importantly, they should reflect the essential nature of the activity (p. 8). Swanwick (1993) proposes an eight level hierarchy of performance abilities, which he believes is helpful in adjudicating all types of musical performance (see Figure 2).

The evaluation process may be defined as the implicit and explicit decisions that lead to an assessment. This process depends on a number of factors, including the training of evaluators, whether assessments are done individually or by panel, the scope of the evaluation, constraints placed on the evaluators, the physical environment, and evaluator expectations.

As noted earlier, the benefits of training programs for judges have not been clearly established. It is often assumed that training in performance assessment can improve reliability, and decrease biases relating to race, gender, or personal beliefs. Performance examiners typically are trained by examining bodies. An important goal of

training is to establish consistent criteria among adjudicators. However, different examining bodies may emphasise different criteria.

Level 1: The rendering is erratic and inconsistent. Forward movement is unsteady and variations of tone colour or loudness appear to have neither structural nor expressive significance.

Level 2: Control is shown by steady speeds and consistency in repeating patterns. Managing the instrument is the main priority and there is no evidence of expressive shaping or structural organisation.

Level 3: Expressiveness is evident in the choice of speed and loudness levels but the general impression is of an impulsive and unplanned performance lacking structural organisation.

Level 4: The performance is tidy and conventionally expressive. Melodic and rhythmic patterns are repeated with matching articulation and the interpretation is fairly predictable.

Level 5: A secure and expressive performance contains some imaginative touches. Dynamics and phrasing are deliberately contrasted or varied to generate structural interest.

Level 6: There is a developed sense of style and an expressive manner drawn from identifiable musical traditions. Technical, expressive and structural control are consistently demonstrated.

Level 7: The performance demonstrates confident technical mastery and is stylistic and compelling. There is refinement of expressive and structural detail and a sense of personal commitment.

Level 8: Technical mastery totally serves musical communication. Form and expression are fused into a coherent and personal musical statement. New musical insights are imaginatively and systematically explored. (Swanwick, 1993, p. 8)

Figure 2: Swanwick's (1993) Criterion Statements for Music Performance

Winter (1993) reported that training programs facilitate the assessment process by highlighting evaluative criteria that are determined prior to the performance. Based on a review of approaches of various examining bodies, Winter designed an assessment tool called the Music Performance Assessment instrument (MPA). The MPA involves assigning a rating to a number of descriptive statements, which are listed under the following headings: technical, pitch, time, interpretation, and overall. Following the assignment of ratings, adjudicators rank performances in order of preference.

Using the MPA, Winter asked qualified musicians and educators to evaluate three piano performances. Adjudicators were classified into each of four categories: untrained and inexperienced; trained and inexperienced; untrained and experienced; and trained and experienced. Trained adjudicators were those who had attended a short preparation course. According to Winter, training in assessment resulted in a greater understanding of the criteria involved, and greater confidence in making a judgement.

The assessment process is sometimes shaped by physical constraints. For example, the distance between the evaluator and the performance, the use of a musical score, and the physical environment, may all influence the set of decisions that lead to an assessment. The assessment process can also be guided by expectations about the performer, either from knowledge of that performer, or from visual impressions of the

performer (e.g., stage presence). In this regard, visual impressions of performers may have a negative or positive effect on an assessment, especially when the performer appears stiff, cold, relaxed, or emotionally involved. As Schumann observed, "If Liszt played behind a screen, a great deal of the poetry would be lost" (cited in Davidson, 1993). We are only just beginning to understand how these types of visual cues provide perceptually relevant information about the performer and the performance. Davidson (1993) found that visual (kinematic) performance cues can signal or clarify the expressive intentions of the performer. Such cues include hand, arm and head movements. An implication of this finding is that it might be appropriate in some circumstances to encourage adjudicators to consider both visual and auditory information when assessing performances.

The final assessment may be reported as a rating, grade, rank, or report. Depending on the form of the final assessment, further analyses may be appropriate. Written reports may be summarised or discussed, while ratings, grades, and ranks may be analysed to ensure that there is adequate consensus among judges.

Analysis of assessment data has the broader purpose of providing feedback to the examining body regarding the procedures, evaluative criteria, evaluative instruments, and adjudicators employed. Such analyses have revealed a number of important results on assessment. Wapnick et al. (1993) reported that reliability is not significantly related to: 1) whether judges can play the instrument involved; 2) the performance ability of judges; 3) the duration of the performance being assessed; 4) whether the judge is listening to live or taped performances; 5) whether a judge has access to the musical score; or 6) whether a judge uses a rating scale. Bergee (1993), however, found that self-evaluations correlate quite poorly with faculty and peer evaluations, a finding that raises questions about the reliability of self-evaluations.

The results of a performance assessment may be reported in isolation or they may be collated with other information. Such results may be used for practical purposes, such as deciding about seating placements or grade level, or they may be used for diagnostic purposes. Providing feedback to performers that is consistent, appropriate and designed to improve future performance may exert a positive influence on a musician's subsequent performances.

