Musical Performance Evaluation: Ten Insights from Psychological Studies of Aesthetic Judgment

PATRIK N. JUSLIN

Abstract: Performance evaluation is of great importance to the development of musical performance. Yet, we know little about the psychological process that underlies such evaluations. In this essay, I argue that performance evaluation is essentially a form of aesthetic judgment, and that recent findings from psychological studies of aesthetics may provide valuable insights. First, I present a preliminary model of aesthetic judgment. Then, I outline a methodological paradigm which has proved useful in capturing the judgment process. This is followed by a consideration of ten insights about aesthetic judgment of music from recent studies. Finally, implications for music education are discussed. I ask whether there is such a thing as 'good' performance evaluation, and – if so – what this might entail. It is proposed that evaluation is a skill that can be trained based on feedback from analytic models which make the aesthetic judgment process transparent to musicians and listeners alike.

Keywords: aesthetics, evaluation, judgment, music, performance, reliability

1. Introduction

Evaluations of music performance occur in a large number of contexts, and have important consequences: They are made when music students apply for admission to a conservatoire, and when musicians audition for a job in an orchestra; they are made, repeatedly, by record producers; they guide consumers and have an effect on music sales and download numbers; they are used in studies of music performance (e.g., in studies of sight-reading ability); and they contribute to musical experiences in everyday life. Evaluations may also be seen in the daily work of music teachers – for example when a teacher gives feedback about a student's performance during a lesson to help the student to improve the performance.

Understandably, then, performance evaluation has received a great deal of attention in the education literature (see Barry, 2009; Bergee, 2003; Fiske, 1983; Parkes, 2010; Schleuter, 1997; Waddell et al., 2019). McPherson and Schubert (2004) provide a helpful discussion of contextual factors that can impact on the evaluation of a performance, with regard to fairness and reliability. However, surprisingly few studies to date have investigated the *psychological* characteristics of the evaluative process itself.

Still, in order to legitimize the use of performance evaluation in various forms of 'high stakes' testing (e.g., in contests, end-of-course examinations, orchestra auditions, and degree classifications), there are some important questions that should be addressed:

- Do different judges agree in their evaluation of a performance?
- Does a judge make stable evaluations of the same performance over time?
- Which criteria do judges rely on in their evaluations?
- Do different judges rely on the same set of criteria?
- Is performance evaluation affected by musical expertise?

Journal of Comparative Literature and Aesthetics Vol. 46, No. 1, Spring 2023 [93-106]

- Do judges understand how they make their evaluations?
- Can the evaluative process be made more transparent?
- Is there such a thing as 'good' performance evaluation?

2. Performance Evaluation as Aesthetic Judgment

In this essay, I will focus on performance evaluation within the context of the *work* vs. *performance* distinction, which is typical of most notated Western classical music. The focus is on evaluating performances as *performances of a particular work*; the object of evaluation is the specific sounds deriving from the performer's activity (Levinson, 1987).

I will further make a distinction between the *measurement* of a performance (which is 'objective') and the *evaluation* (assessment, judgment) of a performance (which is – more or less – subjective). Although some aspects of performance may be measured objectively (e.g., intonation; e.g., Gabrielsson, 1999), it might be argued that the most *important* performance aspects (e.g., interpretation, musicality, expressivity, and originality) cannot be satisfactorily measured in a truly objective manner.

In fact, McPherson and Schubert (2004) labeled it one of the "flawed assumptions" (p. 65) about performance evaluation that the musical value of a performance could be assessed accurately and reliably – that we can somehow access the 'true' value of a performance. One problem is that performance evaluation is affected not only by the music, but also by various non-musical factors (see next section).

However, even if the evaluator would be able to focus only on the *musical* features of a performance, his or her evaluation would still not be objectively accurate in a straightforward sense (which is not to say that we should not strive towards making evaluation as reliable and valid as it can be). This reflects the nature of the evaluative process itself: assessment requires judgments about *value*, and music is commonly regarded as one of the fine arts (Kivy, 1991).

Accordingly, I will argue that music performance evaluation is, ultimately, a form of *aesthetic* judgment, and that we may hopefully gain some valuable insights from studies of aesthetic judgment in music. Clearly, a first step toward better performance evaluation is to understand the underlying psychological process.

I define *aesthetic judgment* here as a process by which the value of a piece of music as 'art' is determined, based on one or more subjective criteria (e.g., novelty, expressivity, and beauty) which relate to properties of the artwork, either its form or its content (Juslin, 2013). Moreover, I submit that aesthetic judgments of music are neither completely 'objective', nor merely 'subjective': They involve psychophysical *interactions* between objective properties of the music and person-dependent impressions of the judge. Thus, *there are no absolute or universal criteria for aesthetic value.* As any historical review would demonstrate, aesthetic norms change over time in society.[1]

Thus, for example, when it comes specifically to musical performance, philosopher Jerrold Levinson (1987) observes that "performances of music are *legitimately* evaluated from a number of different perspectives". He suggests that "there is no *single*, *overriding* point of view concerning performances such that whatever seems good from that point of view qualifies in effect as an *absolutely* good performance of the work"; he concedes that there might well be "a *particular* point of view that is arguably most *central* to evaluative assessment" (p. 75). Yet, "there is no simple answer to how good a performance is" since every evaluation is dependent on "a context of assessment in which certain objectives are taken as paramount" (p. 82).

Does this suggest that aesthetic judgments are necessarily arbitrary, idiosyncratic, and unreliable? No, aesthetic judgments may actually be more systematic than is often assumed, once one takes a closer look at their characteristics. Aesthetic judgments can be statistically modeled. Doing so invites us to consider more closely *when*, *how*, and *why* such judgments of music differ and what – if anything – can be done about it.

