

Towards enhancing student learning and examiner reliability with criterion-referenced assessment in the creative arts: the case of music

Adrian Thomas and Brad Millard

Queensland University of Technology, Brisbane, Queensland

For the past 30 years or so, various attempts have been made to develop criteria from which to judge the quality of solo performances in music. The laudable rationale behind almost all these attempts has been to mitigate against excessive examiner subjectivity, and to do so by defining the criteria against which performances are to be judged. However, until relatively recently there has been relatively little consideration given to specifying appropriate standards associated with each criterion and consequently negligible thought given to enhancing student learning through interaction with the criteria and standards.

This paper reviews and analyses recent approaches to criterion-referenced assessment in the creative arts. It uses the undergraduate music course at Queensland University of Technology as a case study in the development of criteria and standards, tracing the evolution of practice there over a ten-year period. Gaps in the effectiveness for the triangulation of student learning, academic grading and assessment criteria are identified and analysed. From these, a set of governing principles for developing and setting of standards in creative arts units are proposed, and approaches to enhancing student learning through interacting with these standards are outlined.

Introduction

A major preoccupation in the Australian university sector in recent years has been the attempt to write highly detailed descriptive criteria and standards for assessment tasks. This has proved especially problematic for the creative arts sector, and music in particular. A music performance, or indeed any visual and performing arts creative activity, is multi-faceted with many interlocking variables, inhibiting the writing of an explicit and meaningful range of objective standards against which to measure student performance. Indeed, there are suggestions that '[their] alleged explicit clarity, precision and objectivity is largely spurious' (Hussey and Smith, 2002, p. 232), and that 'a single-minded focus on explicit articulation falls short of providing students and staff with common and meaningful knowledge of standards and criteria' (O'Donovan, Price and Rust, 2004, p. 333).

Assessing solo music performance according to pre-determined criteria is a well-established practice. From the early 1970s (Abeles, 1973; Fiske, 1977) and consistently since then, various attempts have been made to develop criteria against which to judge the quality of music performances. The laudable rationale behind almost all these attempts has been to mitigate against excessive examiner subjectivity, and to do so by defining the criteria against which performances are to be judged. However, until relatively recently there has been relatively little consideration given to describing the standards associated with each criterion and consequently negligible thought given to enhancing student learning through interaction with the criteria and standards. Indeed, Johnson (1997) contends that "[while] it is desirable that students and their teachers have sight of the criteria, they are primarily for the use of the examiners". (p. 273) However, if that is the case, then it is probably not worth the hours put into the development of criteria and standards, because in practice examiners tend to refer obliquely to the text-based criteria and rely principally on their own opinion of standard (Stanley, Brooker, & Gilbert, 2002).

Even when attempts were made to construct rubrics for music performance, the approaches used to describe levels of performance and/or to derive an overall result from the levels achieved by the student were open to question.

In particular, the standard practice in criterion-referenced assessment of matching the student performance in each criterion with the relevant published standard, then arriving at an overall result by adding together the various standards, was often incongruent with the intuitive global judgement made by the assessor. In particular, the practice of marking performance quality against criteria such as rhythm, pitch, dynamics, phrasing, tempi and the like, then adding to achieve a result, seemed often at odds with the actual experience of the performance itself. And, as Stanley et al (2002) found, examiners far preferred to address a relatively small set of criteria comprising global statements of achievement rather than focus on a wide range of variables.

The more recent practice of providing detailed text descriptors of standards equating to “outstanding”, “very good”, “good” and “pass”, for each criterion (Dunn, Parry, & Morgan, 2002), while apparently a step forward in creating transparent standards for students and clear guidelines for assessors, may actually have been a retrograde step. Assessors still marked the student achievement measured against each criterion then “added” to arrive at a result. However, assessors often found considerable difficulty in matching what they had heard to the fixed-text descriptors on offer in the ‘Standards Schema’ (Stanley et al., 2002). As a consequence, the standard they circled in the schema was only an approximate description of the actual quality of the performance. Table 1 below shows a possible Standards Schema for a student performance on piano:

