Mitigating the Halo Effect: Managing the *Wow* Factor in Music Performance Assessment

Nathan Zoanetti

Helen Champion

Victorian Curriculum and Assessment Authority, Victoria, Australia

ABSTRACT

This article describes a project undertaken to assist assessors with recognizing and ameliorating the influence of the halo effect when multiple criteria are applied in a rubric to score music performance assessments. The halo effect occurs when impressions of the quality of a performance either overall, or against one or a subset of the criteria, unduly influence the scores assigned to subsequent criteria. The assessment context described in this article is set within the Victorian Certificate of Education (VCE), and specifically the VCE Music Investigation study. This VCE study provides a flexible pathway for students to perform, compose and research diverse traditional, contemporary, classical and popular music styles using instruments ranging from voice, tabla, recorder and ud to orchestral and wind-band instruments, instruments typically found in rock bands and digital instruments in solo and ensemble performances. This diversity must be addressed in practice through valid assessment tasks and the reliable application of assessment criteria. These tasks and criteria must be coupled with appropriate assessor training, guidance and documentation. To support this requirement, experienced assessors reviewed recorded student performances whilst annotating the performance features they referenced when allocating scores against each of ten rubric criteria. Also recorded were the phrases that assessors used to describe well-documented sources of cognitive bias in performance assessment contexts, including the halo effect. These sources of information were distilled in an effort to provide enhanced training and scoring guidance and materials to support assessment validity and reliability.

Keywords: Halo effect, rater effects, performance assessment, rater reliability, rater cognition, rater training

INTRODUCTION

It is well documented that human judgments of performance quality can be susceptible to various sources of subjectivity, bias and construct-irrelevant variance (Thorndike, 1920; Rudner, 1992). In the assessment and educational measurement literature, these matters are often described under the headings of rater effects (Wolfe, 2004) or rater accuracy (Engelhard, 1996). As Baird et al. (2013) summarized, rater effects can be grouped into *inter-rater effects* and *intra-rater effects*. Inter-rater effects pertain to systematic differences between the leniency or severity of raters, or systematic differences in the ways that raters tend to use the available score range. Intra-rater effects occur where a rater assigns one or more scores, which are to some extent influenced by factors unrelated to the assessment construct. Such factors could include the personal preferences of an assessor or the onset of assessor fatigue following a large number of assessments. Under these circumstances, the assigned scores are generally inconsistent with initial impressions of the rater's overall leniency or severity and can adversely impact assessment validity and reliability.

This article focuses on addressing one distinct inter-rater effect: the *halo effect*. The halo effect occurs when impressions of the quality of a performance either overall, or against one or a subset of the criteria, unduly influence the scores assigned to subsequent criteria (Eckes, 2009). This effect typically manifests itself in the form of overly similar scores across conceptually distinct criteria; the level of similarity or uniformity across the criterion scores would be greater than that arising from an unbiased appraisal of the performance. Another commonly cited definition of halo effect is "...a rater's tendency to assign ratees similar ratings on conceptually distinct traits" (Myford & Wolfe, 2004, p. 209). In other words, judgments about the quality of different aspects or dimensions of a performance or product are less independent of each other than would be the case in the absence of the halo effect. If the criteria each reflect distinct and important skills, then the halo effect could result in criterion scores, which do not best reflect the strengths and weaknesses of a given performance. This is notwithstanding that some criteria will be intrinsically more inter-related than others and therefore could be expected to attract relatively more similar scores even in the absence of the halo effect (Latimer, Bergee & Cohen, 2010). In a senior secondary examination context, where a specific curriculum and its embedded outcomes underpin the assessment construct and the respective criteria in the rubric, being able to recover this profile of strengths and weaknesses is arguably more valid and more diagnostically useful than allowing it to be masked by the halo effect. Whether these concerns are as important in other performance contexts, particularly those where preset scoring criteria are not applied (see Sadler, 2009), is left as an open question. Studies into the reliability of judgments of music performance quality have been reported on previously (e.g. Garman, Boyle & DeCarbo, 1991; Bergee, 2003; Latimer et al., 2010). Some such studies have focused on inter-rater reliability measures when rubrics have been