3. A Preliminary Model

In this essay, I will adopt a psychological model outlined by Juslin (2013, 2019), which focuses specifically on aesthetic judgment in music experience: Figure 1. Aesthetic judgment is regarded as one of several psychological mechanisms that may evoke emotions in listeners during music listening.

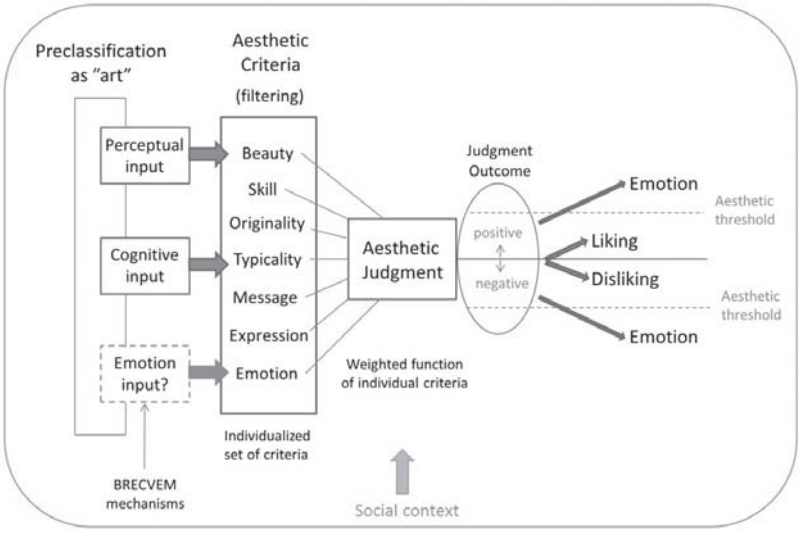

Figure 1

Consistent with many theories in aesthetics (see Levinson, 2003), it is assumed that the aesthetic judgment process begins when the listener adopts an *aesthetic attitude* to the music. This is a particular *mode* of listening that brings a focus on those properties of the music that are regarded as relevant for its value as 'art': The listener's attention is focused on the music and one or more criteria for aesthetic value are brought to bear on the music.

Once an aesthetic attitude has been adopted, both perceptual and cognitive analyses of the music will proceed, providing 'inputs' to the aesthetic judgment process (Figure 1). This can be construed as a continuously on-going process. Aesthetic processing can be influenced by factors in the *music*, the *perceiver*, and the *situation*. Information related to each of these factors is channelled through the perception (e.g., sensory impressions of low-level features), cognition (e.g., input that depends on conceptual knowledge), and emotion (induced by other psychological mechanisms) of the listener (Juslin, 2013). However, whether these inputs will have an effect on the resulting aesthetic judgment depends on the listener's criteria: Figure 1. The criteria serve as 'filters' and determine what information is relevant to the judgment task.

Based on preliminary survey findings regarding aesthetic criteria (Juslin & Isaksson, 2014), the model postulates that judgments involve individual sets of subjective criteria for aesthetic value and a relative weighting of the criteria. The model illustrates that judges can differ in terms of *how many* criteria they use, *which* criteria they use, and how these criteria are *weighted*. The overall judgment represents a weighted function of the specific criteria.

As regards evaluation of *performances* within the context of instrumental teaching, the specific aspects evaluated may involve personal criteria, criteria that have been identified by various authorities, or a combination of these two (for examples, see Thompson et al., 1998). As noted by Gabrielsson (1999), "there are hardly any agreed-upon criteria, neither for what should be judged, nor for how the judgments should be made" (p. 577).

However, McPherson and Schubert (2004) argued based on the published literature that there are at least four general types of competencies, which are commonly assessed by music institutions: *technique* (physiological, physical and instrumental), *interpretation* (e.g., faithful reading of the score; authenticity, in terms of understanding the style or composer intentions), *expression* (e.g., the projection of mood, conveying of structural aspects), and *communication* (e.g., confidence, holding the audience's attention).

An assessment of specific criteria occurs continuously, but judgment outcomes will be produced at particular points in time (cued by significant moments in the music, such as the ending of a piece, or when a teacher needs to provide feedback on a student's performance). If the judgment process indicates that on balance the performance is good, this will result in liking (preference). Conversely, if the process indicates that the music is *not* so good, it will result in disliking. In both of these cases, no emotion is necessarily evoked. If, however, the result of the judgment is that the music is judged as extraordinarily good (or bad), overall or on at least one of the criteria, an emotion (e.g., awe) will be aroused in addition to liking. In this essay, I focus on the judgment process *per se*, rather than on the affect it might induce.

Social context (Figure 1, bottom) refers to the fact that although aesthetic judgments of music are, ideally, mostly influenced by evaluations of criteria related to the music itself, any music evaluation is affected by other factors which impact on the reliability of the evaluation (McPherson & Schubert, 2004). Factors such as visual impressions, social prestige, audience support, and stereotyping could all affect aesthetic judgments. In the following, I will mostly leave out contextual factors and focus on how musical aspects are evaluated.

4. Analyzing Judgments

It has been suggested that judges may be unaware of which criteria they actually use in their assessments (Gabrielsson, 2003). This seems to present us with a problem: How can we capture the listener's judgment strategy if it is mostly tacit and cannot be reported accurately by the listener?

