Implicit and Explicit Assessment in Music Performance: Contemporary Attitudes and Concerns

Filip Maximilian Sievert Lindeskog

Master Thesis – Talent Development and Creativity

s3193608
July 2022
Department of Psychology
University of Groningen
Examiner/Daily supervisor:
Susan Niessen

Abstract

In academic research, structured information collection and mechanical judgment (explicit methods) have been widely accepted as superior to unstructured information collection and holistic judgment (implicit methods) in terms of validity and reliability. However, there generally exists a consensus gap between science and practice concerning which methods are best suited for assessment; practitioners typically have numerous objections to the use of explicit methods, making it challenging to implement evidence-based methods into practice. In the field of creative assessment, particularly music assessment, this issue is further complicated due to a lack of relevant research, and there is presently a great reliance on subjective and implicit methods. To better understand the perspectives of music assessors regarding the suitability of different assessment methods, this study analyzes seven individual interviews and one group interview with staff at a Dutch music conservatory using qualitative thematic analysis. It was found that music assessors viewed intuition and consensual assessment as integral to music assessment, were generally open-minded to using structured information collection methods, found mechanical judgment restrictive, unfeasible, and unsuitable for music assessment, and desired some flexibility in the weighting of subcomponents. Importantly, some music assessors reported a prevalence of ‘bypassing’ mandated mechanical judgment by adjusting the weights so that they align with a preceding holistic judgment.

Keywords: music assessment, thematic analysis, holistic judgment, mechanical judgment, consensual assessment, rubrics

Implicit and Explicit Assessment in Music Performance: Contemporary Attitudes and Concerns

What makes music ‘good’ or ‘bad’? When is it ‘excellent’ and when is it ‘incredible’? And, perhaps most importantly, how can one tell? Questions like these are not only interesting in their own sake but are central to understanding how music assessment works, and how it can be improved. However, a wide variety of assessment contexts are still characterized by a gap between science and practice in regards to which methods are considered most suitable (Neumann, Niessen, Tendeiro, & Meijer, 2021) which underlies an enduring tension of disagreement between practitioners and policymakers (Colwell, 2008; Denis, 2017). Additionally, research specifically targeting practitioners’ attitudes towards assessment in the field of creativity, such as in music performance, has been somewhat lacking (Moran, 2010). With this in mind, this study will conduct a qualitative thematic analysis to investigate the attitudes of contemporary music assessors on different approaches to music assessment, aiming to better understand which methods are considered suitable for this field, and, perhaps equally importantly, which are not considered suitable and why.

Assessing Performance

At its core, assessment is comprised of collecting, analyzing, interpreting and combining information in order to make well-founded decisions (Brookhart & Nitko, 2014). Two common areas in which assessment of performance plays a central role are education and employee selection (Kuncel et al., 2013). These fields regularly employ empirically supported assessment methods such as admission tests and work samples (Cook, 2016; Dalal et al., 2020; Niessen et al., 2016). A wide variety of approaches to assessment exist (McQuarrie & Sherwin, 2013; Rohwer, 1997; Russell & Austin, 2010), and what is considered the most appropriate method can vary with the field in question. Regardless of the context in which assessment is performed, it is critical for the quality of the subsequent decisions that the method exhibits good validity and reliability (Carmines & Zeller, 1979). In general, validity in assessment refers to whether the method meets its intended aim, and reliability refers to the consistency of the method over different trials or assessors (Leung, 2015). Importantly, while a method can be reliable but not valid (e.g., provide consistent results but failing to capture what is intended), a method cannot be valid but unreliable; for a method to be considered valid, it must per definition also be consistent.

Assessing Creative Performance

For creative performance, and music performance in particular, the literature on assessment is somewhat slimmer than for the educational and occupational contexts as the field of creative performance assessment has been comparatively limited in scope (Moran, 2010) and less funded (Runco et al., 2015), but the importance of appropriate assessment in music performance remains equally central (Wesolowski, 2012). What conventional educational and occupational contexts generally have in common, that does not apply equally to more creative contexts such as music education, is the strive for standardization, where some answers or responses are clearly considered correct while others are considered incorrect (Tienken, 2016). For example, a mathematics exam may accept different methods, but they must all lead to the same answer, while a sales job may accept different strategies, but they, too, must all lead to increased revenue. These characteristics notwithstanding, there are aspects of education and work assessment that also require some subjective interpretation. For example, while educational admission decisions to a great extent rely on general, standardized aptitude tests and previous educational results, there is a growing emphasis on less rigid, performance-based assessment as opposed to pencil-and-paper tests (Lai, 2011). Within the occupational setting, an interview may seek to determine if a candidate has ‘good’ qualities for the job, but what constitutes ‘good’ qualities may vary widely depending on the organizational role of the person asked (Conway & Huffcutt, 1997), and employee selection remains characterized by an adamant reliance on subjectivity and intuition (Highhouse, 2008).

Challenges of Music Assessment

In general, assessment of creative performances such as those made in the context of music is even less straight-forward than assessment in educational and occupational contexts due to several reasons. As mentioned, there is generally no agreed upon absolute standard to which one can compare a performance and rate music based on its relative likeness to this standard, making it difficult to determine the appropriateness of assessment outcomes. While this point also holds true to some extent for occupational contexts, contributing to the generally low reliability of job performance assessment (Conway & Huffcutt, 1997), reaching consensus on assessment between music educators and policymakers has proven even more difficult (Denis, 2017).

Further, assessment in music tends to reward divergence of expression, i.e., departing from the norm (Wesolowski, 2012), contradicting an idea of striving towards some absolute model of performance. This contrasts with the traditional educational context, where mainly convergent thinking is taught and, perhaps as a result, creative thinking often diminishes (Noddings, 2013). Consequently, since music assessment operates with this relative lack of objective standards, it becomes particularly influenced by subjectivity (Radocy, 1986; Wesolowski, 2012) which often results in a negative impact on the validity and reliability of assessments relative to more objective and standardized procedures (Kaufman, 2013). For example, subjective judgment is more likely to be influenced by criteria irrelevant to music performance quality, such as behavior, attitude and personality factors (McCoy, 1991), to result in inconsistency in the importance placed on different subcomponents of the performance (Karelaia & Hogarth, 2008), and to lack the transparency which is required for assessees to understand what is expected of them (Wesolowski, 2012).

Altogether, music assessment generally lacks both standardization of the assessees' creative products, product meaning an object or a performance, and of the assessors' approaches to assessing these products. First, there is an exceptionally high variability in performances: assessees can choose between a seemingly infinite combination of different instruments, songs and styles. Second, this is coupled with an exceptionally high variability of interpretation: assessors reliant on subjectivity in judgment, and different assessors could evaluate a given performance in many different ways. The lack of standardization in both of these aspects make music assessment a quite complex task (Denis, 2017; Wesolowski, 2012). However, research on educational and organizational assessment suggest that performance-based assessments, requiring at least some interpretation and subjective judgment, can be improved through structuring the way performance is judged and how judgments are integrated to form a final assessment (Boyle & Radocy, 1987; Kishk et al., 2005; Plakiotis, 2017). While the subjective aspect of individual taste may perhaps be inseparable from music assessment, one step towards implementing some level of structure could be found in employing more explicit approaches and tools that support increases in validity and reliability (Wesolowski, 2012). To better understand how to carry out this implementation, one must consider the concepts of structured versus unstructured information collection, and holistic versus mechanical information judgment.

Structured Versus Unstructured Information Collection

Information, also referred to as data, can be collected in a structured or unstructured manner; here, structure refers to the degree of standardization for which subcomponents of performance are to be judged, and how. In other words, structured information collection aims to assess the same type of information for every performer, in the same manner (Huffcutt et al., 2013). One method of increasing consistency for the type of information to be assessed is disaggregating performance, i.e., dividing it into its constituent parts and explicitly assessing each of these parts, which has been shown to increase inter-rater reliability (Arkes et al., 2006; Conway et al., 1995). While this is a necessary part of structured information collection, unstructured information collection does not include this step (Bergkamp et al., 2020). For increasing consistency in how the information is assessed, one could use explicit and structured assessment tools such as Likert scales, where one numerically rates a statement by level of agreement or disagreement, or criteria-specific scales, such as rubrics (Wesolowski, 2012). Tools like these can improve the quality of assessment by introducing a degree of standardization to the procedure.

An example of structured information collection could be to listen to a series of music performances, taking into account a predetermined set of subcomponents that are deemed relevant. For example, violin performances could comprise intonation, rhythm, and expressivity. In assessing violin performances, assessors would explicitly pay attention to and separately rate these subcomponents in every performance according to a predetermined assessment tool. An example of unstructured information collection could be to listen the same series of violin performances and assess each performance as a whole, without necessity of formally discerning subcomponents or using an explicit assessment tool.

Holistic and Mechanical Judgment

The method of judgment, also referred to as combination, can vary in its level of explicitness and standardization. It concerns how different subcomponents of performance are combined to form an overall judgment. Holistic judgment, often referred to as clinical or intuitive judgment, entails relying upon expert knowledge to combine information for an assessment (Dawes et al., 1989). Colloquially, it could be described as intuitively determining what is important and combining these relevant factors “in your head” to reach a final assessment, whether that be a binary decision of yes or no, or a more continuous rating on a scale. While holistic judgment can be based on separately specified subcomponents of performance, the final judgment is still determined intuitively and as such an explicit information combination policy is not present.