used (Ciorba & Smith, 2009). Other such studies have focused on differences between the levels of reliability of different rating or scoring forms, for example comparing rubrics to previously more commonplace adjudication forms (Norris & Borst, 2007). Yet other such studies have applied particular measurement frameworks and corresponding statistical models to analyze aspects of rater reliability (Wesolowski, Wind and Engelhard, 2015). The cited studies were predominantly evaluative in nature, with a focus on the detection and quantification of various reliability and model fit measures, leading to discussions of implications for improving subsequent assessment administrations. In the current study, the initiative to explore and preemptively address the halo effect was initially motivated by two main sources of evidence. The first was data revealing that the rubric criteria, many of which could be regarded as conceptually distinct (refer to Appendix A), attracted similar scores. The Pearson correlation values between the scores awarded for different pairs of criteria (often referred to as inter-item correlations) all exceeded 0.8 and so could be considered high. While more formal measurement frameworks and statistical models could have been applied to detect the possible presence of the halo effect and other rater effects (e.g. Myford & Wolfe, 2004), this was not within the scope of the present study. Beyond quantitative evidence, the second source of evidence around the possible impact of the halo effect arose through discussions with assessors. These discussions revealed a collective awareness that awarding full marks for all criteria when the overall impression of a performance, or aspects of it, were overwhelmingly positive was difficult to avoid. These sources of evidence, whilst by no means definitive, when taken in combination were deemed sufficient to hypothesize that the halo effect could be impacting scoring processes in this assessment context. To minimize the halo effect in operational assessment programs, Myford and Wolfe (2004) noted the important role of assessor training as a preventative strategy. They recommend efforts to raise awareness of the halo effect and its influence so that assessors can work to consciously ameliorate it. In effect they are advocating the development of effective metacognition, a capacity defined by Kuhn and Dean (2004) as “awareness and management of one’s own thought” (p. 270). It is in this spirit that the present project is being undertaken.

RESEARCH QUESTIONS

The workshop sought to provide insights about the following research questions:

1. How do music performance assessors describe the phenomenon of the halo effect?
2. How do music performance assessors describe potential strategies for mitigating the halo effect?
3. What are the evidentiary features of music performances that assessors refer to when assigning scores to each of the criteria in the prescribed assessment rubric, and how do they describe these?

4. What are the salient differences between performance qualities that distinguish one criterion score from the adjacent criterion scores in the prescribed assessment rubric, and how do assessors describe these?

This article focuses primarily on the first two questions but presents some findings relevant to other questions where they relate to the broader issue of rater effects. Ultimately, by learning how best to raise awareness of halo effects with assessors so that they could consciously avoid it, this research was intended to lead to an increase in the validity of students' scores, and would ensure that the most valid relative ranking of student achievement could also be obtained.

METHOD

A group of eight experienced assessors was invited to a workshop to review audio recordings of student performances whilst annotating the performance evidence they referenced when allocating scores against each criterion. These assessors were specialists in a range of instruments and music styles and each had between three and 20 years of experience assessing music performance examinations.

The task for the day was to listen to, score and discuss recordings of a sample of the previous year's examination performances. The workshop employed two qualitative research methodologies in sequence. The first method was similar to retrospective verbal protocol analysis as described by Ericsson and Simon (1993). This method provided an opportunity to gather information about rater cognition, by asking individual assessors to recount their thinking as they listened to and scored each recorded performance. In advance of listening to each recorded performance, assessors were asked to annotate their thinking whilst applying the scoring rubric to support this approach. This was intended to aid their recall, noting that the validity of retrospective verbal protocols diminishes as the elapsed time since the cognitive task of interest increases. The second method involved a focus group methodology (Krueger and Casey, 2009). This method extended the retrospective verbal reports by inviting group discussion about the qualitative features of each performance, and how these features influenced overall impressions of performance quality and how they could be linked to different levels of each criterion and to the expectations of the senior secondary context more broadly.

The workshop was led by the chief assessor, as would be the case for assessor training. VCAA representatives, including the authors, observed the workshop and took notes. An audio recording was made to enable follow up analysis of the verbal reports and group discussion. This process aimed to identify and record phrases and statements that described cognition and metacognition relevant to the halo effect, and relevant to other aspects of differentiating between adjacent score categories for specific criteria within the rubric.