From cognitive psychology, we might obtain some initial clues about the nature of the aesthetic judgment process and also a useful analytic paradigm for studying such judgments. The paradigm is termed *Judgment Analysis* (Cooksey, 1996). Inspired by Brunswik's (1956) *lens model*, judgment analysts employ multiple regression models to capture how individual judges combine multiple differentially weighted pieces of information to arrive at an overall judgment. Similarly, if we wish to study aesthetic judgment in music, we can ask listeners to rate the aesthetic value of pieces of music that vary in different aesthetic dimensions such as novelty. The aim would be to predict listeners' overall judgments based on criteria.

Although objective features of the music can be manipulated and have an effect on the listener, such effects are 'mediated' by the *perception* of the listener. Previous studies found that subjective impressions (e.g., subjectively perceived complexity) are better predictors of responses than objective measures (e.g., objectively measured complexity; cf. Hargreaves & North, 2010). Hence, it makes more sense to use subjective impressions of criteria – as rated by listeners – as predictors of overall judgments in multiple regression analyses.

The strength of this method is that a listener's judgment strategy can be extracted in the statistical analysis independently of any conscious awareness of criterion use that the listener may have. The judgment strategy is revealed by the complex statistical interdependencies of overall judgments and rated criteria. Thus, multiple regression models can provide measures of the different aspects of the judgment process (as illustrated in Figure 1).

5. Ten Insights From Studies of Aesthetic Judgment

Armed with the model and the analytic paradigm proposed above, our research group has investigated different aspects of the aesthetic judgment process in a series of studies. In the following, I will

consider some potentially important insights from these studies as well as some pioneering studies in music education, which align nicely with research conducted in psychology. Not all of the studies focused specifically on aspects of music performance, but it is a reasonable assumption that the underlying cognitive process will not be radically different if a judge focuses merely on the performance, as opposed to the music as a whole (e.g., regardless of focus, the process will necessarily involve a weighting of criteria which reflects the processing constraints of human cognition).

(1) Judges generally show low inter-rater agreement

Because aesthetic judgments of music have important consequences in many contexts, it appears relevant to investigate whether judges are reliable (cf. Fiske, 1983; Juslin et al., 2021; Manturzewska, 2011). First, we need to consider *inter-judge* reliability: Do different listeners make similar (overall) aesthetic judgments of the same music? In two listening tests (Juslin et al., 2016, 2021), listeners varying in musical expertise judged the overall aesthetic value of 72 musical excerpts from 12 different genres, selected by means of a *stratified random sampling* procedure to obtain a reasonably broad and representative sample of music. In order to obtain a measure of the inter-rater agreement in the listeners' judgments, we computed the intraclass correlation coefficient. The coefficients for three groups (.18, .24, and .16, respectively) were all indicative of 'poor' inter-rater reliability.

A music educator may perhaps object that the above studies did not focus specifically on performances or that the judges were not experienced enough. However, our results may not be explained away that easily. A previous study of performance evaluation offers rather similar results. Maturzewska (2011) analyzed 2156 ratings given by 28 members of the jury of the Sixth International F. Chopin Piano Competition in Warsaw, where they evaluated 77 performances of one of Chopin's Polonaises. The members of the jury were "music experts of the highest international reputation" (p. 99). Maturzewska observed very large individual differences in judgments.

Reflecting on ratings of particular performances, Manturzewska (2011) asked "which judge was 'right': the one who gave performance B 14 points, i.e., fair, or the one who gave 25 points, the highest possible rating? And who is right for performance E? The judge who gave 1 point (almost the worst rating possible) or the one who rated the performance at 16, i.e., good?" (p. 104). Manturzewska was puzzled as to why members of the jury differed to such an extent: "What are the reasons for these inconsistencies?" (p. 104).

(2) Judges show moderate intra-rater reliability

A second aspect of judge reliability concerns the *intra-rater* reliability: Does the same listener make similar judgments over time? Here, previous findings are somewhat mixed. In one of our listening tests, one of the music examples was – unbeknownst to the participants – repeated (appearing 24 excerpts apart, to reduce memory effects; Juslin et al., 2021).

In order to obtain an estimate of intra-judge reliability (or stability), we tested the mean difference in rating of the repeated piece between the first and second trial, using a dependent samples *t*-test. Although the piece was rated slightly higher in aesthetic value the second time than the first time (Cohen's $d = 0.132$), the difference was not statistically significant; and the test-retest reliability was $r = .713$. (A reliability of greater than or equal to .70 is usually considered 'acceptable'.)

However, the test-retest reliability might be reduced if the judgments are farther apart in time. Höfel and Jacobsen (2003) studied the temporal stability of aesthetic judgments of visual (graphic) patterns. In a first session, psychology students were asked to judge 252 patterns. In a second session, the same participants were asked to categorize 80 of the same patterns again.

Owing to external circumstances, the time span between these sessions varied from one day to 14 months, which made it possible to study the temporal stability. When the time span was only a few days, the two judgments were relatively stable. (In the most trivial case where the two ratings of a repeated stimulus occur *very* close in time, a judge may simply *remember* both stimulus and score.) However, with longer time spans, judgments differed significantly.

In addition, in accordance with the interactionist view of aesthetics described in Section 2 (i.e., that judgments depend on both characteristics of the *stimulus* and characteristics of the *judge*) intra-rater reliability can also be moderated by the type of stimuli judged. Fiske (1983) studied the intra-judge reliability of experienced musicians, when they rated a set of *different performances of the same piece*. This is arguably a more 'difficult' set of musical excerpts to judge. Unbeknownst to the judges, some of the performances were repeated so that the raters provided two ratings of these performances. Fiske discovered *very* low correlations between the first and second set of ratings of the same performances. He argued that judges may have applied the criteria inconsistently, but as explained below, most judges are quite consistent.