Criteria	Outstanding	Very Good	Good	Pass	Fail
TECHNIQUE Desired attributes: A performance that demonstrates control of and accurate response to: Rhythm Pitch Dynamics Articulation Tone quality	You displayed a consistent and compelling command of all the technical attributes needed to realise the music.	You displayed a consistent command of all of the technical attributes needed to realise the music.	You displayed a command of most of the technical attributes needed to realise the music. Attributes requiring attention to enable you to move to a higher standard of performance are marked in the Technique column and detailed in the attached report.	You displayed a command of some of the technical attributes needed to realise the music. Attributes requiring attention to enable you to move to a higher standard of performance are marked in the Technique column and detailed in the attached report.	You displayed an insufficient command of the technical attributes needed to realise the music. A detailed report is attached.
ARTISTRY Desired attributes: A performance that is expressive, stylish, and well characterised	The music you performed was exceptional in its artistry: it was expressive, appropriate to its style, well-characterised and compelling	The music you performed was consistently expressive, appropriate to its style and well-characterised	The music you performed was mostly expressive, appropriate to its style and well-characterised	The music you performed was sometimes expressive, appropriate to its style and quite well characterised	The music you performed was insufficiently expressive, stylish and characterised

Table 1. Possible Standards Schema for piano performance

Although such a schema is concise and provides a summative précis of a performance, there are inherent problems. For example, In an actual performance situation, an examiner may view the technique displayed as “Outstanding”, but lacking a little in intensity/ imagination/ involvement, thus rating artistry as “very good”. Another performer presenting the same music might be rated “Very Good” for technique but ‘Outstanding” for artistry. The performances have the same rating outcome, but are of quite different quality. Now it may be appropriate to use an additive method to calculate tennis player rankings—based on the number of tournaments they played, their results in these tournaments and the rating of the tournaments themselves , but it is problematic to assess something as complex as a music performance in this manner (Swanwick, 1996).

To an extent, this problem in music and across the creative arts can be modified by allocating a different weighting

to Technique and Artistry as the student progresses through the course. That is, strong technical skills should be an assumed part of the student's armoury as the course nears its end, so greater weight can be given to artistry. And if achieving an appropriate balance between technique and artistry is the most difficult task facing the expert examiner (Johnson, 1997), depending as it does on judgement and experience (p. 279), communicating this to students effectively is doubly difficult. Indeed, it is highly doubtful if text-based criteria alone can hope to achieve in students an understanding of the complexities of a performance. As well, the usual requirement of universities that examiners give a numerical or letter grade to the performance tends to exacerbate their relative disregard for text descriptors of standard. If progress is to be made, the experience at QUT and elsewhere (Blom & Poole, 2004; McPherson & Thompson, 1998) would suggest that far more needs to be done in interrogating the criteria and standards, and that students and staff need to be equally involved in this interrogation.

Assuming that the examiners return the schema with the standard achieved clearly indicated, the student is then aware of the grade achieved and the components that contributed towards it. However, in tertiary music performance examinations, the difficulties of reporting via a standards schema such as that above are exacerbated when the norm is for students to present a number of pieces of different style and character. A solution tried and abandoned at QUT some years ago was to rate each piece separately, then award an overall grade based on the mean. This caused a raft of problems, ranging from disputes about the relative difficulty of pieces to student expectations that short pieces would be valued less than longer ones. Clearly, as a recent study at the Sydney Conservatorium of Music found, "procedures...need to be established to facilitate examiners' use of criteria where the standard of a student's performance varies widely during a recital" (Stanley et al., 2002).

A similar situation applies to student work in the visual arts and in music composing, where students present a folio of disparate works completed during the semester.

Criterion-referenced assessment of music performance at QUT

At Queensland University of Technology, all end of semester solo performance music examinations have been criterion-referenced since the mid-1990s. Influenced by research that questioned the previously mentioned additive approach to judging music performance, (Mills, 1991; Swanwick, 1996), which takes the view that a performance is more than the sum of its parts, QUT adopted a global model for assessment, drawing closely on Swanwick's eight-level criterion statements for music performance (1996). These were reduced by two to match the grading levels required by the university. Table 2 shows the resultant schema, with "A" = outstanding, "D" = pass level and "F" = fail.