The group discussed this evidence and noted some common language they considered suitable for adoption by other assessors, with the intention of providing a basis for criterion-referenced rather than solely impressionistic discussions of performance quality. Assessors were asked to reflect not only on the evidence they referenced, but also on the judgment processes they applied. Through this process a number of phrases were derived for describing the occurrence of, or conscious avoidance of, the halo effect. Importantly, these phrases were in the everyday language of musicians, not language couched in theoretical terminology from the assessment or educational measurement lexicon. The words and phrases were therefore considered suitable for being recast as user-friendly guidance for the broader assessor group.

As a preliminary goal, this approach aims to increase the quality of discussion amongst assessors through a focus on evidence about salient performance features and performance differences against each prescribed criterion, rather than a focus on agreeing in general terms about the overall quality of a performance. Even if assessors first come to a shared conclusion about the overall quality of a performance, the language being developed here is intended to assist assessors in unpacking the features of the performance, referenced against the relevant criteria. To facilitate this, the approach being undertaken is concerned with co-constructing an inventory of common language to support discussions of performance quality, focusing on each criterion in turn. The approach specifically targets the halo effect by adding a layer of metacognitive guidance that assessors will be asked to refer to when recording scores for each performance. This guidance is field-tested during each examination period and it is envisaged that it will be integrated into assessor training and support materials for other learning areas if there is evidence of its effectiveness. To support implementation of this additional, written guidance, a brief presentation and explication of the halo effect is now delivered as part of the regular assessor training meetings.

CURRICULUM AND ASSESSMENT CONTEXT

The VCE is the main senior secondary qualification in Victoria, Australia. Each study in the certificate is assessed by a mix of coursework and external examinations.

VCE Music Investigation Units 3–4 were introduced in 2011 to provide a flexible option for students interested in both performance and composition. Over the four years that the study has been offered, students have presented a wide-range of performance programs across traditional, contemporary, classical and popular music styles and using traditional, digital, orchestral, rock and other instruments. This diversity poses issues of equity and rigor that must be addressed in practice; there is a need for assessment tasks and criteria which support reliable scoring whilst catering for the range of instruments, styles and abilities. Further, these tasks and criteria must be coupled with appropriate assessor training and documentation. Recent assessor workshops and training meetings are

providing insights into how improvements might be made to further strengthen the equity and rigor of the assessment program.

VCE Music Investigation is a performance/portfolio style study, typically undertaken in the final year of secondary schooling. The study challenges students to explore an area of personal musical interest and develop their abilities as performers through research and composition, improvisation and/or arrangement. Students may choose any instrument or instruments for their performance and elect to work as a soloist or as a member of group. The curriculum is provided in the VCE Music study design, a framework from which schools design learning programs. Details of the curriculum and assessment specifications are available at www.vcaa.vic.edu.au

The device of a *Focus Area* is used in the curriculum to organize the student's work across the school year. The Focus Area might be about the practice of a performer or performers, a music style or genre, or the use of an instrument in a particular context. Predictably, Music Investigation students choose diverse areas of musical interest as the basis of their study. For quality assurance, VCAA reviews and approves Focus Areas early in the school year. Table 1 shows a cross-section of the Focus Areas chosen by students.

TABLE 1: sample focus areas

CLARINET	The manipulation of various clarinet techniques in romantic and neo-classical clarinet pieces.
RECORDER	How performers and composers create interest in the melodic line in a range of styles through the use of performance techniques and conventions, such as diminution, ornamentation, free embellishment, and both neo-classical and avant-garde variation writing.
VOICE	Animated Hollywood musicals: artistry and interpretation.
VIOLIN	How changes in the construction of the violin since the baroque period have allowed for the emergence of new techniques.
ELECTRIC BASS	The development of pitch-based rhythmic support seen through bass including its evolution into a melodic instrument.
TABLA	Defining North Indian Tabla repertoire in a modern Western context.

Performance examination structure

The performance examination for Music Investigation contributes 50% of the total study score. The examination requires students to present a program of at least four works that demonstrates their understanding of the Focus Area they have studied throughout the year. The maximum time allowed for a solo performance is 25 minutes. For students who are assessed as a member of a group the time varies from 25 minutes where only one student is assessed in the group, to 40 minutes where five to six students are assessed. A group may have between two and eight performers. Solo performances are assessed by two assessors, one of whom is an instrument and/or genre specialist and the other, a coordinating assessor who works across a number of instruments and/or genres. Group

performances are assessed by two or more assessors depending on the number of students to be assessed so that the performance of each assessed performer can be considered by two assessors. As with the panel for solo performance assessments, each panel includes an instrument specialist/s and a coordinating assessor.