A more likely explanation could be that there were strong *order effects* in the test (e.g., Flôres & Ginsburgh, 1996), which influenced the listeners' impressions (and thus ratings) of the criteria, and that these order effects had an overriding effect on the overall judgments, as compared to the subtle differences between fairly similar performances of the same piece.

(3) Judges are usually internally consistent

In this particular context, the term *consistency* (sometimes also called *cognitive control* in the judgment-analysis literature; Cooksey, 1996) refers to the degree to which the judge is able to execute a judgment strategy in a consistent manner across different cases (as opposed to the *same* case, like when estimating intra-rater reliability).

Note that high consistency is necessary (but not sufficient) in order for judges to have high (interrater and intrarater) reliability. And if a judge is consistent in executing a specific judgment strategy (i.e., high cognitive control), an analyst will find it is easier to model and predict his or her judgments across cases than if the judge is inconsistent. Thus, for instance, in Judgment Analysis, a large multiple correlation is usually regarded as indicative of a high internal consistency of the judge (Cooksey, 1996).

In the studies mentioned earlier (Juslin et al., 2016, 2021), we modeled music listeners using multiple regression analyses that aimed to predict the judges' overall judgments based on the ratings of individual criteria. The mean multiple correlation ($R = .88$) showed that the models provided a good fit to the data; that the judgment process was systematic and mainly additive; and that the judges were very consistent. Thus, to explain low inter-rater reliability in aesthetic judgments, we need to look beyond judges' internal consistency.

(4) Judges tend to rely on only a few criteria

Although preliminary survey data indicated that a potentially large number of criteria may be involved (Juslin & Isaksson, 2014), there is reason to suspect that individual judges actually utilize a much smaller number. To start with there are working memory limitations suggesting that judges should not use more than about four criteria (Cowan, 2010). General studies of human judgments have found that judges in different domains tend to use a small number of criteria, fewer than the judges themselves report (Brehmer & Brehmer, 1988). Is the same true of aesthetic judgments?

A fair indication of which criteria listeners rely on in their aesthetic judgments is given by the beta weights (ß) of the predictors (i.e., the aesthetic criteria) in each listener model. In accordance with previous studies, a predictor may be considered used by the judge if its beta weight is significant (Harries et al., 2000). Although the number of criteria varied, we found that most judges relied on between one and three criteria ($M = 2.38$ and 2.29, respectively, in two samples; Juslin et al., 2016, 2021).

(5) Individual differences between judges reflect the use of different criteria

An early indication that this is the case came from a study by Thompson et al., (1998), which adopted a new approach, the so-called *repertory grid technique*. Five adjudicators were asked to report the constructs (or criteria) they used to compare and evaluate six expert performances of a Chopin

etude. The adjudicators were then asked to rate the performances, using these same constructs. The results revealed that the adjudicators used different criteria, but since not all adjudicators rated the same criteria, the ratings are not directly comparable.

Further evidence comes from our studies of listeners' aesthetic judgments (Juslin et al., 2016) where judges rated musical excerpts on criteria selected based on aesthetics (cf. Juslin, 2013; Levinson, 2003), as well as survey data of performers and listeners (Juslin & Isaksson, 2014). Table 1 shows some examples of data for individual judges, in terms of both multiple correlations and beta weights. Note the strikingly different weighting schemes for the criteria of different listeners. Thus, for instance, listener 2 seems to rely much on *emotion*, but not on *expression*, whereas the opposite is true for listener 1. How could we expect two listeners to make a similar overall judgment, if they rely on different criteria?

		Predictors (Criteria)						
Listener	***R***	**Beauty**	**Originality**	**Expressivity**	**Skill**	**Emotion**	**Message**	**Typicality**
1.	.95	.13	-.09	.65*	.28*	.09	.02	.17*
2.	.89	.05	.32*	-.31	.44	.43	.24	-.16
3.	.64	.35	.06	.05	.15	.06	.29	.13
4.	.96	.33*	.06	.13	.60*	-.08	-.12	-.03
5.	.91	.35*	.23*	-.03	.40*	.31*	-.19	.04
6.	.97	.04	.59*	.45*	.12	-.28*	.10	.16*

Table 1: Examples of individual regression models of aesthetic judgments by six listeners

Note: *R* refers to the multiple correlation of the regression models, which indicate the extent to which overall aesthetic judgments could be predicted based on a linear combination of the criteria for aesthetic value. Criterion data indicate the relative weight of the criteria for each listener in the judgment process (* indicates that a beta weight was statistically significant, $p < .05$).
[From Juslin et al. (2016). Adapted with permission from the American Psychological Association.]

(6) Criteria emphasized by music institutions may not overlap with those of listeners

Although some music institutions have explicitly identified performance criteria to be used in assessments (McPherson & Schubert, 2004), a crucial question is how these criteria compare to those used by regular music listeners in actual judgments of performances. I am not currently aware of any study that has systematically compared the two. (In fact, there is hardly any research on how 'ordinary' listeners (e.g., audiences) judge a performance.) One can, however, make an informal comparison of the criteria emphasized by music institutions (as listed by McPherson & Schubert, 2004, p. 3) with the criteria rated as most important by listeners in questionnaire research (Juslin & Isaksson, 2014) or that correlate most strongly with actual judgments (Juslin et al., 2016, 2021).