CRITERIA FOR THE ASSESSMENT OF SOLO PERFORMANCE MUSIC EXAMINATIONS AT QUT FACULTY OF CREATIVE INDUSTRIES 1996-2000

Overall, an air of confident mastery of technical, musical and interpretive elements is conveyed. The performance is stylistic and compelling, with refinement of expressive and structural detail and a sense of personal commitment.	A
There is a developed sense of style and an expressive manner drawn from identifiable musical traditions. Technical, expressive and structural control are consistently demonstrated.	B
A secure and expressive performance that contains some imaginative touches and which indicates deliberate attention to detail. Dynamics and phrasing are deliberately contrasted to or varied to generate structural interest.	C
The performance is generally secure technically and there are instances of effective expression. Melodic and rhythmic patterns are repeated with matching articulation, but the interpretation is fairly predictable.	D
Control is sometimes inconsistent and there is not much evidence of expressive shaping or structural organization.	E
The performance is erratic and inconsistent. Forward movement is unsteady and variations of tone colour or loudness appear to have neither structural nor expressive significance.	F

Table 2. QUT first adaptation of Swanwick's Criterion Statements for Music Performance

These criteria were used by staff and students unchanged until 2000. Feedback to students after their recitals included a marked copy of the criteria with their result clearly shown, either by circling the relevant statement and/or the grade. As well, a detailed report on each piece presented was appended.

Discussions with examiners and students over the years revealed a range of issues. A major difficulty for examiners in particular was the limiting aspect of the criteria. In practice, the only criteria statements they found to reflect an actual performance were the A and F standards, and the latter to a lesser degree. Of concern were the qualifying statements in the other criteria. For example: "...dynamics and phrasing are deliberately contrasted to or varied to generate structural interest" was either found to be not applicable to the performance, or an incomplete reflection of other aspects of the performance that led to the grade. Further, the use of the singular "performance" was problematic, given that one piece might be performed considerably better than another. For students, the standards schema did not serve its intended specific purpose of guiding them towards excellence in performance; rather, it was treated as an incidental. Accordingly, as shown in Table 3 below, the schema was revised in 2000 and used until 2003.

**CRITERIA FOR THE ASSESSMENT OF SOLO PERFORMANCE MUSIC EXAMINATIONS AT QUT
 FACULTY OF CREATIVE INDUSTRIES 2000-2003**

Overall, an air of confident mastery of technical, musical and interpretive elements is conveyed throughout the recital. An outstanding level of performance.	A+	A	A-
There is a developed sense of style and an expressive manner drawn from identifiable musical traditions. Technical, expressive and structural control are consistently demonstrated throughout the recital. A very high level of performance. A very high level of performance.	B+	B	B-
Secure and expressive performance(s) that contain some imaginative touches and that indicate deliberate attention to detail. A high level of performance.	C+	C	C-
Performance(s) are generally secure technically and there are instances of effective expression . Instrument control is occasionally inconsistent. A satisfactory level of performance.	D+	D	D-
Control is often inconsistent and there is not much evidence of expressive shaping or structural organisation in the performance(s). Just below a satisfactory level of performance.	E+	E	E-
Performance(s) are erratic and inconsistent. An unsatisfactory level of performance.		F	

Table 3. QUT second adaptation of Swanwick's Criterion Statements for Music Performance

Examiners found this schema, with many of the qualifying statements from the first iteration removed, more suitable for assessment purposes. They regarded the reductive standards as more clearly expressed and, importantly, comparable to statements educated musicians might make about a professional performance. The ability to indicate whether the performance was at the lower, mid, or high range of the relevant standard gave them a greater degree of flexibility to exercise subjective judgement, which they were able to justify in the detailed written comments on each piece. Further, the use of the plural "performances" enabled them to take a more holistic view of the recital. However, despite these improvements from the examiners' perspective, the change in criterion statements had no marked improvement in student performance, or a reduction of unhappiness with results.