Performance examination assessment criteria

The assessment criteria for the Music Investigation end-of-year performance examination are derived from the key knowledge and key skills described in the study design. The criteria are published on the VCAA website along with explanatory annotations. This strategy supports transparency and connection across the learning process. Only criteria 2 through 11 are considered here. The first assessment criterion is related to compliance as opposed to performance quality and was therefore not relevant for this work. These criteria are listed in Appendix A. Each criterion is scored out of a maximum of 10 by each assessor. Assessors are encouraged to discuss the features of performances with relation to the published criteria, but they are expected to assign scores independently of each other.

FINDINGS

Language used by assessors to describe the halo effect

Most of the discussion pertaining to the halo effect was directed towards the top end of the performance distribution. This was consistent with the sentiment that assessors believed the halo effect could emerge most readily when they were drawn in by a high quality performance. One assessor noted that when their initial impression of a performance was that it was “amazing”, it took a more conscious effort to refocus on unpacking the evidence against the criteria. In their words: “That’s my challenge, to move away from that [initial impression] and say, hey, but by this criterion...” This realization is exactly the sort of metacognitive process that this workshop and ongoing training activities aimed to foster.

A variety of phrases were informally uttered or more formally proposed by assessors, based on their reflections after listening to a number of high scoring recordings of examination performances. These phrases are recounted in this section and have been separated thematically.

It was agreed that central to mitigating the halo effect was the idea that each criterion must consciously be judged or referenced as part of unpacking evidence from an overall performance. Consequently, the first statement to be noted for the inventory was:

Judge each criterion on its merits

This statement encapsulates the intent of the workshop but also provides a link to the annual training that assessors undertake.

The next statement captured here references the role of an assessor juxtaposed with the usual roles that assessors as musicians and teachers undertake, that contribute significant experiences to the expertise they bring to bear in their judgments:

As musicians and teachers you sometimes [experience] the emotive response...[but in this context you need to] bring the assessor into play...

Again, this statement adds value to the inventory as both a truism and a reminder of the assessment principles that underpin the annual assessor training workshops.

A third set of statements, this time targeting exceptional performances, particularly those that were exceptional in at least one distinct respect, included

I'll never hear anything like this again, but then, if they haven't met that criterion, ...

This is the best I've seen in 10 years. I may never see something like this in another 10 years, but is it actually full marks for this [criterion]?

It is everybody's challenge...to stand back...because that sort of response...I hear that in myself so often when I've heard amazing performances.

Aspects of what I've heard today are the very best I've heard but...

These statements demonstrate assessor awareness that whilst the context for the performances is an examination to award scores that contribute to student achievement in a senior secondary certificate of education, a proportion of the student cohort demonstrate levels of achievement that are commensurate with those expected in the early years of tertiary music study. The statements also demonstrate awareness that the assessment process is criterion referenced, and that the various criteria set expectations for conceptually distinct aspects of musical performance.

A fourth set of statements focused on the idea that the awarding of a maximum score for a criterion should trigger an explanation in terms of the evidence apparent in the performance.

Challenge each of your 10s

Justify your 10s

These statements are simplistic yet they describe in everyday language the sorts of stop-checks or metacognitive guidance that might mitigate the halo effect where it seemed most likely to emerge; that is, in instances where a performance has the wow factor in this context. After being considered collectively, a number of statements of metacognitive guidance were distilled to cover the key themes and to leverage the language captured through the group discussions.

Metacognitive guidance to mitigate the halo effect

Following a review of the words and phrases used by assessors in relation to the halo effect, an initial set of short and straightforward guidance statements was constructed. Again, the focus was on the higher scores for each criterion. At the time of writing, these were:

TABLE 2: Assessor advice card

Aspects of this performance were amazing but I still need to:

1. *Assume the role of Assessor*
2. *Judge each criterion on its merits*
3. *Justify each of my 10s*

These statements were drafted as a basis for developing an extra resource, possibly in the form of a cue card, for assessors to bring with them to the performance examinations. In the words of one assessor, these were described as *checking/red-light sentences*, which could be referred to and could be used to prompt robust discussion between assessors observing the same performance. Given the need to contain the administrative burden for assessors, one goal was to ensure that this guidance remained succinct.