Comparisons reveal that music institutions and ordinary listeners share a concern with *technical skill* and *expression*, but that listeners (including musicians) rate *originality* highly and that they also consider it important to be *emotionally moved*; none of these criteria are, it seems, explicitly emphasized in music education. In contrast, *authenticity* (with regard to the composer's intentions or the style) is emphasized by music institutions, but does not seem to play an important role in judgments by adjudicators or listeners.

Further data comes from studies of critical reviews of music performances. Alessandri et al. (2016) content-analyzed 100 reviews of recordings of Beethoven sonatas, published in *Gramophone* between 1934 and 2010. By utilizing a new combination of data reduction and thematic analysis, they extracted some consistent themes in the reviews (e.g., novelty, style, emotion, technical skill). These themes seem to correspond better with the criteria suggested by aestheticians and studies of musicians and music listeners than with the ones emphasized in rubrics or scales used by music institutions.

(7) Judges have generally low self-insight

In order for a teacher to be able to provide informative feedback to a student, the teacher has to infer what it is that needs to be addressed; and this, in turn, will depend on the teacher's aesthetic judgments – as well as on his or her self-insight. *Self-insight* refers here to the degree to which judges have an accurate understanding of their own judgment strategy: do they know which criteria they rely on and how these are weighted?

One clue comes from studies of cognition, which show that people are often unable to correctly explain the basis of their own judgments and preferences (Nisbett & Wilson, 1977). Because judgments are usually intuitive, and based on underlying processes which cannot be elucidated through introspection, judges develop their own implicit causality theories, which correspond poorly with their actual behavior. In their review of multiple-cue judgment tasks, Brehmer and Brehmer (1988) observed that "most subjects in most studies show little insight into their own judgmental processes" (p. 107).

We examined self-insight in a musical setting, using a two-part study design. In the first part, listeners were required to fill out a brief questionnaire and to rate the relative importance of various criteria for their aesthetic judgments of music. In the second part, the same listeners took part in a listening test where they rated 50 pieces of music from various genres regarding ten criteria, as well as overall aesthetic value. To explore self-insight, we compared subjective ratings of criterion importance with statistically recovered patterns of criterion weighting from the regression models of actual judgments (Juslin et al., 2021).

What we found was that most listeners showed a low level of self-insight concerning their own judgment strategies – as shown by the fact that little variance was shared between subjective ratings of criterion importance and objective measures recovered from judgment models ($M = 32\%$); that is, 'objective' indices of which criteria actually went together with the listeners' overall judgments indicated that they frequently relied on criteria they did not emphasize in their self-report or that they emphasized criteria that they did not actually use.

However, we also noticed that there were wide individual differences in self-insight: For some judges, the 'subjective' ranking of criteria corresponded well with the 'objective' ranking; for others there were considerable discrepancies. Notably, individual level of self-insight was correlated with variance accounted for in the judgment model ($r = .43$), which suggested that judges with greater self-insight may also be more consistent.

(8) Judges' musical expertise has a limited influence on their judgments

One factor that might help to explain individual differences in aesthetic judgments and criterion use is expertise. People tend to believe that experts possess an ability to make finer discriminations than non-experts. We examined the role of musical expertise in our listening tests, where the listeners judged musical excerpts (Juslin et al., 2021). Expertise was indexed in three ways: the listener's (1) extent of formal music education, (2) experience of playing a musical instrument and (3) frequency of focused music listening. To our surprise, judges with a high level of musical expertise did *not* display higher inter-rater agreement, greater internal consistency, or more self-insight.

However, we did find one meaningful link with musical expertise: The number of years a listener had played an instrument was positively correlated with the number of criteria he or she used to make judgments ($r = .33$); that is, the longer the listener had played an instrument, the larger number

of criteria he or she used. This is consistent with previous research showing that experts use a larger number of criteria than non-experts (Brehmer & Brehmer, 1988).

(9) (Some) experts may be less subject to 'biases' in their judgments

Perhaps, there is another category of musical expertise, which may have a much larger impact on aesthetic judgments than those considered so far? Lundy (2010) has examined the inter-judge reliability amongst professional music critics. His study featured 5.161 randomly chosen albums in popular music, covering nine genres (of popular music) and ratings by 352 critics compiled from books of reviews, thus resulting in a total of 15.220 album ratings.

Pairs of critics who had rated at least 30 of the same albums were compared to obtain an estimate of consensus. Lundy used a correlation index because it effectively taps into a similar *pattern* of ratings of two judges across albums. Using this index, Lundy found that 87% of the critic pairs showed statistically significant and positive correlations. In contrast, no significant negative correlations were found. The average correlation, which was moderately positive ($r = .49$), meant that only 24% of the variance was shared between the critics.

Even so, critics might still show greater consensus than lay listeners. In a subsequent study, Lundy and Smith (2017) compared the aesthetic judgments of 50 randomly selected albums by professional critics with those by non-professional undergraduates. Critics, who had previously reviewed these albums, showed relatively consensual ratings (mean $r = .61$), and their distributions approached normality. Conversely, relatively few of the lay listeners' ratings were positively correlated (mean $r = .08$) and their distributions deviated more from normality. Strikingly, none of the non-professionals' ratings was correlated with the critics' ratings.

One possible explanation is that lay listeners are more 'biased' in their judgments, than are professional critics. Lundy (2016) provided a list of 11 biases, which "commonly operate in aesthetic judgment, especially among laypersons" (p. 8) – for instance being influenced by unequal levels of familiarity across works; being unduly influenced by one's place in history; basing one's judgment of an artwork on others' reactions to it; assuming that entire genres of artworks are virtually *all* good or bad; basing one's judgments predominately on the topic of an artwork; having unjustifiably negative attitudes toward certain types of art associated with an 'out-group' (distinguished from one's 'in-group'); basing one's judgment on idiosyncratic characteristics of the self that are not relevant (e.g., memories); and making a judgment when one is not in a mental state that is conducive to competent appraisal (e.g., intoxicated).