An alternative to the above is mechanical judgment, often referred to as actuarial or statistical judgment, and it differs from the holistic approach on several important points. First and foremost, a defining feature of mechanical judgment is that the combination of information always relies on a formal procedure, such as an algorithm or set of rules (Dawes et al., 1989). In other words, it entails using a predetermined formula to combine the relevant information for an assessment rather than doing so intuitively, as in holistic judgment. Second, the input of the formula consists of separately and quantitatively rated subcomponents. As such, mechanical judgment necessarily includes formally defining different subcomponents of what is being assessed in contrast to holistic judgment in which this is a matter of choice. Importantly, the subcomponents are also assigned weights in the formula based on how important they are deemed to be; using the earlier example of a violin performance and its subcomponents, intonation could account for 25%, rhythm for 35%, and expressivity for the remaining 40% of the total assessment score. In this manner, the mechanical process aims to increase validity and reliability by reinforcing the consistency of the information combination procedure. Naturally, how the subcomponents are assessed will still be influenced by subjective interpretation, but even partially standardizing the combination procedure has been shown to mitigate some of the subjective variation between assessors (Kuncel et al., 2013).

Overall, a solid foundation of empirical evidence suggests that mechanical judgment is superior to holistic judgment in terms of validity and reliability, in particular because it mitigates some of the inherent subjectivity in holistic approaches (Dawes et al., 1989; Grove & Meehl, 1996; Grove et al., 2000). This evidence spans a wide range of contexts such as graduate school success (Wiggins & Kohen, 1971), psychiatric diagnosis (Brown et al., 1989) and criminal behavior (Hall, 1988). Worth mentioning, however, is that while mechanical judgment may aspire to be the more objective alternative, subjective judgment often enters the equation through the necessity of deciding upon what subcomponents to use and how much weight is given to each. Alternatively, the subcomponents and weights can be derived from statistical methods when applicable (Dawes, 1979), but how to select these statistical methods still necessarily involves some degree of subjective judgment.

Applied Approaches of Information Collection and Judgment Policies

A commonly used assessment approach in creative contexts such as music assessment is the Consensual Assessment Technique (CAT; Amabile, 1982), based on unstructured holistic judgment. The fundamental assertion of CAT is that the best available judgment for the quality of a creative product are the opinions and judgments of the current experts in the relevant field (Baer & McKool, 2009). In other words, those who are most likely to provide an accurate evaluation of a creative product are those who have already accumulated a high level of expertise relating to this category of products. While CATs application originally lay in conducting creativity research, Baer and McKool (2009) suggests that it is well-suited for virtually any field of creativity evaluation, from more concrete phenomena such as scientific research designs and theories to more abstract ones such as creative products and performances.

Baer & McKool (2009) describes a typical CAT procedure as beginning with a person being asked to create something, such as a music performance, and subsequently having the aforementioned experts assess it. The experts conduct these assessments independently of each other, as not to contaminate the results by social influence, and the phenomena in question to be measured is the creative product itself, rather than notions of some underlying creativity-driving traits in the person or product. Whatever assessments are made should be represented by a point on a scale, such as a grade, optimally consisting of at least three points to allow for sufficient range in the ratings. Ultimately, an independent assessment is made by each expert holistically and the final assessment reflects a combination of these, e.g., by averaging.

While CAT offers the advantage of generous flexibility, its validity and reliability has been debated with a high level of discrepant research findings. The array of evidence presented by Baer & McKool (2009) suggests high agreement between experts (Conti et al., 1996; Hennessey, 1994; Kaufman et al., 2004; Runco, 1989) that also remains quite stable over time (Baer, 1994). Nevertheless, CAT is based on unstructured holistic judgment; as earlier mentioned, the inherent subjectivity of this approach has been widely critiziced for its low validity and reliability in many domains, in great part due lack of structure and standardization regarding what information is considered and how it is rated (Dawes, 1979; Dawes et al., 1989; Grove & Meehl, 1996). However, to the author’s best knowledge, no studies have directly compared CAT to a structured, mechanical approach, and thus no definitive conclusions can be drawn about this. As it stands, CAT remains the most common framework for evaluating creative performance, such as music (Baer & McKool, 2009).

An alternative to CAT in music assessment is the emerging approach of rubrics, a structured performance scale with criteria that are tailored to a specific context (Wesolowski, 2012). When developing rubrics, one generally breaks down the product to be assessed into its essential subcomponents, complete with formal descriptions for each level that can be achieved on a scale (Stevens & Levi, 2005). In addition, rubrics can be shaped in a holistic or analytic fashion (Quinlan, 2006). In holistic rubrics, the creative product is assessed based on its subcomponents ultimately only rated on a single scale, therefore only providing a single grade that represents the assessor’s impression as a whole. An advantage of holistic rubrics is that they are often broad enough to be adaptable to a range of assessment situations; however, the lack of specificity prohibits detailed information about the creative performance

(Wesolowski, 2012). In analytical rubrics, the assessor takes all subcategories into account by having separate rating scales for each, and the final assessment is derived from the grade on each subcomponent, computed by a formula. While the tailoring of this approach allows for the more detailed information that holistic rubrics lack, the consequence is loss of flexibility, making a given analytical rubric only suitable for a quite narrow range of contexts compared to a holistic one (Wesolowski, 2012). Utilizing rubrics, especially analytical ones, has been shown to result in higher validity and reliability compared to unstructured approaches (Baptiste, 2008; Ciorba & Smith, 2009; Latimer et al., 2010; Norris & Borst, 2007) and could therefore constitute a promising method for improving the quality of music assessment.

As may be apparent, the two forms of rubrics are reminiscent of the broader holistic and mechanical approaches to assessment in general. While the core of rubrics primarily addresses the method of information collection rather than combination, analytical rubrics are a suitable tool for subsequent mechanical combination as it provides the subcategories and respective ratings necessary for a subsequent formula-based final assessment.

Objections to Structured and Mechanical Judgment

While academic research presents strong evidence for the superiority of structured information collection and mechanical judgment in a wide range of assessment contexts, real-world assessments in, e.g., educational, occupational, and clinical contexts are still primarily based on holistic methods with varying degrees of structure (Kuncel et al., 2013; Ryan & Sackett, 1987), and this subjectivity extends to music assessment as well (Wesolowski, 2012). As a whole, it seems that the implementation of more explicit structure, especially when coupled with mechanical judgment, is met with a degree of resistance from some practitioners; Dietvorst et al. (2018) succinctly labeled this as “algorithm aversion”. What are the underlying causes for this aversion?

First, an objection to the use of structured mechanical judgment could be that practitioners generally have a desire of accounting for the uniqueness and personal contexts of individuals (Longoni et al., 2019; Newman et al., 2020); connected to this idea of uniqueness, they are prone to believe that they can recognize and assess exceptions to the norm in a way that mechanical methods could not (Dietvorst et al., 2018; Guay & Parent, 2018; Hoffman et al., 2017). Perhaps given this notion of unique individuals and contexts coupled with practitioners' ability to implicitly capture this uniqueness, it has been suggested that the use of unstructured holistic judgment better fulfills practitioners' need for autonomy, which is described as being fulfilled "when people experience choice and control over processes" (Neumann, Niessen, Tendeiro, & Meijer, 2021). According to Self-Determination Theory (Deci & Ryan, 2000), autonomy is a fundamental human need, and studies exploring interventions for incorporating mechanical judgment in assessment have found that the most effective outcomes were reached when assessors' autonomy was strengthened (Kaplan et al., 2001; Neumann, Niessen, & Meijer, 2021). As such, practitioners' aversion to mechanical judgment may be partially rooted in a percieved threat to their need for autonomy.

Second, a reason for resistance to structured mechanical approaches may be found in the widespread belief that people can become near-perfectly proficient in making intuitive assessments given enough experience (Highhouse, 2008). In general, empirical evidence has argued against this notion (Camerer & Johnson, 1991; Dawes et al., 1989; Grove et al., 2000; Sherden, 1998), but its enduring prevalence may underpin the earlier discussed practitioner reliance on expertise in assessment (Dawes, 1979; Kuncel et al., 2013), which becomes particularly central in the Consensual Assessment Technique (Baer & McKool, 2009).

Third, a set of objections can be found in Dawes (1979) discussion, and subsequent rebuttal of, common arguments against mechanical judgment. One of them concerns individuals' personal experiences of successes following use of holistic judgment. One may be able to vividly recall several examples of when someone made an excellent assessment "off the top of their head". Surely these instances occur, but one has to take into account the well-documented phenomenon of confirmation bias, where wanting to believe in something drives us to seek out confirming evidence and dismiss rejecting evidence (Nickerson, 1998). Naturally, proponents of mechanical judgment are also subject to this phenomenon, but their assertion rests on a more solid foundation of empirical evidence. In conjunction with the above, personal experiences of holistic successes may be influenced by the availability heuristic (Tversky & Kahneman, 1974), i.e., the ease of accessibility for memories of successes as they stood out, paired with the relative difficulty of accessing memories where holistic judgments produced insignificant outcomes. As Nisbett et al. (1976) demonstrated, vivid single instances such as these can have a greater influence on attitudes than more rigorous statistical data based on far more instances. Finally, a recurring objection concerns the idea that complex phenomena, such as individuals and the performances they produce, cannot be completely accurately represented by numbers. This much is true; however, neither can they be with holistic methods. In essence, every kind of measurement is a form of approximation, and it is desirable that we employ one that is less incorrect than others.