Unpacking performance evidence against the criteria

During the workshop, raising awareness about the halo effect and providing metacognitive guidance to avoid it were seen as productive initial strategies for safeguarding reliability and validity. In parallel with the development of these strategies, detailed discussions of performance evidence were facilitated after each recorded performance had been listened to. The purpose of these discussions was to reinforce common expectations around each criterion score point. This was viewed as important for supporting assessors in making consistent and evidence-based determinations of the criterion scores that each performance should receive. Of particular importance was discussion about the nature of salient differences in performance quality evidence that would strongly indicate the appropriateness of a particular score category over the adjacent score categories. A sample of the language used by assessors to describe the quality of evidence against the various criteria is tabled in the third column of Appendix A. Eliciting and documenting this extra layer of detail about performance evidence is part of a larger program of work focusing on rater effects, but within which this exploration of the halo effect has taken place. All of these streams of work are complementary as they each promote the centrality of the criteria and aim to forge valid and reliable links between the quality of evidence apparent in observed performances and the patterns of scores allocated.

As the assessors reviewed the range of performances, discussions focused on two issues often evident in performances with a wow factor. These issues were agreed to have relevance irrespective of differences in instruments, styles, genres and repertoire selected by the students. First was the matter of *glitches* and second was the matter of repertoire selection. Discussion about glitches focused on evidence in the performance such as occasional wrong notes or slight inconsistencies in intonation, for example:

Range jumps impressive given their difficulty

A couple of stumbles don't constitute a reduction in marks. It's still a 10 because...

Pretty much flawless...in execution and cleanness

They're 10s and they've shown that so many times that we have to give them a 10, or do we pull it back because they did in fact fall down here or there very slightly?

Can a student make up for errors?

The presence of this phenomenon in live music performance was acknowledged and the group agreed on guidelines to assist in determining the level at which glitches should impact on scoring. The rationale underpinning these guidelines focused on the wow effect and the extent to which the glitches or inaccuracies interrupted the flow and intensity of the performance for the audience. For example, in relation to accurate pitching such as pitching of high notes by a trumpet player, it was agreed that an occasional lack of clarity did not have a disruptive effect on the listener and that a performance where the student did make some small errors could communicate a wow effect. However, repeated examples of missed high notes could cause the listener to focus on the high notes and not on the music, in fact creating an effect of *wow*, *but...*

The second issue of repertoire selections for the examination program included consideration of the range of performance and expressive techniques available in the music works selected by the student to demonstrate their skill against the criteria. For example:

Certain sameness to program in terms of dynamic and technique

Didn't allow [him] to fully exploit the criteria

Student didn't choose to push [himself]

The skill in the part [he] demonstrated was very high but [he] could have allowed a section to show a higher level of technique.

These comments acknowledge that the qualities a performer can demonstrate in performance are in part dictated by the repertoire selected and that whilst students and their teachers need to consider how repertoire offers opportunities or imposes limitations to demonstrate skill against each of the assessment criteria, the assessors' task is to score the performance they hear. It was agreed that in some cases an initial response might be "wow" but consideration of the range of qualities evident in the performance could mitigate this response in relation to one or more criteria.

CONCLUSIONS

This paper describes initial steps undertaken to strengthen support for assessors around known challenges in judging the quality of complex, extended performances against prescribed criteria. More specifically, strategies for addressing the rater effect known as the halo effect have been outlined in some detail. Whilst the implementation of these strategies is in its early stages, participating assessors anecdotally reported that: assessor support for the project and direction is strong; assessors feel comfortable with the process because the additional material is being derived from authentic conversation and not imposed; and, the discussions are acknowledging the diversity of performances the students present in response to the curriculum.

This program of work continues to build on and refine the assessor training programs. Guidance materials need to be as rigorous as possible and responsive to what is understood and documented about threats to reliability and validity in the judgment of performance quality. Other strategies to mitigate rater errors will also be implemented concurrently to strengthen the overall approach. The primary method, which is already a staple of assessor training, relies on the use of exemplar recordings specifically chosen to cover different regions of the available score range. Some recording might be chosen with specifically targeted performance idiosyncrasies, or musical features that are relatively new, for example arising through the use of new technologies in music performance examinations. These exemplars can be used to identify assessors who do not conform to the judging standard, and can be used to identify a range of raters errors such as leniency or severity, halo effect, restriction of range, and central tendency at the point of assessor training. Consideration will also be given to how the effectiveness of this evolving program might be evaluated using longitudinal, quantitative data from the assessment program, should such an evaluation be tractable.