Only some of the listed biases have been the subject of study in music, though recent studies indicated, for instance, that lay listeners tend to be more influenced by familiarity in their aesthetic judgments than are professional critics (Lundy et al., 2019), and that they are also more subject to personal idiosyncrasies (Lundy et al., 2018), such as basing a judgment on feelings of nostalgia. It can perhaps be tempting to think that if we can only remove such biases from lay listeners' judgments, their ratings will become virtually identical to those of professional critics. As Lundy (2016) argued, "disagreements are not expected to disappear, but they should decrease when the background noise is reduced, and people are disagreeing about aesthetic factors only" (p. 21).

(10) Inter-rater reliability can be enhanced by enforcing a reliance on similar criteria

It has been suggested that critics show higher inter-rater agreement than lay listeners because they tend to converge on more similar sets of criteria in their judgments (cf. Juslin, 2019). This raises an important question: Can inter-rater reliability be enhanced by forcing judges to use the same set of criteria? Indeed, several studies of performance evaluation in music education have reported that formal rating scales can be helpful (Fautley & Colwell, 2018; Latimer et al., 2010; McPherson, 1995; Parkes, 2010).

Results show, for example, that fixed criteria may help a music faculty to grade more consistently in jury settings, and that they grade with higher reliability if they use particular criteria, as opposed

to only giving a global grade based on an overall impression of a music performance (Bergee, 2003). This is, presumably, because a standardized rating scale helps judges to recall all relevant criteria.

Standardized rating scales are not without problems, however. As Parkes (2010) notes, teachers may find it hard to agree on descriptors or may 'adapt' new descriptors to their own existing evaluation process. There may also be resistance to the task of verbally labelling the most important aspects of performance (Thompson et al., 1998) – that is, the aesthetic aspects. There are further preliminary results indicating that the use of segmented scales can influence the relative balance of evaluation, by producing higher ratings for technical aspects and lower ratings for expressive aspects (Iusca, 2014). Thus, although a standardized scale could clearly have a positive effect on judge reliability, it raises questions about the *validity* of such ratings.

6. Implications for Music Education

What are the implications of the reviewed research for the field of music education? Manturzewska (2011) argued that "we must be very cautions in accepting point scores for performance even when they are given by people with the highest level of competence in instrumental performance" (p. 107). Fiske (1994) was even more blunt, claiming that the "evaluation of a performer does not mean anything, until we know how reliable the judge was who evaluated that performance" (p. 76).

In my estimation, musical performance evaluation will tend to be reliable as long as the assessment involves basic-level instrumental teaching, where the focus is primarily on identifying technical deficiencies of a performance, particularly when evaluators rely on a standardized rating scale (McPherson, 1995). Somewhat paradoxically, when the level of performance reaches beyond basic technical competence, the inter-rater agreement might actually decrease because comparisons will then instead hinge more on ill-defined, subtle and elusive criteria such as 'musicality', 'expression' and 'originality' (*aesthetic* aspects). Here, even experts may not agree about criteria (Thompson et al., 1998).

Indeed, a surprising finding was the overall absence of effects of musical expertise on aesthetic judgment (Juslin et al., 2021). Such judgments may reflect more general cognitive characteristics of an individual that cut across domains. For instance, some individuals may be more reflective and insightful or less biased than others *in general*. Another increasingly plausible possibility is that musical performance evaluation is a *unique* skill that could (and should) be *trained*, alongside the skills of music performance (Waddell et al., 2019). Such a proposition also raises an important question: Is there such a thing as 'good' evaluation – or aesthetic judgment – and, if so, what does this entail?

These are complex issues, and even if we limit ourselves to a psychological perspective, the answers are not obvious. However, what may perhaps be argued is that a 'good evaluator' is someone who: (a) has good self-insight concerning his or her own judgment strategy; (b) is able to apply his or her strategy in a consistent manner across cases; (c) shows good temporal stability in repeated ratings of the same stimulus; and (d) is able to minimize all biases and to disregard irrelevant factors in the evaluation context.

What we *cannot* specify, however, is precisely which criteria a 'good evaluator' should focus on, or how they should be weighted in the overall judgments. (This was, obviously, the *main* factor influencing agreement among judges in previous studies; e.g., Juslin et al., 2016, 2021). In a way, this is the crux of the matter: even if judges have good self-insight and are as consistent as humanly possible, and if they show temporal stability and can remove all biases and contextual effects, there will *still* be disagreements in aesthetic judgment, simply because even the foremost experts will never agree 100% on a precise set of relevant aesthetic criteria and their relative weighting. In the absence of absolute or universal criteria for aesthetic value, standards of performance evaluation will inevitably be somewhat relative and provisionary.

It might be tempting to equate 'good aesthetic judgment' with 'consensus with expert ratings' (Lundy et al., 2019), though apart from the fact that expert judgments correlate only moderately, there is the problem that those composers that are hailed by today's critics may well have been

lambasted by experts back in the day. Even in our current time, it is not that unusual for a professional critic to drastically reappraise a music album within just a couple of years. Expert judgment, it seems, is a shaky ground upon which to base any notion of an *absolutely* good judgment.