The Current Study

Given the complex nature of assessment in music, and practitioners' pervasive resistance to the more empirically-supported structured mechanical judgment, the process of improving validity and reliability of assessment procedures must take into account the generally unstructured and holistic tradition of creative performance assessment. While academic research findings may contribute greatly to this development, the perspectives of those who assess music in practice will have to be considered as to eventually reduce the gap between practice and science, and a first step in this would be understanding their attitudes towards different assessment methods. With this goal in mind, this study will conduct a qualitative thematical analysis of contemporary music assessors’ viewpoints on structured versus unstructured and holistic versus mechanical approaches to assessment. The aim is to clarify the current challenges of translating academic research about music assessment into practical applications, and explore the limitations of different approaches to information collection and judgment within this context. The research questions are as follows:

1. What are the attitudes of people working with music assessment concerning structured versus unstructured and holistic versus mechanical approaches?
2. What are the future outlooks of people working with music assessment concerning structured versus unstructured and holistic versus mechanical approaches?
3. What are the concerns of people working with music assessment concerning obstacles to improving assessment methods?

As this study is exploratory in nature, formal hypotheses are not specified. However, based on the previously discussed high prevalence of resistance to mechanical judgment, it is expected that a majority of participants will express negative views towards its implementation.

Method

This study will analyze transcripts of seven individual interviews and a group interview by employing thematic qualitative analysis, based on the framework of Braun and Clarke (2006). They describe it as a method for “identifying, analyzing and reporting patterns (themes) within data” (p. 3), with its benefits including flexibility, ease of use, and independence of specific theoretical commitment. Its steps and key components are summarized in the Analysis section.

Participants

A total of ten individuals participated in the study, with one individual participating in both an individual interview and the group interview (see Table 1 for demographic details). Participants were recruited among the staff employed at the Prins Claus Conservatory in Groningen, The Netherlands. The recruitment method was convenience sampling through email distributed by the head of the classical and jazz departments. In the email, recipients were informed about the study and could choose to opt in. Recruitment and interviewing were carried out by two people, with the group interview and two individual interviews being conducted by the author, and the remaining five individual interviews by a psychology

Table 1

Participant Characteristics

Department	Jazz		Classical	
	n	%	*n*	%
Gender				
Male	6	75.00	1	50.00
Female	2	25.00	1	50.00
Experience (years)[1]				
<5	0	0	2	100.00
5-14	1	12.50	0	0
15-24	4	50.00	0	0
>24	3	37.50	0	0
Experience (mean)	23.90		3.00	
Age (mean)	58.80		31.50	

[1] Experience working with music assessment

master student colleague. Individual participants are labelled and referred to in this study with letters A through J, where the subscript “$_i$” denotes an individual interview and the subscript “$_g$” denotes the group interview. Participants A_g through D_g, E_i, and F_i were interviewed by the author, and participants A_i and G_i through J_i were interviewed by the student colleague. All participants signed an informed consent form. No compensation was offered for participation. The study protocol was approved by the Ethical Committee of Psychology, University of Groningen.

Materials and Procedure

The interviews and group interview conducted by the author were based on a semi-structured protocol of five questions and four sub-questions (see Appendix A for the individual interview protocol and Appendix B for the group interview protocol). The interviews by the student colleague were conducted in the context of an internship with the main aim of surveying conservatory employees’ impression of a University of Groningen study they had recently participated in. While these interviews were informed by this protocol, they were not explicitly based on it; in a given interview, some questions partly or wholly unrelated to the current study were asked, and only parts of the protocol questions were asked.

The semi-structured protocol contained the following questions:

1. What is the goal of music assessment?
2. How is valid assessment achieved in music performance?
    a. What approach (holistic and/or mechanical) to assessment results in more valid assessments of music performance? Why?
3. How is reliable (consistent) assessment achieved in music performance?

    a. What holistic and/or mechanical approach to assessment results in more reliable (consistent) assessments of music performance? Why?

4. How is fair assessment achieved in music performance?
    a. What holistic and/or mechanical approach to assessment results in more fair assessments of music performance? Why?
5. What obstacles do you see in realizing valid and fair assessment of music performance?
    a. How does this relate to holistic or mechanical approaches?

Analysis

First and foremost, Braun & Clarke (2006) argue that a key component of conducting qualitative analyses is being wholly transparent about one's analytical choices. As such, the subsequent descriptions of the analysis steps will be interspersed with descriptions of how they were implemented in this study. Before the steps themselves, the authors raise a tentative warning that deeply familiarizing oneself with literature related to the topic before conducting the analysis may narrow one's viewpoint; the idea is that one may become too focused on some predetermined aspects of interest before the present data is taken into account, and as such may be less likely to notice other aspects of importance outside of this scope (Braun & Clarke, 2006). However, a contrary position comes from Tuckett (2005), in which he argues that an early literature review may in fact make one more inclined to notice the finer details in the data. For this study, an initial review of relevant literature prior to conducting the analysis was necessary as prior knowledge was deemed too limited, and due to study planning necessities (e.g., submitting proposal and drafting interview questions). That said, Braun and Clarke's (2006) warning was heeded: while a single hypothesis has been presented so far, an open mind was kept for any and all themes to emerge.

Phase 1: Familiarizing Oneself with the Data

The first step of thematic qualitative analysis begins with reading (in this case, listening) through one's material. In doing so for this study, emerging ideas were pondered to inform the analysis to come, in line with Braun and Clarke's (2006) recommendation. Examples of this included preliminary notions of themes and how they could potentially relate to the literature presented in the introduction; here, two of the main themes and three of the sub-themes emerged (1, 1b, 3, 3a, and 3c; see Results)[1]. Following this, the interviews were manually transcribed, a process that is thought to contribute to the quality of analysis: it is an effective approach to familiarizing oneself with the material (Riessman, 1993) and can be seen an early act of interpretation where creation of meaning starts (Bird, 2005). Indeed, listening to the recordings after the fact generated new interpretative perspectives; for example, since the cognitive effort of attentively conducting the interviews was no longer present, finer details in the participants' wordings and tones appeared more salient.

Further, Braun & Clarke (2006) recommends that the transcription process is guided by the values of retaining the information needed, remaining 'true' to its original nature and being practically suited for conducting the analysis (Edwards, 1993). The author's interpretation of this was to transcribe close to the entirety of the recordings rather than the parts deemed 'most relevant' alone, only omitting those parts at the very beginning and end of the interviews that did not relate to the subject at hand, but rather consisted of social formalities. The audio was, however, not transcribed verbatim: for clarity, single utterances of filler words such as 'like' and 'you know' were kept, but repeated utterances of the same filler words (or filler noises such as 'uh') within the span of a few sentences were omitted. Note

[1] The above numbering refers to (1) 'Necessity of Intuition and Consensual Assessment' with its sub-theme (1b) 'Trusting Expertise', and (3) 'Obstacles to Mechanical Judgment' with its sub-themes (3a) 'Holism' and (3c) 'Restrictiveness'.

that when expressions such as ‘you know’ were used in the literal and more functional sense, ‘you know’ gauging consensus of some previous point, they were not omitted.

Phase 2: Generating Initial Codes

The second step involves generating codes, that is “the most basic segment, or element, of the raw data or information that can be assessed in a meaningful way regarding the phenomenon” (Boyatzis, 1998, p. 63). In this way, one organizes bits and pieces of the data into groups that may later turn into themes (Tuckett, 2005; see Phase 3). For this study, the coding was part theory-driven and part data-driven. Theory-driven refers to the data related to the research questions being explicitly sought after and coded relating to these questions, and data-driven refers to this study also being exploratory in regards to themes not directly related to the research questions. This step was characterized by a cyclical nature: as new codes were defined in interview B, interview A was re-read and re-coded to accommodate for these, with some segments having their previous codes changed. The cyclical coding continued throughout: as new codes were defined in interview C, interviews A and B were re-read and re-coded, and so on.

At the end of this step, it was deemed that an excess of codes had been created, numbering close to 100. Since this presented cognitive difficulties in terms of project overview and consistent application of codes, codes were either merged or discarded until roughly 40 codes remained. The criteria for discarding were that the code either occurred too seldomly, was too specific compared to the other codes, or judged not relevant to assessment. An example of a code that was discarded was “Dislike” of mechanical judgment with only six mentions; what was judged relevant for the study was not the mere presence of a dislike, but the *reasons* for it.

Phase 3: Searching for Themes

The third step comprises analyzing and sorting the previously generated codes into different potential main themes and the sub-themes within them. Some codes may fit into a single theme, some in more than one theme, and some not into any specific theme. The analysis ultimately generated codes in all three of these categories, leading to several initial main themes and sub-themes, and some additional discarding of codes by the criteria described in the previous step; at this point, 33 codes remained. The matter of deciding how broad a theme could be before it had to be broken down into sub-themes proved difficult. To balance detail with accessibility for both author and reader, it was decided that any main theme had to consists of at least one and at most four sub-themes.

Phase 4: Reviewing Themes

For the fourth step, Braun & Clarke (2006) identifies two phases of reviewing themes, both involving re-reading the entire data set. First, one reviews the individual codes contained in each potential theme and asks oneself if they are consistent with the theme and with the other codes in the same theme; in short, one determines whether they fit well. If yes, one continues with the second step described below, and if no, one would either revise the theme, create a new theme in which to gather these codes, or discard the codes. In this analysis, four main themes and their sub-themes were revised into main themes (2) ‘Appropriate Degree of Structure’ and (4) ‘Personalized Assessment’, and their sub-themes, respectively (see Results).

Second, one moves up a level by reviewing the individual themes to see if they similarly fit with the data set as a whole, then checks the individual codes for any additional data that may have been initially overlooked and (re)codes it. Here, it is possible that yet more themes emerge, although this was not the case in the present analysis. A contributing reason to this may be that in hindsight it was realized that step one and two were implicitly carried out concurrently rather than as two sequential steps. Concluding this stage, and in line with Braun & Clarke’s (2006) recommendation, refinement was terminated when it was perceived that value was no longer added; here, this point was reached when a fourth and final re-reading did not result in any changes to the theme structure.