Conclusions

A key intention of this review of literature has been to propose a model that could be used to structure further research in the area. As is evident from our review, existing literature on music performance assessment is extremely limited. Our model therefore serves the purpose of alerting researchers to some of the main gaps in research that need to be filled.

A number of important conclusions emerge from our review. First, a more holistic approach is needed in the research literature; one that attempts to not only clarify the measurement instruments and musical contexts in which various types of performance assessments take place, but other related factors, such as the personal, social and cultural biases that influence a judge's assessment. More research is also needed to explain the range of performer characteristics that impact on an assessment. Preliminary evidence suggests that the cliché 'mind over body' does have some validity, especially in terms of the cognitive mediational processes that impact on a musician's ability to perform in an efficient and effective manner. Added to this is the extent to which the performer is made aware of the evaluative criteria that will be used to assess the performance. This single dimension may be an important component of improving performance, and of alleviating some of the conflicting perceptions that can occur between evaluators and musicians, and also between teachers and their students.

In the end, perhaps the single most important variable is the judge, for as Fiske aptly reminds us: "An evaluation of a performer does not mean anything until we know

how reliable the judge was who evaluated that performance". (Fiske, 1994, p. 76) As researchers we need to obtain further evidence on the variety of elements that directly influence the decision making process of music assessors, and how these are mediated by social, cultural, personal and professional factors. To what extent is a judge able to provide reliable assessments under differing contexts and for differing purposes, such as a music examination, audition and competition? In what situations and in what ways can training help alleviate some of the problems of inter or intra-judge reliability? These questions, plus a host of others, remain unanswered, despite over 30 years of research in the area.

References

- Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, 21(3), 246-255.
- Abeles, H.F., & Porter, S.Y. (1978). The sex-stereotyping of musical instruments. *Journal of Research in Music Education*, 26, 65-75.
- Austin, J. R. & Vispoel, W.P. (1992). Motivation after failure in school music performance classes: The facilitative effects of strategy attributions. *Bulletin of the Council for Research in Music Education*, 111, 1-23.
- Bergee, M.J. (1988). Use of an objectively constructed rating scale for the evaluation of brass juries: A criterion-related study. *Missouri Journal of Research in Music Education*, 5(5), 6-25.
- Bergee, M.J. (1989). An objectively constructed rating scale for euphonium and tuba music performance. *Dialogue in Instrumental Music Education*, 13, 65-86.
- Bergee, M.J. (1992). A comparison of faculty, peer, and self-evaluations of applied brass jury performances. *Journal of Research in Music Education*, 41(1), 19-27.
- Berry, W. (1989). *Musical Structure and Performance*. Yale University Press: New Haven.
- Bradley, I.L. (1972). Effect on student musical preference of a listening program in contemporary art music. *Journal of Research in Music Education*, 20, 344-353.
- Brodsky, W. (1996). Music performance anxiety reconceptualized: A critique of current research practices and findings. *Medical Problems of Performing Arts*, 11(3), 88-98.
- Bruce, R. & Kemp, A. (1993). Sex-stereotyping of children's preferences for musical instruments. *British Journal of Music Education*, 10, 213-217.
- Cantor, J. R. & Zillman, D. (1973). The effect of affective state and emotional arousal on music appreciation. *Journal of General Psychology*, 89, 97-108.
- Catell, R.B. & Anderson, J.C. (1953). The measurement of personality and behavior disorders by the IPAT Music Preference Test. *Journal of Applied Psychology*, 37, 446-454.
- Cooksey, J.M. (1977). A facet-factorial approach to rating high school choral music performance. *Journal of Research in Music Education*, 25, 100-114.
- Davidson, J. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, 21, 103-113.
- Duerksen (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education*, 20, 268-272.
- Elliott, C. A. (1995). Race and gender as factors in judgements of musical performance. *Bulletin of the Council for Research in Music Education*, 127, 50-56.
- Fiske, H. E. (1977). The relationship of selected factors in trumpet performance adjudication. *Journal of Music Education Research*, 25(4), 256-263.
- Fiske, H. E. (1978). The effect of a training procedure in music performance evaluation on judge reliability. Ontario Educational Research Council Report.
- Fiske, H. E. (1983). Judging musical performances: Method or madness? *Update*, 7-10.
- Fiske, H. E. (1994). Evaluation of vocal performances: Experimental research evidence. (74-103). In G. Welch & T. Murao (Eds.), *Onchi and Singing Development*, London: David Fulton Publishers.