An alternative approach could be to evaluate judges in terms of whether they are able to apply criteria in accordance with a formal rating scale for performance evaluation (of a music institution, for instance). However, assume that we enforce adherence to a fixed set of criteria and even their relative weightings (to the degree that judges are really able to fully implement those), such that an 'ideal' music performance is clearly implied and inter-rater reliability can be maximized: Are the resulting judgments still fully *valid*? Are they still *aesthetic* judgments that reflect music as a *creative* art form?

Fautley and Colwell (2018) noted that some performance aspects (e.g., fingering) can be straightforward to evaluate, whereas other aspects (e.g., whether a performance is musical or original) might involve more difficult judgments. In an attempt to increase reliability, it is all too easy to fall back on criteria that are easily assessable, but that are not necessarily valid in measuring aspects of musical learning (Fautley & Colwell, 2018). As observed by Thompson et al., (1998), "criteria used by examining boards for the assessment of music students may be insufficient for the assessment of performances at the highest level of musicianship" (p. 154).

Allowing for the fact that even experts may take different views on aesthetic values that are, in some sense, equally valid (Levenson, 1987), does a 'forced-consensus approach' to the criteria mean that validity is sacrificed to obtain a better estimate of (inter-rater) reliability? If so, is this acceptable or desirable? And if we evaluate performance according to a fixed scale, does this imply that there is only *one* 'correct' performance of a work, such that any notion of 'interpretation' becomes meaningless? If teachers apply a fixed judgment strategy where they, in effect, 'teach to the test', will the resulting (identical) performances appeal to the audience? I do not have the answers to these questions, but I think they deserve reflection and debate.

Findings on aesthetic judgment in music also relate to more practical matters. Thus, for instance, data on what criteria listeners actually rely on most in their judgments of music can have pedagogical implications, for instance, by showing what aspects need specific attention in instrumental teaching. Moreover, the finding that judges tend to have low self-insight (see Section 5) has a crucial implication: If a music teacher's explicit understanding of how he or she makes aesthetic judgments differs from the way in which he or she actually makes those judgments, the teacher may in effect provide misleading instructions to the student.

In addition, my preliminary observation that the criteria included in formal rating scales (or rubrics) used in by educational institutions may not quite overlap with the criteria used by music listeners or critics raises important questions for music education: What criteria should be used, and on what basis? Gabrielsson (2003) argued that the criteria used in assessment of student performance are dominated by technical aspects (e.g., intonation, rhythmic accuracy). Listeners, in contrast, may emphasize expression, originality and emotion. This is significant since pedagogical documents (including rubrics) serve as models for music performance and guide students. As argued by Gabrielsson (2003), "much work remains to establish adequate criteria for the evaluation of music performance" (p. 257).

Even if educators converge on a more adequate set of criteria, performance evaluation remains a skill; and it seems that few music educators receive any formal training in grading performances (e.g., Waddell et al., 2019). Similarly, the training of musicians does not seem to include systematic knowledge about which criteria music listeners and critics use to rate a performance.

Here, the previously proposed analytic paradigm of Judgment Analysis might come in handy. Idiographic regression models of aesthetic judgments could perhaps help to make the judgment process more *transparent* for music teachers, such that they understand it, and – if needed – might alter specific aspects of it. We found that judges who had greater insight into their own judgments were also more consistent. This suggests the possibility that increasing a judge's self-insight (e.g., via feedback) may also improve his or her internal consistency.

Winter (1993) reported findings indicating that the training a music examiner receives prior to the performance assessment session may be more important in producing consistent judgments than amount of previous music examining experience. Similarly, McPherson and Schubert (2004) noted that training may be key to alert the evaluator of subconscious biases (for further discussion, see Lundy, 2016). Waddell et al. (2019) presented a novel tool, "The Evaluation Simulator", to study and train performance evaluation by means of an immersive virtual environment. Our lab has found that computer feedback based on multiple regression models might enhance a performer's communication of emotions (e.g., Juslin et al., 2006). It seems plausible that one can develop similar performance-enhancing computer interventions that involve feedback based on models of aesthetic judgments.

Understanding the process of assessment is, arguably, a key to enhancing one's musical performance (McPherson & Schubert, 2004). Hence, Duke and Byo (2018) suggested that the goal of instrumental teaching should be not only to change each learner's performance for the better, but also to change each learner's *perception* of her own performance: "if learners must rely on the teacher to indicate what sounds good and what does not, and what needs to happen next after every performance trial, then there is little that learners can do on their own time in individual practice" (p. 9). To enhance this process of aesthetic judgment might be one of the most valuable contributions researchers can make towards the goal of helping music students to develop their full potential as music performers.

Uppsala University, Sweden

Notes

[1] Leech-Wilkinson (2006) explains that "a prominent note in a score that in 1910 was emphasized by sliding up to it from the note below, in 1950 might have been emphasized by vibrating on it, and in 1990 by increasing and decreasing its amplitude" (p. 60). Though the inclination to emphasize certain notes remains constant, the means to achieve this differ. Thus "almost every aspect of performance style has changed over the past century" (p. 42).

Works Cited

Alessandri, E., Williamson, V. J., Eiholzer, H., & Williamon, A. (2016), A critical ear: analysis of value judgments in reviews of Beethoven's piano sonata recordings. *Frontiers in Psychology, 7,* 391.

Barry, N. H. (2009). Evaluating music performance: Politics, pitfalls, and successful practices. *College Music Symposium, 49/50,* 246-256.

Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education, 51,* 137-150.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 75-114). Elsevier Science Publishers.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* University of California Press.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* Academic Press.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19,* 51-57.