To provide a greater degree of information to students prior to and after assessment, the criteria and standards were revised once more, aimed at striking a balance between specificity and global description. In this iteration, current to the present day, the variants on the Swanwick criteria (1996) have been abandoned, for three principal reasons. First, the various statements shown in figures 2 and 3 are not criteria; rather, they are statements of standard, with criterion implications. Second, the standards do not compare like with like, which is in opposition to recent research (Tierney & Simon, 2004) which indicates that attributes should be addressed consistently from one level to the next across the scale. Third, students indicated that the statements were too summative in nature; they provided little formative guidance to them as they proceeded with the preparation of their program during the semester. Accordingly, the schema in figure 4 below, based to a degree on criteria promulgated by Stanley et al, (2002), was constructed and is in current use:

CRITERIA	STANDARD A	STANDARD B	STANDARD C	STANDARD D	STANDARD E	STANDARD F
1. Technical achievement Weighting 30%	The performance was consistently flawless in every respect.	The performance was nearly always flawless, but occasionally minor details were not completely under control	The performance was nearly always under control, but minor details were not completely under control in places, which detracted somewhat from the overall impression	The performance was under control overall, but there were several instances of errors or slips and/or loss of control which adversely affected the general impression at times	The performance was under control in a broad sense only, but numerous errors or slips were evident and/or loss of control which adversely affected the general impression	The performance was significantly affected by inconsistent control of technical elements. Evidence of this may be in the form of a large number of slips, errors or significant loss of control
2. Musical effectiveness Weighting 30%	Consistently demonstrates comprehensive understanding and skill in these attributes, resulting in musically compelling performances	Nearly always demonstrates substantial understanding and skill in these attributes, resulting in musically convincing performances	Usually demonstrates understanding and skill in these attributes, resulting in performances that are musically satisfying in most parts	Sometimes demonstrates understanding and skill in these attributes, resulting in performances that are musically satisfying in some parts	Demonstrates understanding and skill in these attributes in a broad sense only, resulting in performances that are musically satisfying on occasion	Only occasionally or never demonstrates understanding and skill in these attributes, resulting in musically unsatisfying performances overall
3. Creativity Weighting 20%	Consistently demonstrates compelling outcomes in this criterion	Nearly always demonstrates convincing outcomes in this criterion	Often demonstrates convincing outcomes in this criterion	Generally demonstrates satisfying outcomes in this criterion	Only occasionally demonstrates satisfying outcomes in this criterion	Rarely or never demonstrates satisfying outcomes in this criterion
4. Presentation and Documentation Weighting 20%	Consistently demonstrates comprehensive understanding and skill in these attributes, resulting in a compelling presentation	Nearly always demonstrates substantial understanding and skill in these attributes, resulting in a presentation with considerable impact	Usually demonstrates understanding and skill in these attributes, resulting in a presentation with a positive impact in most parts	Sometimes demonstrates understanding and skill in these attributes, resulting in a presentation with a positive impact in some parts	Seldom demonstrates understanding and skill in these attributes, with positive impact on occasion	Rarely or never demonstrates understanding and skill in these attributes, with little positive impact

1. Technical achievement - accuracy in control of musical elements as appropriate within your chosen domain, including tempo, rhythm, beat/pulse, pitch, intonation, balance, articulation, dynamics, tone colour, space, vibrato, other techniques as required in the music
2. Musical effectiveness - Includes as appropriate: expressive, stylistic and structural sophistication, sound production, projection and nuance, communication with other performers
3. Creativity – unique, interestingly different musical creativity and artistic individuality within your chosen domain
4. Presentation and Documentation – audience communication/engagement. Clarity, appropriateness and effectiveness of spoken communication, written communication (supplied “write-up” and copy of sheet music/score) and stage presence. Write-up should be 750 - 1000 words.

Table 4. QUT current Criterion Statements for Music Performance

“But I still don’t understand why I received that grade...”