Phase 5: Defining and Naming Themes

In step five, one defines what the core of each theme is, and the themes as a group. The essential part here is not just the ‘what’ of one’s data, but the ‘why’: why are the chosen data points relevant for this theme? Why are they supportive of the theme? Why are they part of this theme, and not another? It is important that each theme receives an in-depth analysis where these questions are answered, where the research questions are related to, and where the main theme (and its sub-themes) is put in perspective to what story the data set tells as whole. One also needs to be clear about what a theme is not, as this assists in determining if one needs to refine a theme further to be more specific and demarcated from other themes. In this analysis, the ‘why’ questions above were approached by writing a pre-draft of the results section, intermittently comparing the wording of the emerging draft with the main themes and sub-themes to see if they could be accurately described. The description of main theme (4) “Personalized Assessment” (see Results) was deemed ambiguous, and the coded material was revised and re-arranged again. A particular difficulty of arranging the data into main themes and sub-themes was appropriate demarcation; in a sense, the sub-themes of (1b) ‘Trusting Expertise’ and (3a) ‘Holism’ seemed present to some extent in all of the main themes, but had to be placed within one or the other as they were not deemed to warrant main themes of their own. Although these sub-themes eventually found their way into a main theme, a truly satisfactory resolution was not achieved as they were still perceived to reasonably fit into more than one main theme.

Results

Several themes emerged from the analysis, containing both positive and negative views on both structured versus unstructured information collection and holistic versus mechanical judgment. Overall, based on mention frequency both within and between interviews, there were more positive than negative expressions towards holistic judgment. For mechanical judgment, the pattern was reversed with a majority of negative expressions. Interestingly, while holistic judgment was clearly preferred, all participants expressed at least some positive views towards mechanical judgment. Further, views on structure in information collection were skewed towards the positive, although all participants expressed opinions that both supported and opposed explicit structure in some manner. This section will describe the participants' views grouped in a total of four different themes and their nine subthemes. The major themes were labeled as follows: (1) Necessity of Intuition and Consensual Assessment, (2) Appropriate Degree of Structure, (3) Obstacles to Mechanical Judgment, and (4) Personalized Assessment.

Necessity of Intuition and Consensual Assessment

A central theme that was expressed by all participants to varying degrees was the notion that there are certain ethereal qualities of music that are implicitly perceivable to those with experience and expertise, but which cannot be easily explained, therefore necessitating the use of consensual assessment[2]. Precisely what those qualities constituted remained somewhat ambiguous, but they could be interpreted as being related to emotional impact, artistic intention, or an 'X-factor':

[2] One clear deviation from the Consensual Assessment Technique procedure as described by Baer & McKool (2009) is that after completing the individual assessments, the assessors of Prins Claus Conservatory discuss the outcomes with each other which could constitute "justification", something the above referenced work discourages. In the conservatory, rare cases of a point difference of more than one and a half between the lowest and highest individual assessment triggers a re-vote that may or may not result in an altered final assessment.

> “Because also what is not so clear with the assessment criteria, and which I find very important, I think everybody, it’s . . . the ability to perform, to make you feel something. And not everybody has this, but if you have it or not have it, it’s like day and night. It’s also very difficult to grade, because it might be that they can do that. So, you were like, you got really emotional while they were playing. It was inspiring. I mean, that’s everything that you want. But at the same time, maybe his technique was not the best, his presentation, his sound. So then if you look only at the assessment criteria, actually, this is not a very good performer. But actually, it was the best performance of the day.” (J_i)

Within this theme were two sub-themes that touch upon different components relating to the role of intuition in music assessment, labeled as concerning (1a) Difficulty of Defining, and the consequent need for (1b) Trusting Expertise.

Difficulty of Defining

Four participants expressed that while they may be highly aware of the relevant criteria, both personal ones and those specified by the conservatory, many assessors find it very difficult to define and verbalize them. Importantly, the participants who held this view did not consider this difficulty to be a deficit, but simply a trait that varies between assessors, seen as having little to no impact on their ability to validly assess. One participant said the following:

> “There are high quality musicians that cannot intellectually describe what it is that they’re doing. They know exactly what they’re doing. But they just, they’ve never really thought about it in intellectual terms to be able to describe it. But in this case, what you need is someone who can clearly understand what it is that makes a musician

> valid, or performance valid . . . I mean, it’s like I say, it’s all very, very holistic and very messy. At the same time, that doesn’t make it any less valid. You know, it’s just harder to prove it to people that don’t hear it. And that’s what it comes down to. You have to be able to hear it.” (E_i)

Trusting Expertise

As the first quote suggested, it was believed that some qualities of music and musicians are ‘special’ in the sense that they are potentially able to shape the impression of the whole to a point where other, more tangible aspects are considered less important. Five participants expressed that to recognize and accurately take these qualities into account, one needs to rely on the experts of the field, and the associated institution’s ability to select experts with this competence:

> “If you’re very experienced in music and have dealt with it all your life, the more experience you have, and the more knowledge you have, the more you are equipped to judge . . . Sometimes you have to trust the experience and knowledge and expertise of the jury . . . some people don’t like it because it could be dangerous to [exchange] trust in people [for] rules that you can write down on paper. I think that relying on the professionalism of artists, from the working fields, that’s the most important thing . . . To trust that expertise of individual people who are working as professionals in the field.” (G_i)

> “Certainly, taste plays a role, but it’s not a question of, oh, this person plays the way I like, so that’s great. This person plays well, but not the way I like, so it’s not good. That’s why I say you need somebody, you need teachers with a certain kind of vision,

> certain kind of view of the whole picture . . . Like I say, not every person could or would be a good assessor. But it’s all people work . . . Teachers can choose faculty that are at a high level, that has a vision enough to understand what people are doing, and if that’s at a high level, and if those people can really contribute . . . It’s messy. But that’s the only way I would know how to do it.” (E_i)

Concerning the reliability of consensual assessment, it was acknowledged by four participants that differences in taste, knowledge of certain instruments or personal emphasis on certain aspects results in some subjective biases in assessment. However, these participants expressed that this effect is adequately mitigated by the averaging of the jury members’ individual assessments:[3]

> “And also this thing, again, that it’s subjective, and that you can very much hear once people start to give feedback, what is important to them? And I think that is, of course, nice, that everybody thinks different. That’s why if there will be much more jury members, it would be even more different things, and then it will be more of an *average of reality*.” (J_i, emphasis added)

> “Yeah, it’s the fact that there’s multiple jury members, that’s the security. That gives us security because maybe one person could make a mistake, but then there’s three or four more there, and then it becomes clear that the average will bring a good outcome.” (G_i)

[3] The current jury structure for final examinations at Prins Claus Conservatory comprises five people: the chairman, two teachers who play the same instrument, one teacher who plays another instrument, and one external member who is not affiliated with the institution.

One participant told of “calibrating sessions”, where assessors informally assess a recording of an older performance without knowing the previously reached outcome beforehand. When comparing the previous outcome to their own final assessment, they were most often found to agree:

> “I think again, because we have so many assessors, day to day disbalance or something can be smoothed out by that. But also by calibrating sessions . . . We had an accreditation and we looked at our files and the dossiers of the students and checking material, what was in the file, and then also there was a recording of the final [exam]. And I think we were with four or five people. Just for the fun of it, we looked at bits and pieces of the concert the students had on file . . . And then we just said all, I think this was a seven and a half, and the other one, no, this was a nine, or this is definitely an eight… And then we looked into the file where all the transcripts were and then the feedback and of course, the grade. And I think 90% of the time we were spot on. And one or two times we were like half a point off. And we were a group of people who weren’t at those exams, and pretty much came to the same conclusion as the jury on that evening. So that’s when we thought, oh, okay, I think we are on a good track here, on the right track with this kind of calibration, or letting different groups of assessors judge the same exam on a different date on a different time and see if they come with the same conclusion. I think this helps a lot.” (B_i)

Appropriate Degree of Structure

Six participants expressed positive views on using explicit criteria, including specifying subcomponents. In fact, at time of writing, Prins Claus Conservatory already prescribes eight separate criteria as a basis for their assessment; however, when asked, none of these six participants could name all of them. Perhaps as a symptom of this, they expressed that there should be more clarity regarding the criteria in use:

> "I had a little chat with my colleague from the exam committee and I told him about these criteria, and a lot of colleagues are not aware of it. It might be wise to give them some input before all these exams start, in front of it instead of during the exam. Maybe in two weeks, there's a period of exams, and then two weeks before that you can inform all colleagues, okay, you have to notice these criteria." (A_i)

> "I would say before an exam period, together with all the examiners, take a look at the criteria, because there's always a list of criteria being handed to the examiners, usually too late. At the backside of the of the form you need to fill in your grade, but if you would have that before an exam period and if you would have a discussion about this then maybe in the holistic way of looking at it, when you get a little bit more mechanical… because you're more aware of all the different aspects and you have already talked to your fellow examiners about what is important." (F_i)[4]

The above views notwithstanding, two participants expressed a prevalence of dislike for relying upon explicit criteria as part of the assessment procedure as it was deemed unnecessary. One participant described it in the following words:

> "What I noticed with lots of teachers is that they don't want this list with criteria . . . It's not about music, it's not how you should assess music . . . The thing is we work like musicians, and musicians always think they know the criteria in their head. We

[4] Note that this participant appears to confuse structured information collection with mechanical judgment, likely as a consequence of inadequate explanation on part of the researcher (see Discussion).

> know how to grade because we have experience in the field and in school. So hardly anyone will have a look at the list of criteria." (A_g)

This theme contained a sub-theme that related to the level of explicit structure found in rubrics, and was therefore labeled as concerning (2a) Clarity and Transparency of Rubrics.