REFERENCES

- Baird, J., Hayes, M., Johnson, R., Johnson, S., Lamprianou, I. (2013). *Marker effects and examination reliability: A comparative exploration from the perspectives of generalizability theory, Rasch modeling and multilevel modeling*. Oxford University Centre for Educational Assessment. Retrieved October 1, 2013 from <http://www.ofqual.gov.uk/files/2013-01-21-marker-effects-and-examination-reliability.pdf>
- Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education*, 51, 137-149
- Ciorba, C. R., & Smith, N. Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research in Music Education*, 57, 5-15.
- Eckes, T. (2009). *Many-facet Rasch measurement*. Retrieved October 1, 2013, from <http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf>.
- Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33(1), 56-70.
- Ericsson, K. A. & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge, MA.

- Garman, B. R., Boyle, J. D., & DeCarbo, N. J. (1991). Orchestra festival ratings: Interjudge agreement and relationships between performance categories and final ratings. *Research Perspectives in Music Education*, 2, 19–24.
- Krueger, R.A., & Casey, M.A. (2009). *Focus groups: A practical guide for applied research* (4th Ed.). Thousand Oaks, CA: Sage Publications.
- Kuhn, D. & Dean, D. (2004). A bridge between cognitive psychology and educational practice. *Theory into Practice*, 43(4), 268–273.
- Latimer, M. E. J., Bergee, M. J., & Cohen, M. L. (2010). Reliability and perceived pedagogical utility of a weighted music performance assessment rubric. *Journal of Research in Music Education*, 58(2), 168-183.
- Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E.. V. Smith, & R. M. Smith (Eds). *Introduction to Rasch measurement* (pp. 518–574). Maple Grove, MI: JAM Press.
- Norris, C. E. & Borst, J. D. (2007). An examination of the reliabilities of two choral festival adjudication forms. *Journal of Research in Music Education*, 55, 237–251.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 469–477.
- Rudner, L. M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation*, 3(3). Retrieved September 22, 2013 from <http://PAREonline.net/getvn.asp?v=3&n=3>.
- Sadler, R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159-179.
- Victorian Curriculum and Assessment Authority (VCAA), 2010. *VCE Music Study Design*, Retrieved October 8, 2013, from <http://www.vcaa.vic.edu.au/Pages/vce/studies/music/musicinvestigateunits3-4/musicinvestigateindex.aspx>;
- Victorian Curriculum and Assessment Authority (VCAA), 2010, *VCE Music Investigation Performance Examination Specifications and Criteria*, Retrieved October 8, 2013,

from <http://www.vcaa.vic.edu.au/Documents/exams/music/mus-invest-specs-crit-w.pdf>

Wesolowski, B., Wind, S.A., & Engelhard, G. (2015). Rater Fairness in Music Performance Assessment: Evaluating Model-Data Fit and Differential Rater Functioning. *Musicae Scientiae*, 19(2), 147-170.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35-51.

APPENDIX A

VCE Music Investigation Criteria 2–11 with published annotations and sampled assessor comments from the assessor workshop and training meetings

Criterion	Annotations	Sample inventory of language used by assessors (in development)
2. Skill in performing accurately and with clarity	<ul style="list-style-type: none"> <i>precision of pitch, rhythm, articulation, dynamics and phrasing as notated or within the styles performed</i> <i>clarity of passage work, tone production, timing, phrasing and articulation as appropriate to each instrument</i> 	Program has a high degree of difficulty High level of accuracy and consistency in every work Possibly a few incidental errors 'in the pocket' impeccable precise focused well-projected authoritative consistently centred intonation flawless clarity executive very secure
3. Skill in performing a range of techniques with control and fluency	<i>Within the context of the Focus Area:</i> <ul style="list-style-type: none"> <i>the program selected contains a range of instrumental or vocal techniques that reflect the instrument(s) expressiveness and versatility and are performed with control and fluency</i> <i>techniques are performed securely and with dexterity.</i> 	Mastery demonstrated Thorough Comprehensive Complete Sophisticated Extremely safe Technical prowess Virtuositic Great versatility Pizzazz and understanding demonstrated Outstanding dexterity and flexibility
4. Skill in producing a range of expressive tonal qualities relevant to the Focus Area	<i>the ability to perform a range of tonal qualities throughout the program</i> <ul style="list-style-type: none"> <i>the quality and projection of tone production throughout a large dynamic range as appropriate to the instrument and relevant to</i> 	Broad range of tone colors, extremely well balanced Almost professional tone across the full dynamic range Piercing <i>ppp</i> Projection Enticing range of tonal colors Sensuous