QUT students are able to access the criteria and standards documents from the start of each semester, and their end-of-semester examination attempts to replicate a public recital situation. This is particularly so for final year students, who are required to venture beyond the safety of the music department milieu. They hire performance venues around the city, dress for the occasion, run a box office to extract money from obliging parents and friends, and generally perform at their very best. Two examiners are seated discreetly and separately among the audience and provide extensive written comments on each piece presented. The audience response is unfailingly enthusiastic, and the experience is nearly always a positive one for the students—until the release of results. Then, some students are unable to reconcile the acclamation they receive at the public performance with the detailed written report that accompanies the grade awarded by the examiners. Attempts at explanation and conciliation are usually brushed aside. The student feels let down. Clearly there is a lack of congruency between the meaning of the published criteria and standards, the student understanding of the quality of his or her performance, and the standards of the examiners. With respect to the latter, it is futile to expect examiners to be totally objective in their assessment of performance. Laming (1990) observed that

Assessment of ability in musical contexts is normally subjective Judges, who are often music professionals, listen to prepared performances and then rate these according to implicit or explicit criteria. It is well established that such assessments, even if offered by experienced and trained judges, have limited reliability. (p. 241)

Summary: writing and using graded standards in units with high creative/ re-creative/ artistic components

To enhance student learning, achievement and examiner reliability, a three-pronged approach is suggested. First, the criteria and standards should be written in clear, unambiguous language. Each standard should be expressed concisely, but free of the myriad variables and wordy descriptors that make accurate and/or meaningful reporting difficult. This is not easy to do, and it is a truism that “making expectations clear and explicit is problematic in ... interpretive and applied fields where grading criteria cannot be too precise or they will constrain student performance” (Parry, Hayden and Speedy, 2000, p. 4). The current criteria sheet in use at QUT (Table 4, above) is an attempt to provide clarity of standards and expectations to students and, at the same time, to allow for the exercise of professional judgement by staff.

Second, and most importantly, students must constantly interact with the criteria and standards throughout the semester, so that gradually a clear notion of ‘quality’ and ‘standard’ is embedded in consciousness. To assist in this process, peer assessment of student work in a non-threatening environment has been identified as contributing to an appreciation of the assessment process itself and to assist in understanding of the strengths and weaknesses of their own performance (Blom & Poole, 2004; D. Hunter, 1999; D. Hunter, and Russ, M, 1996). By itself, however, the peer assessment process may lead to ultimate disappointment when the student finds that examining staff do not share the assessment standards. To help address this difficulty, QUT plans for students to view videotaped performances approved for training purposes by past students at the same year level. Referring continually to the published criteria and standards under staff guidance, students will discuss the qualities of the relevant performance and interact with examiners. In this way, they should begin to understand the meaning of the criteria and standards and be able to assess their own performance standard. The degree to which students are made aware of criteria and standards has been identified as “ 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 students (McPherson & Thompson, 1998).

Last, examiners should undergo training sessions with the criteria and standards prior to the examination period. Where possible, they should also engage in dialogue with students in workshop sessions. For example, making frequent reference to the criteria and standards to be used in examination, they could lead sessions in non-threatening situations—such as the video seminars proposed above.

References

- Abeles, H. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, 21(3), 246-255.
- 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(01 - Mar 2004), 111-125.
- Dunn, L., Parry, S., & Morgan, C. (2002, 28-30 August, 2002). *Seeking quality in criterion referenced assessment*. Paper presented at the Learning Communities and Assessment Cultures, Northumbria.
- Fiske, H. (1977). The relationship of selected factors in judgements of musical performance. *Journal of Music Education Research*, 25(4), 256-263.
- Hunter, D. (1999). Developing peer-learning programs in music. *British Journal of Music Education*, 16(1), 51-63.
- Hunter, D., and Russ, M. (1996). Peer assessment in performance studies. *British Journal of Music Education*, 13, 67-78.
- Johnson, P. (1997). Performance as Experience: the problem of assessment criteria. *British Journal of Music Education*, 14(3), 271-282.
- McPherson, G., & Thompson, W. (1998). Assessing Music Performance: Issues and Influences. *Research Studies in Music Education*(10), 12-24.
- Mills, J. (1991). Assessing Musical Performances Musically. *Educational Studies*, 17(2), 173-181.
- Stanley, M., Brooker, R., & Gilbert, R. (2002). examiner Perceptions of Using Criteria in Music Performance Assessment. *Research Studies in Music Education*(18), 43-52.
- Swanwick, K. (1996). Teaching and Assessing. *Newsletter of the Special Research Interest Group in Measurement and Evaluation (MENC)*, 18(Winter), 6-9.
- Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research and Evaluation. PAREonline.net*