Clarity and Transparency of Rubrics

While the term 'rubrics' was not explicitly used in the interviews, the concept was discussed through its partial aspect of dividing performance into subcomponents. Five participants found this helpful in assessment, especially in the sense of maintaining attention towards the entirety of subcomponents:

> "It helps us to stay focused, because when you only listen and you have nothing on paper, then you might miss something. So yeah, definitely. It definitely goes hand in hand [with music assessment]. It's very good to do it." (G_i)

> "I think that if all the examiners set the same goal, have seen something about all the different criteria on that list that is handed out to us . . . We are more aware of all those different criteria everyone is looking at, listening to the same things. For instance, if you have five criteria, and as an examiner, I think two criteria are the main criteria, and I don't look at the other three, then I might give it a certain grade, because I only look at those two criteria. But if you convince everybody that they should look at all five criteria, then there is automatically more [consideration] of these five criteria." (F_i)

Interestingly, participant A, who expressed negative sentiment towards explicit criteria in the group interview, also had something quite positive to say about them in the individual interview:

> "I think it's helpful. And because I'm doing a lot of writing with the exams, it gives us tools to get it on paper. And there's also always a third party that has to read these protocols. In that case, it's really helpful to use them. To get the story clear on paper, a transparent story." (A_i)

Obstacles to Mechanical Judgment

When it came to explicitly and numerically rating subcomponents, all participants, regardless of their positive or negative views on structured information collection, expressed mostly negative sentiments towards it at some point. One described this quantification as neglecting the emotional aspect of music while attempting to break down a whole into parts that no longer represent the whole:

> "If you quantify everything, if you try to measure everything, you take out a lot of the emotions, but you also make it very fragmented. And you're trying to get hold or grasp of something, which in our opinion as artists, maybe isn't so easily caught in numbers. It might work with quantitative research, and I don't know, in law or in physics or something. But in music, I doubt it." (B_g).

> "Because sometimes, even if the intonation is terrible, it still can sound great. Even the best world-famous musicians can have a bad technique. They're like, world class musicians, and they touch everybody around the world, and everybody wants to buy

> the record, loves them. And then in our school, they would fail. That's weird, you know? Yeah, it's just with arts, it's not like science. It's different." (G_i)

Another point regarding rating subcomponents made by four participants was that in the moment of assessment it detracts from the music performance. When describing captivating performances, teachers mentioned that they occasionally forgot that they were there in the role of judges, and this was seen as a positive effect that spoke to the high quality of the creative product:

> "Before the concert, of course, it's all written on the thing. But sometimes in a concert, I'm listening, and I'm kind of getting into the performance part and then sometimes we might forget that we're there to be a judge. But of course, we are assessors there to make appropriate assessment . . . This is a whole discussion, the difficulty of listening to a piece of music and also thinking of things, thinking of comments or thinking of these things that you're going to tell the student later. And then the next tune happens, the next song, and then you kind of forget what you're going to say in the first tune. I think that sometimes that might happen. So I don't know if it's something that we would be able to do, making notes while the piece is playing, during the whole concert." (C_g)

> "I also found it interesting what C said, that sometimes you forget, you're being an assessor in a concert setting . . . So there's somebody performing on stage, there's audience, lights on, there's music and you sit, and if you forget that you're an assessor at those times… Usually that's a great concert. That's what we're looking for." (B_g)

In addition, this theme contained three sub-themes relating to either a notion of music assessment being intrinsically holistic, or to different types of impracticalities. They were labeled as concerning (3a) Holism, (3b) Unfeasibility, and (3c) Restrictiveness. Additionally, a point about ‘cheating the system’ of mechanical combination was raised by two participants, resulting in the minor section of (3d), Bypassing the algorithm.

Holism

A critique of using mechanical judgment to arrive at a grade through individually rating subcomponents was a view that the parts of a whole interact, making the sum of the whole more than its parts, an idea is known as metaphysical holism (Kitchener, 1982). This perspective was discussed in five of the individual interviews as well as the group interview, and it was expressed with slight variations: either that the parts are inseparable from the whole or that even if separate subcomponents were to be rated there would be essential parts that could not be specified:

> “. . . you can put a number on how in tune someone plays or a number on how expressive they played. But you can’t really. I mean, it’s always been holistic . . . you can’t have one without the other, it has to be the whole thing, or it won’t work.” (E_i)

> “Because it’s about what we talked about before, music is emotion. And it’s lots more than only things that you can specify. And when you only assess things that you can specify, you’re off the hook or… not the way to judge in my opinion.” (A_g)

Unfeasibility

Four participants reported that employing mechanical combination would be unfeasible. A main reason for this was the notion that the different subcomponents are reliant upon each other and thus become non-representative of the performance when assessed in isolation, relating to the previously discussed notion about an intrinsically holistic nature of music:

> “. . . jazz music has always been an oral tradition, which means while you can look at different aspects of it individually, mechanically, as you say, the judgment is always holistic. You can’t have intonation without good timing, you can’t have phrasing without intonation… I mean, you can say, oh, the intonation needs to be better. But there are some players that are very expressive, and their intonation isn’t great. And some players whose intonation is perfect, but it doesn’t interest you. So for me, after thinking about it quite a bit, because this is, yeah, this is a hot topic, I’ve really come down to the conclusion that you have to look at it holistically. Music is the definition of holistic, [integrative] assessing. And that mechanical assessing of weighted criteria? Yeah, the timing is twice as important as the intonation, half as important… No idea. For me, that’s onbegonnen werk [a hopeless task], as they say in Dutch . . . It’s really messy. But again, I just can’t see how it could be done another way, and I think it’s always been done that way. (E_i)

Restrictiveness

Regardless of what subcomponents could be specified and rated, it was argued by four participants that an aggregation of ratings could never accurately represent a performance; holistic judgment was deemed a necessity to capture what mechanical judgment could not:

> “You can use the same criteria for everyone, but you can’t give the same value on each one. A very simple example, I think Michael Jackson could probably not read

> notes . . . So if you say like, well, reading notes is 20% of your grade, that means Michael Jackson can never get more than 80%. But that doesn't make sense, because he's a great artist, he's one of the best artists ever in the history of music. And we can value him. Maybe later on, we will say like, well, Michael Jackson and Beethoven are equally genius. Beethoven could really read notes, you know? But he was not a great performer, a show man or something. So could Beethoven not graduate in our school? Because he doesn't know how to interact with the audience? That doesn't make sense. He should definitely graduate with Cum Laude if Beethoven was our student." (G_i)

Bypassing the Algorithm

While at most a peripheral point based on its few mentions, two participants brought up their personal and anecdotal experiences about mandated use of mechanical combination outside of the conservatory context, such as in music competitions. In these situations, use of a predetermined mechanical combination with its assigned weights was bypassed by first holistically reaching a final assessment, then adjusting the scores in the formula to match this, rendering the mechanical component redundant:

> "Over the years, we've done also separate grades, then still, as a jury member, I make the grades separate so that the outcome will be anyway what I would give without. So the answer to your question is yes, I like it better as a reminder than to grade them separately." (G_i)

> "Look, I've been in juries of competitions often. And I used to have this competition where you also have criteria and this and this and then there was balanced by taking two times this grade, one time the other and then divided by two. So what happens is,

> you refer to this and then you get a grade, and then you look at all the people and then you're like, well, but this person has higher than the other one, although I don't think that that was true. So, I notice that at some point, I was like, okay, I want to have that grade so I make a calculation in my head what to give for this. Things to give to get to that grade. So that felt a bit unnatural." (I_i)

Personalized Assessment

Perhaps by being a rather small institution with just over 300 students, there is time and resources for a more individualized form of assessment at the conservatory. In line with this, seven participants described some level of criteria adaptation they regularly make depending on which performer they are assessing, and on what is perceived important specifically for that performer's development. Essentially, some components are judged as more central for one performer, while different ones are judged as more central for another; in other words, the implicit weights are used in a flexible manner. The objection consists of the notion that if a mechanical combination policy would be mandated, it would not allow for this flexibility and consequently would take away a freedom of judgment that is regarded as important:

> "We should use all the criteria, I don't know if it's eight or nine or seven. But they are not equally important. They're not all one eighth of the total outcome. And it can also be different for each exam, some criteria can be more important than others, it depends on the situation. So I think that we should keep that freedom for the judges." (G_i)

Branching out, this theme includes two sub-themes each relating to a different aspect of personalization: (4a) Flexible Criteria, and (4b) Norm Divergence.

Flexible Criteria

Five participants expressed that being able to adapt the importance of subcomponents was necessary to accomodate the wide variation of assessees, whether that regarded their instrument, style, their intentions with the music piece or with their future music careers. Some had this to say:

> “For myself, it’s really important to know what a student wants in the future . . . Some students just want to play in an orchestra, for instance, technical, really good trumpet player, and you have to be able to play together and with a certain sound and ability to listen to the people around you. But when your future idea is being a soloist in for instance a jazz ensemble, then these criteria are kind of different, in my opinion . . . in this case, you have to write a personal list of criteria, because maybe it’s up to the student to write down what do I want to do, and what will my learning outcomes be at the end of the study? And maybe the teacher and the students can make lists of criteria so that you can assess these criteria, but it’s only for this student. Well, that’s really different input than we do now. It’s only an idea.” (A_i)

> “There have been performances of very, very old musicians, who were hardly able to play, would you give a 3 or 4 in 1 to 10? But because of their intention, you sense what the musical meaning is. Performance is worth, like, an 11. But that’s the strange thing with art. Also, don’t forget the instruments, some instruments sound great. Other instruments don’t sound too great. So you can see someone’s taking everything out of the instrument, but in the end, sound is not great. Someone else uses a Salivarius or some better instrument, but the arrangement doesn’t sound so good. What can you do?