Criterion	Annotations	Sample inventory of language used by assessors (in development)
	<i>the selected repertoire</i> <ul style="list-style-type: none"> • <i>the ability to vary tone in a way that is appropriate within the Focus Area</i> 	Very expressive, reflecting emotional response to the music
5. Skill in articulating and phrasing	<i>Within the context of the Focus Area:</i> <ul style="list-style-type: none"> • <i>creation of purposeful shape through artistic variation of expressive elements including tonal quality, tempo, phrasing, articulation, dynamics and texture</i> • <i>communication beyond the source material (notation and/or reference recording of a work by using appropriate nuances and performance techniques</i> 	Sophisticated, well-defined, refined, elegant, informed Exceptional clarity Dexterous changing between [articulation – tonguing, bowing] techniques Phrasing – wide variety of shaping Very clean and appropriately weighted Sensitive phrasing Wide range of dynamics with transitions thoughtfully managed Different dynamic palettes evident across the program
6. Skill in differentiating the structures and textures within each work of the Focus Area	<ul style="list-style-type: none"> • <i>the performance communicates understanding of the variety of structures and textures contained in the program</i> • <i>effective differentiation of the structures and textures within each work of the program</i> • <i>the variety of structures and textures in the program is representative of structures and textures that are characteristic of the Focus Area</i> 	Highly informed Awareness of issues evident Depth of understanding demonstrated Excellent sense of direction using contrasts to differentiate
7. Skill in differentiating the musical lines in the selected works as	<i>As appropriate to the instrument and/or instrumental context:</i>	Consistent Fluent Synchronization

Criterion	Annotations	Sample inventory of language used by assessors (in development)
appropriate to the Focus Area	<ul style="list-style-type: none"> • <i>the delineation of the main musical ideas and balancing of the musical parts within each work</i> • <i>internal communication and synchronization in the performance parts.</i> 	Sophistication and sensitivity displayed in balancing the musical lines
8. Skill in presenting an interpretation of the works that is informed by historical and/or contemporary practices and conventions relevant to the Focus Area	<ul style="list-style-type: none"> • <i>across the program, demonstrating understanding of performance practice/s relevant to the Focus Area to present informed interpretations, including demonstrating an understanding of original performance practices of each work presented; and/or</i> • <i>interpreting works within the Focus Area using current stylistic conventions</i> 	Convincing depth Subtlety Nuances Study of past players in the style is evident Insightful Perspicacious Depth of understanding
9. Skill in performing with musicality through creativity and individuality	<ul style="list-style-type: none"> • <i>demonstration of individuality and creativity within performance practice appropriate to the Focus Area. This may involve improvisation and/or a new realization or interpretation</i> • <i>perform in a way that demonstrates elements of personal interpretation and communicates understanding of the Focus Area</i> 	playing the work not the instrument engaging, captivating satisfying integrity command of technical and expressive aspects of the performance stylistically appropriate throughout
10. Skill in demonstrating how the musical works in the	<ul style="list-style-type: none"> • <i>the cohesiveness of the program in relation to the Focus Statement</i> 	a diversely varied and stylistically aware program relevance to the focus area

Criterion	Annotations	Sample inventory of language used by assessors (in development)
program are representative of the Focus Area	<ul style="list-style-type: none"> • <i>the degree to which the student succeeds through their performance in making connections between works in the program and the Focus Statement such as factors that unify the works or realizing characteristics relevant to the Focus Area</i> • <i>the extent to which the program demonstrates diversity appropriate to the Focus Area</i> 	highly evident insightful, informed interpretation
11. Skills in the presentation of a cohesive program relevant to the Focus Area	<ul style="list-style-type: none"> • <i>use of relevant musical elements of performance presentation</i> • <i>the flow of the program (which may include brief verbal commentary)</i> • <i>effective musical communication with audience</i> 	Commanding presence Fluent, highly organized presentation Polished Excellent poise The whole package of a well-structured program presented in a convincing manner Issues for consideration: <ul style="list-style-type: none"> - endurance - performance contouring - maximizing degree of difficulty considerations