Duke, R. A., & Byo, J. L. (2018). Building musicianship in the instrumental music classroom. In G. E. McPherson & G. L. Welch (Eds.), *An Oxford handbook of music teaching and learning. Vol. 3* (pp. 165–183). Oxford University Press.

Fautley, M., & Colwell, R. (2018). Assessment in the secondary music classroom. In G. E. McPherson & G. F. Welch (Eds.), *An Oxford handbook of music education. Vol. 2* (pp. 257–271). Oxford University Press.

Fiske, H. E. (1983). Judging musical performances: Method or madness? *Update: Applications of Research in Music Education, 1,* 7–10.

Fiske, H. E. (1994). Evaluation of vocal performances: Experimental research evidence. In G. Welch & T. Murao (Eds.), *Onchi and singing development* (pp. 74–103). David Fulton Publishers.

Flôres, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking? *Journal of the Royal Statistical Society, 45,* 97–104.

Gabrielsson, A. (1999). The performance of music. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 501–602). Academic Press.

Gabrielsson, A. (2003). Music performance research at the millennium. *Psychology of Music, 31,* 221–272.

Hargreaves, D. J., & North, A. C. (2010). Experimental aesthetics and liking for music. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 515–546). Oxford University Press.

Harries, C., Evans, J. St. B. T., & Dennis, I. (2000). Measuring doctors' self-insight into their treatment decisions. *Applied Cognitive Psychology, 14,* 455–477.

Höfel, L., & Jacobsen, T. (2003). Temporal stability and consistency of aesthetic judgments of beauty of formal graphic patterns. *Perceptual and Motor Skills, 96,* 30–32.

Iusca, D. (2014). The effect of evaluation strategy and music performance presentation format on score variability of music students' performance assessment. *Procedia - Social and Behavioral Sciences, 127,* 119–123.

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews, 10,* 235–266.

Juslin, P. N. (2019). *Musical emotions explained.* Oxford University Press.

Juslin, P. N., Ingmar, E., & Danielsson, J. (2021). Aesthetic judgments of music: Reliability, consistency, criteria, self insight, and expertise. *Psychology of Aesthetics, Creativity, and the Arts.* Advanced online publication. https://doi.org/10.1037/aca0000403

Juslin, P. N, & Isaksson, S. (2014). Subjective criteria for choice and aesthetic value of music: A comparison of psychology and music students. *Research Studies in Music Education, 36,* 179–198.

Juslin, P. N., Karlsson, J., Lindström, E., Friberg, A., & Schoonderwaldt, E. (2006). Play it again with feeling: Computer feedback in musical communication of emotions. *Journal of Experimental Psychology: Applied, 12,* 79–95.

Juslin, P. N., Sakka, L. S., Barradas, G. T., & Liljeström, S. (2016). No accounting for taste? Idiographic models of aesthetic judgment in music. *Psychology of Aesthetics, Creativity, and the Arts, 10,* 157–170.

Kivy, P. (1991). Is music art? *The Journal of Philosophy, 88,* 544–554.

Latimer, M. E., Jr., 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.

Leech-Wilkinson, D. (2006). Expressive gestures in Schubert singing on record. *Nordic Journal of Aesthetics, 33,* 51–70.

Levinson, J. (1987). Evaluating musical performance. *Journal of Aesthetic Education, 21,* 75–88.

Levinson, J. (2003). Philosophical aesthetics: An overview. In J. Levinson (Ed.), *The Oxford Handbook of Aesthetics* (pp. 3–24). Oxford University Press.

Lundy, D. E. (2010). A test of aesthetic consensus among professional modern music critics. *Empirical Studies of the Arts, 28,* 243–258.

Lundy, D. E. (2016). Decontaminating taste: Minimizing nonaesthetic biases in aesthetic appraisal. *Review of Arts and Humanities, 5*, 1–16.

Lundy, D. E., Allred, G. E., & Peebles, B. L. (2019). How good is this song? Expert versus nonexpert aesthetic appraisal. *Psychology of Aesthetics, Creativity, and the Arts, 13*, 293–304.

Lundy, D. E., Hinners, C. T., Stephens, L. A., & Whitton, J. R. (2018). What it means to be (un)professional: the presence of nonaesthetic bias within differing levels of music and film expertise. *Psychology of Aesthetics, Creativity, and the Arts, 12,* 205–215.

Lundy, D. E., & Smith, J. L. (2017). It's tough to be a critic: Professional versus nonprofessional music judgment. *Empirical Studies of the Arts, 35,* 139–168.

Manturzewska, M. (2011). The reliability of evaluation musical performance by music experts. *Interdisciplinary Studies in Musicology, 10,* 97-109.
McPherson, G. E. (1995). The assessment of musical performance: Development and validation of five new measures. *Psychology of Music, 23,* 142-161.
McPherson, G. E., & Schubert, E. (2004). Measuring performance enhancement in music. In A. Williamon (Ed.), *Musical excellence: Strategies and techniques to enhance performance* (pp. 61-82). Oxford University Press.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.
Parkes, K. A. (2010). Performance assessment: Lessons from performers. *International Journal of Teaching and Learning in Higher Education, 22,* 98-106.
Schleuter, S. L. (1997). *A sound approach to teaching instrumentalists.* Schirmer.
Waddell, G., Perkins, R., & Williamon, A. (2019). The Evaluation Simulator: a new approach to training music performance assessment. *Frontiers in Psychology, 10,* 557.
Winter, N. (1993). Music performance assessment: A study of the effects of training and experience on the criteria used by music examiners. *International Journal of Music Education, 22,* 34-39.