> What was the grade? That’s the complexity of that. I will not give such a high grade, because she is not using the instruments to the max.” (I_i)

> “. . . for example, if it’s innovative, if it’s something that’s never done before, should that be the requirement? And if so, if you’re not innovative, can you not graduate with high grades? I think that you cannot say that because it could be it could make the grade really high if you’re innovative. But you could also have a really high, equally high grade, if you’re not innovative, nothing new but you do it fantastic with a lot of emotion and beauty. So that’s why I say in some points, some cases and some criteria have more value than in other cases in other exams. And it’s completely based on the judgment of people. And usually, university researchers don’t like that. But it’s not people that you bring from the street and you ask them, oh, do you like this performance? No, it’s people who have been in the industry for years and years and know what they’re talking about.” (G_i)

Norm Divergence

One aspect of music performance previously mentioned in the introduction is norm divergence. In music, as in creative performance in general, norm divergence is often considered positive and consequently rewarded (Wesolowski, 2012). In speaking about mechanical combination, three participants expressed that it would not be able to validly accommodate for innovation and out-of-the-box performances.

> “I hope the one who is assessing you and your paper is open minded to see that it is something you cannot put in a box. It’s not one way that fits all types of artists.” (G_i)

> “And the way it is now, how the past 20 years actually went on with learning outcomes, the competencies, focus, curriculum and education, reforms of the past seven, eight or ten years... makes it much less fun to listen or to make music in the school system than it used to be . . . If you compare it to what’s going on now with everybody in the books and competence this, competence that, and also teachers being very, yeah… You get restrictions all along, because then there comes a student who’s really creative, but it doesn’t fit to the criteria. What do you do with it? And then how do you adjust that? And when somebody, especially artists, they tend to think out of the box. And if you make the box too narrow, no artists will fit in there. I don't know. That’s the challenge. To have a balance in these two things, maybe only a couple of criteria, then a holistic approach works better.” (B_g)

Discussion

This study sought to investigate music assessors’ attitudes towards and concerns about different assessment methods, specifically structured versus unstructured information collection and holistic versus mechanical judgment. In doing so, it aimed to clarify a set of challenges of improving music assessment as to increase validity and reliability, taking into account the gap between approaches often suggested in academic research and approaches currently used in music assessment.

In general, assessors at the Prins Claus Conservatory deemed it necessary to at least partially rely on an intuitive approach through consensual assessment. This was based on the notion that there are some special components of music that are not easy, or perhaps not possible, to explicitly define. To perceive and subsequently be able to able validly and reliably assess these special components, one had to rely on intuition and expertise. This aligns well with practitioners exhibiting a strong preference for relying on expertise (Kuncel et al., 2013) and the high prevalence of subjectivity in music assessment (Wesolowski, 2012) as well as in other fields such as employee selection, in which there exists an adamant reliance on intuition and the idea that people can become highly proficient in intuitive assessment through experience (Highhouse, 2008). Participants claimed the expertise of music assessors was perhaps not clearly demonstrable to an ‘outsider’, but clearly so from one expert to another, which was the basis of a good validity in their approach. The averaging of individual assessments over multiple judges was in turn proposed to be the basis for a good reliability.

All participants expressed at least some openness towards, and occasionally even desire for, structured information collection. Formally, the conservatory already employs subcomponents as basis for grading, but as was described these subcomponents were not always taken into consideration by assessors. Since adhering to a structured information collection method, e.g., employing rubrics, tends to result in higher reliability relative to unstructured methods (Arkes et al., 2006; Conway et al., 1995), an important question is how to best facilitate greater adherence to structure by practitioners. One idea suggested by some participants was that discussing the criteria and how they are assessed more frequently could lead to both a greater understanding and awareness of them, and a greater appreciation for the purpose of employing this approach.

The topic of mechanical judgment was characterized by stronger, more negative views. The general consensus was that it was not a suitable approach for assessing music as it would be unfeasible to implement, restrictive of assessors’ freedom in judgment, and ultimately unable to validly capture the ‘whole’ of a performance by only looking at the parts. These results could be linked to previously discussed belief that a holistic approach can account for ‘exceptions to the rule’ that would otherwise be inaccurately assessed with mechanical methods alone (Dietvorst et al., 2018; Guay & Parent, 2018; Hoffman et al., 2017). Further, as some participants spoke of holistic preference in terms of ‘freedom’, the negative views may be connected with the perceived threat to assessor autonomy that mechanical judgment may incite (Neumann, Niessen, Tendeiro, & Meijer, 2021). It is also worth reiterating Dawes' (1979) discussion of the claim that numbers could never accurately represent the whole of a person, a view that was clearly expressed by participants in this study. While the claim does appear reasonable, one must remember that the same applies to holistic judgment; assessment is generally an approximation, and the question of which method to use does not necessarily address which one is 'truly objective', but which one comes closest to this ideal.

Lastly, most of the participants expressed that they assess the criteria with some level of flexibility, i.e., assigning different importance to criteria based on which performer is being assessed. The rationale for this was that each individual has their own priorities and intentions for their future music career, and so some subcomponents are deemed less important for one context and more important for another, e.g., rhythm would be more important when aiming for an orchestra position compared to one as a soloist. While this could be an example of the generally found lack of consistency when applying weights intuitively (Dalal et al., 2020; Dawes, 1979; Kuncel et al., 2013), one should consider the question of whether this could potentially be a positive in this specific context. The flexibility is intentional, clearly communicated to the student, and ultimately aims to serve the student by considering goals and intentions from their personal perspective. Also, this desired flexibility may be part of the equation of satisfying autonomy in practitioners, as discussed by Neumann, Niessen, Tendeiro, and Meijer (2021). Of course, the idealized intention alone does not guarantee any improvement in validity or reliability, especially as flexible criteria weights directly contradict a core principle of mechanical judgment. The question is nevertheless worth considering as some flexibility could possibly allow for better accomodating those performances that fall 'out-of-the-box', creative and skillful in their own manner. Perhaps one could approach music assessment with set weights for certain criteria deemed fundamental, while retaining some personal flexibility for others (although what is deemed fundamental in this context would, again, be difficult to determine without subjectivity).

Altogether, this study suggested that music assessors' attitudes on the discussed approaches to assessment differed to varying extents, but were ultimately pervaded by a desire for some degree of assessor discretion in evaluating music performance. For most participants, this desire could seemingly be fulfilled within the bounds of structured information collection, but the same could not be said of mechanical judgment.

Limitations

A first clear limitation of this study is the small sample size. The participant recruitment process had a significantly lower response rate than expected, possibly due to some level of fatigue in the participant pool given that they had already been prompted to participate in two related studies within the last few months. In addition, all participants were sampled from a single institution, which results in limited generalizability. What could be said at most is that since the participants were reportedly well-connected within the national music education community, the results suggest that similar attitudes are somewhat likely to also be found elsewhere within the Netherlands.

Another limitation concerns the interviews being conducted by two different researchers, and for separate purposes. As a result, the researcher becomes a possibly confounding variable in the sense that one could have elicited responses different to the other. In particular, the author of this study conducted interviews based on a set of semi-structured questions (see Appendix A and B), while the student colleague did not. Still, the entirety of pre-written questions was not asked in all of the author's interviews; it proved difficult to impose this structure without obstructing a natural flow of responses, especially given the author's limited experience in interviewing. Additionally, the conversational dynamics clearly varied in that the author's interviews elicited longer answers between questions, while the student colleague's interviews contained a greater quantity of back-and-forth communication. This created a difference in the process of coding for themes: as viewpoints were explicitly repeated more often in the student colleague's interviews, they ended up containing a greater quantity of individual codes relative to those of the author. Worth considering is also the possibility that the group interview format elicited different viewpoints than the individual interviews given the greater social influence present.

In connection to the previous limitation, the interviews were single-handedly coded by the author. It is possible, perhaps even likely, that another coder would have defined different codes, coded the individual segments differently, and arranged the segments into different themes. Given this, the conclusions that would have been reached by another coder could have been markedly different from those reached by the author. Further, had the interviews been coded by multiple individuals, it would have been possible to cross-check interpretations and additionally calculate and report a measure of inter-rater reliability, both of which were impossible in this study.

Additionally, this study was initially planned with a focus on specifically holistic versus mechanical judgment, as can be seen in the interview protocol questions mainly addressing these approaches (see Appendix A and B). Therefore, this study's explicit inclusion of structured versus unstructured information collection was a result that followed interviews naturally leading to this topic. While structured versus unstructured information collection was briefly touched upon when defining terminology for the participants, the explanation was implicitly intertwined with mechanical versus holistic judgment, rather than demarcated as a topic of its own. Thus, it is likely that the difference between information collection and combination was not entirely clear for some participants, potentially leading to imprecise interpretations during analysis.

Lastly, the interviews did not adequately answer all research questions. In particular, ‘future outlooks’ of music assessors were rarely elaborated on in the interviews, leading to insufficient material for generating relevant codes, or to form a theme specifically addressing this question. While conducting fully structured interviews may have made it easier to elicit answers clearly related to this research question, semi-structured interviewing was chosen as to allow greater freedom for participants to speak about what mattered most to them.

Future Research

Given the gap between music assessment methods suggested by academic research and those currently employed in practice, especially considering the prevalent tension between practitioners and policymakers (Denis, 2017), it is important to better understand music assessors’ viewpoints so that more common ground on assessment methods of choice can be established. Since this study only contributed viewpoints from a very small sample, future research could investigate practitioner’s attitudes on a far larger scale, both geographically and between different educational levels such as primary, secondary, and upper secondary education. Importantly, these studies should utilize trained interviewers for better consistency in questions between interviews, and should be analyzed by multiple, trained coders, allowing for quantitative measures of inter-rater reliability and cross-checking of interpretations. Further, a greater emphasis could be placed on comprehensively explaining and demarcating the differences between structured versus unstructured information collection and holistic versus mechanical judgment. In this way, participants would likely be able to provide clearer answers and researchers could rely less on interpretation as to the intended meaning of a given statement.