- Flores, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking? *The Statistician*, 45(1), 97-104.
- Goolsby, T.W. (1995). Portfolio assessment for better evaluation. *Music Educators Journal*, 82, 39-44.
- Hunter & Russ (1996). Peer assessment in performance studies. *British Journal of Music Education*, 13, 67-78.
- Johnson, O. & Knapp, R.H. (1963). Sex differences in aesthetic preferences. *Journal of Social Psychology*, 61, 279-301.
- Kemp, A. (1982). The personality structure of the musician. III. The significance of sex differences. *Psychology of Music*, 10(1), 48-58.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87(1), 77-107.
- Lehman, P. R. (1968). *Tests and Measurements in Music*. Prentice Hall, Inc. Englewood Cliffs, New Jersey.
- Lidral, K. A. (1997). The solo contest for jazz improvisors. *Jazz Educators Journal*, 29(6), 40-43.
- MacKenzie, C. G. (1991). Starting to learn to play a musical instrument: A study of boys' and girls' motivational criteria. *British Journal of Music Education*, 8, 15-20.
- McPherson, G. E. (1989). Cognitive mediational processes and positive motivation: Implications of educational research for music teaching and learning. *Australian Journal of Music Education*, 1, 3-19.
- McPherson, G. E. (1993). Factors and abilities influencing the development of visual, aural and creative performance skills in music and their educational implications. Doctor of Philosophy, University of Sydney, Australia. Dissertation Abstracts International, 54/04-A, 1277. (University Microfilms No. 9317278).
- McPherson, G.E. (1994). Factors and abilities influencing sightreading skill in music. *Journal of Research in Music Education*, 42, 217-231.
- McPherson, G. E. (1995a). Redefining the teaching of musical performance. *The Quarterly Journal of Music Teaching and Learning*, VI(2), 56-64.
- McPherson, G.E. (1995b). The assessment of musical performance: Development and validation of five new measures. *Psychology of Music*, 23, 142-161.
- McPherson, G. E. (1996). Factors influencing the assessment of musical performance. Unpublished paper presented at the 22nd World Conference of the International Society for Music Education, Amsterdam, Holland, July, 1996.
- Mills, J. (1991). Assessing musical performance musically. *Educational Studies*, 17(2), 173-181.
- Morgan, J. & Burrows, B. (1981). Sharpen your edge on choral competition. *Music Educators Journal*, 67(8), 44-47.
- Namba, S, Kuwano, S., Hatoh, T., & Kato, M. (1991). Assessment of music performance by using the method of continuous judgment by selected description. *Music Perception*, 8(3), 251-276.
- Radocy (1976). Effects of authority figure biases on changing judgments of musical events. *Journal of Research in Music Education*, 24, 119-128.
- Reed, R. (1990). Here comes the judge: Solo string adjudication. *American String Teacher*, summer, 46-49.
- Ross, M., Radnor, H., Mitchell, S., & Birtton, C. (1993). *Assessing achievement in the arts*. Buckingham: Open University Press.
- Saunders, C. T. (1993). The assessment of music performance: Techniques for classroom and rehearsal. *Newsletter of the Special Research Interest Group in Measurement and Evaluation (MENC)*, R. Colwell & R. Ambrose (Eds.), 15, Spring, 7-11.
- 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.
- Sheldon, D. A. (1994). The effects of competitive versus noncompetitive performance goals on music student's ratings of band performances. *Bulletin of the Council for Research in Music Education*, 121, 29-41.
- Swanwick, K. (1996). Teaching and assessing. *Newsletter of the Special Research Interest Group in Measurement and Evaluation (MENC)*, R. Colwell & J. Roberts (Eds.), Winter, 18, 6-9.

- Thompson, W.F., Diamond, C.T.P., & Balkwill, L. (In Press). The adjudication of six performances of a Chopin Étude: A study of expert knowledge. *Psychology of Music*.
- Wapnick, J., Flowers, P., Alegant, M., & Jasinskas, L. (1993). Consistency in piano performance evaluations. *Journal of Research in Music Education*, 41(4), 282-292.
- Watkins, J. G., & Farnum, S. E. (1954). The Watkins-Farnum performance scale. A standardised achievement test for all band instruments. Winona, Minnesota: Hal Leonard Publishing.
- Welch, G.F. (1994). The assessment of singing. *Psychology of Music*, 22, 3-19.
- Whybrew, W. E. (1971). *Measurement and Evaluation in Music*. Wm.C. Brown Company Publishers, Dubuque, Iowa.
- Winter (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, 34-39.
- Zervoudakes, J., & Tanur, J. (1994). Gender and musical instruments: Winds of change? *Journal of Research in Music Education*, 42(1), 58-67.
- 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.

About the Authors

Dr. Gary McPherson is course coordinator for music education at the University of New South Wales in Sydney, Australia. He is currently Treasurer for the International Society for Music Education, an Associate Editor for *Psychology of Music* and on the Editorial Advisory Board for the *Council for Research in Music Education*. He is also co-editor for *Research Studies in Music Education*.

Dr. William Forde Thompson is Associate Professor in the Department of Psychology, Atkinson College, York University, Toronto, Canada. He is cross-appointed to the Department of Music, York University, and publishes in the area of music cognition. He is on the Editorial Advisory Board of *Psychomusicology*.