While only being a minor point in this study, the reported phenomenon of assessors ‘bypassing the algorithm’ in situations of mandated mechanical combination could be a promising avenue for future research. To the author’s best knowledge, no study has explicitly investigated the prevalence of this phenomenon; if it would show to be present in a significant proportion of mechanical assessment situations, whether relating to music performance in particular or creativity assessment in general, it would have the potential to seriously confound studies on the subject as the method used would only appear to be mechanical but in essence still be holistic.

Conclusion

While perhaps not all of the benefits of mechanical judgment can be implemented into a method acceptable to practitioners in music assessment, participant's openness to structured forms of information collection could be a promising entry point for introducing empirically supported methods that result in higher validity and reliability. Some standardization and consistency is better than none, and regardless of what the future of music assessment holds, the author advises that approaches to assessment developed in academic settings take into account the views and needs of practitioners to reach an acceptable compromise that draws on the best of both contexts.

Acknowledgements

I would like to express my deep gratitude to my supervisor, dr. Susan Niessen. The path towards finishing this project certainly came with far more bumps and difficulties than expected, and your support stayed exceptional throughout. Thank you for your time, your extensive feedback, your understanding for my situation, and your belief in my ability.

I would also like to thank my fellow student and colleague, Carla Finsterbusch. Your data was invaluable for this thesis, and collaborating with you was a joy.

Finally, I would like to thank my other student colleague, Maria León. Your insights and suggestions were helpful and refreshed my perspective, and this thesis surely improved through our interactions.

References

Amabile, T. M. (1982). Social psychology of creativity: a consensual assessment technique. *Journal of Personality and Social Psychology, 43,* 997–1013. https://doi.org/10.1037//0022-3514.43.5.997

Arkes, H. R., Shaffer, V. A., & Dawes, R. M. (2006). Comparing holistic and disaggregated ratings in the evaluation of scientific presentations. *Journal of Behavioral Decision Making, 19,* 429–439. https://doi.org/10.1002/bdm.503

Baer, J. (1994). Performance assessments of creativity: Do they have long-term stability? *Roeper Review, 7,* 7–11. https://doi.org/10.1080/02783199409553609

Baer, J. & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In: C. S. Schreiner (Ed.), *Handbook of Research on Assessment Technologies, Methods, and Applications in Higher Education* (pp. 65–77). Information Science Reference. https://doi.org/10.4018/978-1-60566-667-9.ch004

Baptiste, L. (2008). Managing subjectivity in arts assessment. In L. Quamina-Aiyejina (Ed.), *Reconceptualising the Agenda for Education in the Caribbean: Proceedings of the 2007 Biennial Cross-Campus Conference in Education, 23–26 April, 2007, School of Education, UWI, St. Augustine, Trinidad and Tobago* (pp. 503–509). School of Education, UWI.

Bergkamp, T. L. G., Niessen, A. S. M., Hartigh, den, R. J. R., Meijer, R. R., & Frencken, W. G. P. (2020). (On)terecht buitenspel gezet. Sportprestaties voorspellen door systematische en gestructureerde beoordelingen. *SportGericht, 74,* 36–40.

Bird, C. M. (2005). How I stopped dreading and learned to love transcription. *Qualitative Inquiry, 11,* 226–248. https://doi.org/10.1177/1077800404273413

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage.

Boyle, J. D., & Radocy, R. E. (1987). *Measurement and evaluation of musical experiences*. Schirmer Books.

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, *3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M., & Nitko, A. J. (2014). *Educational assessment of students*. Pearson.

Brown, G. G., Spicer, K. B., Robertson, W. M., Baird, A. D., & Malik, G. (1989). Neuropsychological signs of lateralized arteriovenous malformations: Comparison with ischemic stroke. *Clinical Neuropsychologist*, *3*, 340–352. https://doi.org/10.1080/13854048908401483

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195–217). Cambridge University Press.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage.

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. https://doi.org/10.1177%2f0022429409333405

Colwell, R. (2008). Music assessment in an increasingly politicized, accountability-driven educational environment. In T. S. Brophy (Ed.), *Assessment in music education: Integrating curriculum, theory, and practice* (pp. 3–16). GIA Publications.

Conti, R., Coon, H., & Amabile, T. M. (1996). Evidence to support the componential model of creativity: Secondary analyses of three studies. *Creativity Research Journal*, *9*, 385–389. https://doi.org/10.1207/s15326934crj0904_9

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331–360. https://doi.org/10.1207/s15327043hup1004_2

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565–579. https://doi.org/10.1037/0021-9010.80.5.565

Cook, M. (2016). *Personnel selection: Adding value through people – A changing picture* (6th ed.). Wiley-Blackwell.

Dalal, D., Sassaman, L., & Zhu, X. (2020). The impact of nondiagnostic information on selection decision making: a cautionary note and mitigation strategies. *Personnel Assessment and Decisions*, *6*, 54–64. https://doi.org/10.25035/pad.2020.02.007

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582. https://doi.org/10.1037/0003-066x.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674. https://doi.org/10.1126/science.2648573

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*, 227–268. https://doi.org/10.1207/s15327965pli1104_01

Denis, J. M. (2017). Assessment in music: A practitioner introduction to assessing students. *Update: Applications of Research in Music Education*, *36*, 20–28. https://doi.org/10.1177/8755123317741489

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorith maversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*, 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Edwards, J. A. (1993). Principles and contrasting systems of discourse transcription. In J. A. Edwards and M. D. Lampert (Eds.), *Talking data: Transcription and coding in discourse research* (pp. 3–31). Lawrence Erlbaum Associates.

Ericsson, K. A., Roring, R. W., & Nandagopal, K. (2013). Giftedness and Evidence for Reproducibly Superior Performance: An Account Based on the Expert-Performance Framework. In S. B. Kaufman (Ed.), *The Complexity of Greatness: Beyond Talent or Practice* (pp. 137–190). Oxford University Press.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Guay, J. P., & Parent, G. (2018). Broken legs, clinical overrides, and recidivism risk: An analysis of decisions to adjust risk levels with the LS/-CMI. *Criminal Justice and Behavior*, *45*, 82–100. https://doi.org/10.1177/0093854817719482

Hall, G. C. N. (1988). Criminal behavior as a function of clinical and actuarial variables in a sexual offender population. *Journal of Consulting and Clinical Psychology*, *56*, 773–775. https://doi.org/10.1037/0022-006x.56.5.773

Hennessey, B. A. (1994). The Consensual Assessment Technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal*, *7*, 193–208. https://doi.org/10.1080/10400419409534524

Highhouse, S. (2008). Stubborn Reliance on Intuition and Subjectivity in Employee Selection. *Industrial and Organizational Psychology, 1,* 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Hoffman, M., Kahn, L. B., & Li, D. (2017). Discretion in hiring. *The Quarterly Journal of Economics*, *133*, 765–800. https://doi.org/10.1093/qje/qjx042

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment*, *21*, 264–276. https://doi.org/10.1111/ijsa.12036

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*, 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Kaplan, S. E., Reneau, J. H., & Whitecotton, S. (2001). The effects of predictive ability information, locus of control, and decision maker involvement on decision aid reliance. *Journal of Behavioral Decision Making, 14,* 35–50. https://doi.org/10.1002/1099-0771(200101)14:1<35::aid-bdm364>3.0.CO;2-d

Kaufman, J. C., Baer, J., & Gentile, C. A., (2004). Differences in gender and ethnicity as measured by ratings of three writing tasks. *Journal of Creative Behavior, 39,* 56–69. https://doi.org/10.1002/j.2162-6057.2004.tb01231.x

Kishk, M., Pollock, R., Atta, J., & Power, L. (2005), A structured model for performance assessment in property management. *Journal of Financial Management of Property and Construction*, *10*, 159–170. https://doi.org/10.1108/13664380580001073

Kitchener, R. F. (1982). Holism and the Organismic Model in Developmental Psychology. *Human Development*, *25*, 233–249. https://doi.org/10.1159/000272811

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, *98*, 1060–1072. https://doi.org/10.1037/a0034156

Lai, E. R. (2011). Performance-based assessment: some new thoughts on an old idea. *Bulletin Pearson Education*, *20*, 1–4.

Latimer, M. E., 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*, 168–183. https://doi.org/10.1177/0022429410369836

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. *Journal of Family Medicine and Primary Care*, *4*, 324–7. https://doi.org/10.4103/2249-4863.161306

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*, 629–650. https://doi.org/10.1093/jcr/ucz013

McCoy, C. W. (1991). Grading Students in Performing Groups: A Comparison of Principals' Recommendations with Directors' Practices. *Journal of Research in Music Education*, *39*, 181–90. https://doi.org/10.2307%2f3344718

McQuarrie, S. H., & Sherwin, R. G. (2013). Assessment in Music Education: Relationships Between Classroom Practice and Professional Publication Topics. *Research & Issues in Music Education*, *11*, 1–15.

Moran, S. (2010). Creativity in school. In K. Littleton, C. Wood, & J. K. Staarman (Eds.), *International handbook of psychology in education* (pp. 319–359). Emerald.

Neumann, M., Niessen, A. S. M., & Meijer, R. R. (2021). Implementing evidence-based assessment and selection in organizations: A review and an agenda for future research. *Organizational Psychology Review*, *11*, 205–239. https://doi.org/10.1177/2041386620983419

Neumann, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2021). The autonomy-validity dilemma in mechanical prediction procedures: The quest for a compromise. *Journal of Behavioral Decision Making*, *34*, 1–20. https://doi.org/10.1002/bdm.2270

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*, 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Niessen, A. S. M, Meijer, R. R., & Tendeiro, J. N. (2016). Predicting Performance in Higher Education Using Proximal Predictors. *PloS ONE, 11*. https://doi.org/10.1371/journal.pone.0153663

Nisbett, R. E., Borgida, E., Crandall, R., & Reed, H. (1976). Popular induction: Information is not necessarily normative. In J. Carrol & J. Payne (Eds.), *Cognition and social behavior* (pp. 113–134). Erlbaum. https://doi.org/10.4324/9781315802879

Noddings, N. (2013). Standardized curriculum and loss of creativity. *Theory into Practice*, *52*, 210–215. https://doi.org/10.1080/00405841.2013.804315

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*, 237–51. https://doi.org/10.1177/002242940705500305

Plakiotis, C. (2017). Objective Structured Clinical Examination (OSCE) in Psychiatry Education: A Review of Its Role in Competency-Based Assessment. In: P. Vlamos (Ed.), *GeNeDis 2016. Advances in Experimental Medicine and Biology*, *988* (pp. 159–180). Springer. https://doi.org/10.1007/978-3-319-56246-9_13

Quinlan, A. M. (2006). *A Complete Guide to Rubrics: Assessment Made Easy for Teachers of K–College*. Rowman & Littlefield Education.

Radocy, R. E. (1986). On Quantifying the Uncountable in Musical Behavior. *Bulletin of the Council for Research in Music Education*, *88*, 22–31.

Riessman, C. K. (1993). *Narrative analysis*. Sage.

Rohwer, D. A. (1997). The challenges of teaching and assessing creative activities. *Update: Applications of Research in Music Education, 15*, 8–12. https://doi.org/10.1177%2f875512339701500203

Runco, M. A. (1989). The creativity of children’s art. *Child Study Journal*, *19*, 177–189.

Runco, M. A., Hyeon Paek, S., & Jaeger, G. (2015). Is creativity being supported? Further analyses of grants and awards for creativity research. *Creativity Research Journal, 27*, 107–110. https://doi.org/10.1080/10400419.2015.992692

Russell, J. A., Austin, J. R. (2010). Assessment practices of secondary music teachers. *Journal of Research in Music Education, 58*, 37–54. https://doi.org/10.177/0022429409360062

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology*, *40*, 455–488. https://doi.org/0.1111/j.1744-6570.1987.tb00610.x

Sherden, W. (1998). *The fortune sellers: The big business of buying and selling predictions*. Wiley.

Stevens, D. D., & Levi, A. A. (2005). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Stylus.

Tienken, C. H. E. D. (2016). *Defying standardization: Creating curriculum for an uncertain future*. Rowman & Littlefield.

Tuckett, A. G. (2005). Applying thematic analysis theory to practice: A researcher's experience. *Contemporary Nurse*, *19*, 75–87. https://doi.org/10.5172/conu.19.1-2.75

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *184*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wesolowski, B. C. (2012). Understanding and developing rubrics for music performance assessment. *Music Educators Journal*, *98*, 36–42. https://doi.org/10.1177/0027432111432524

Wiggins, N., & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology*, *19*, 100–106. https://doi.org/10.1037/h0031147

Appendix A

Individual Interview Protocol

Introduction

Meet the participant in video conference through Google Meet. Welcome the participant and thank them for their participation. Reiterate that the interview will be recorded as outlined in the information and informed consent form, then transcribed and used in research anonymously. Explain the rationale behind the current study based on the following: knowledge and understanding of assessors' attitudes towards different assessment methods are critical for improving these methods, and the insights of the participants contribute valuable data to this. Explain the format of the current study based on the following: an approximately 45-minute semi-structured interview where the researcher will ask a series of open questions about the participant's attitudes towards different assessment methods. Definitions of terms will be provided for clarity when necessary, and the interview will be moderated by the researcher for purposes of time management. Ask if any clarifications are needed so far.

Interview

During the interview, ask the following questions in the given order.

1. What is the goal of music assessment?

Before question 2, explain the concepts of holistic judgment, mechanical judgment, and validity using the following definitions:

Holistic Judgment

Combining information about performance aspects implicitly (i.e., by intuition “in your head”) to determine a final judgment. For example, an assessor uses a 1 to 5 point scale to rate a violin performance, where 1 = beginner and 5 = master. The scale can be with or without explicitly defined aspects of importance, or separate ratings of these aspects. The assessor determines the final judgment by *thinking* about what level of the scale the performance aspects correspond to.

Mechanical Judgment

Combining information about performance aspects explicitly (i.e., through a pre-determined procedure) to determine a final judgment. For example, an assessor uses the scale in the previous example to rate a violin performance, using explicitly defined aspects of importance and separate ratings of each of these aspects. The assessor determines the final judgment by *calculating* what level of the scale the performance aspects correspond to through entering the separate aspect ratings in a formula that contains pre-determined weights for each aspect.

Validity

How well the examination outcomes (grades) represent the quality of students’ music performance and are not influenced by unintended or irrelevant factors.

2. How is valid assessment achieved in music performance?
    a. What approach (holistic and/or mechanical) to assessment results in more valid assessments of music performance? Why?

Before question 3, explain the concept of reliability using the following definition:

Reliability

How consistent the examination outcomes (grades) are for the same students, assessed by different examiners. Validity is not possible without reliability, as valid assessment needs to be consistent.

3. How is reliable (consistent) assessment achieved in music performance?
    a. What holistic and/or mechanical approach to assessment results in more reliable (consistent) assessments of music performance? Why?

Before question 4, explain the concept of fairness using the following definition:

Fairness

Equality of treatment and transparency of the assessment method so that students know what they are being graded on and how their grade is determined.

4. How is fair assessment achieved in music performance?
    a. What holistic and/or mechanical approach to assessment results in more fair assessments of music performance? Why?

5. What obstacles do you see in realizing valid and fair assessment of music performance?
    a. How does this relate to holistic or mechanical approaches?

Conclusion

Ask the participant to briefly share if and how the preceding discussion changed their ideas of the topics covered, and what questions they would have added to a discussion of this kind. Ask if anything is unclear at this point, and if not, thank them for their participation and conclude the interview.

Appendix B

Group Interview Protocol

Introduction[5]

Meet the participants in video conference through Google Meet. Welcome the participants and thank them for their participation. Reiterate that the interview will be recorded as outlined in the information and informed consent form, then transcribed and used in research anonymously. Explain the rationale behind the current study based on the following: knowledge and understanding of assessors' attitudes towards different assessment methods are critical for improving these methods, and the insights of the participants contribute valuable data to this. Explain the format of the current study based on the following: an approximately 90-minute semi-structured interview where the researcher will ask a series of open questions about the participants' attitudes towards different assessment methods that can be answered by, and subsequently actively discussed between, any and all group members. Definitions of terms will be provided for clarity when necessary, and the interview will be moderated by the researcher for purposes of time management. Ask if any clarifications are needed so far.

Interview

Ask the participants to digitally raise their hands when they wish to speak and to actively provide space for other voices to be heard, and reiterate that the interview as a whole will be moderated by the researcher throughout. Ask if any clarifications are needed so far. During the interview, ask the following questions in the given order.

1. What is the goal of music assessment?

[5] The video conference for the group interview was preceded by a presentation from a student colleague discussing the results of their previous study, which the present participants had participated in.

Before question 2, explain the concepts of holistic judgment, mechanical judgment, and validity using the following definitions:

Holistic Judgment

Combining information about performance aspects implicitly (i.e., by intuition “in your head”) to determine a final judgment. For example, an assessor uses a 1 to 5 point scale to rate a violin performance, where 1 = beginner and 5 = master. The scale can be with or without explicitly defined aspects of importance, or separate ratings of these aspects. The assessor determines the final judgment by *thinking* about what level of the scale the performance aspects correspond to.

Mechanical Judgment

Combining information about performance aspects explicitly (i.e., through a pre-determined procedure) to determine a final judgment. For example, an assessor uses the scale in the previous example to rate a violin performance, using explicitly defined aspects of importance and separate ratings of each of these aspects. The assessor determines the final judgment by *calculating* what level of the scale the performance aspects correspond to through entering the separate aspect ratings in a formula that contains pre-determined weights for each aspect.

Validity

How well the examination outcomes (grades) represent the quality of students’ music performance and are not influenced by unintended or irrelevant factors.

2. How is valid assessment achieved in music performance?

   a. What approach (holistic and/or mechanical) to assessment results in more valid assessments of music performance? Why?

Before question 3, explain the concept of reliability using the following definition:

Reliability

How consistent the examination outcomes (grades) are for the same students, assessed by different examiners. Validity is not possible without reliability, as valid assessment needs to be consistent.

3. How is reliable (consistent) assessment achieved in music performance?
   a. What holistic and/or mechanical approach to assessment results in more reliable (consistent) assessments of music performance? Why?

Before question 4, explain the concept of fairness using the following definition:

Fairness

Equality of treatment and transparency of the assessment method so that students know what they are being graded on and how their grade is determined.

4. How is fair assessment achieved in music performance?
   a. What holistic and/or mechanical approach to assessment results in more fair assessments of music performance? Why?

5. What obstacles do you see in realizing valid and fair assessment of music performance?
    a. How does this relate to holistic or mechanical approaches?

Conclusion

Ask the participants to briefly share if and how the preceding discussion changed their ideas of the topics covered, and what questions they would have added to a discussion of this kind. Ask if anything is unclear at this point, and if not, thank them for their participation and conclude the interview.