Journal of Arts & Humanities

Volume 15, Issue 03, 2026: 01-05
Article Received: 05-01-2026
Accepted: 10-03-2026
Available Online: 20-03-2026
ISSN: 2167-9045 (Print), 2167-9053 (Online)
DOI: http://dx.doi.org/10.18533/journal.v15i3.2652

When formal criteria fail: Judgment and validity in arts assessment

Eugene Seow[1]

ABSTRACT

Assessment in the arts increasingly relies on formalized criteria and rubrics to ensure transparency and consistency. However, in judgment-based disciplines, this reliance raises persistent questions about validity. This short comment examines the epistemic limits of formal assessment instruments in contexts in which evaluative judgment is central to disciplinary expertise. Drawing on assessment theory and humanities perspectives on judgment, the article argues that formalization can misrecognise expert judgment as error rather than capacity, producing validity failures that cannot be resolved through improved specification or calibration. The analysis reframes validity as an inferential problem rather than a procedural one and clarifies why judgment must be understood as an assessment instrument rather than a methodological weakness. The article contributes a conceptual account of assessment formalism that helps explain recurring tensions in arts evaluation and opens space for more defensible approaches to judgment-based assessment.

Keywords: Assessment validity; judgment; arts assessment; criteria-based assessment.

1. Introduction

Across higher education, assessment practices have increasingly relied on formalized criteria, rubrics, and scoring schemes to promote transparency, consistency, and accountability. These instruments promise to render evaluative decisions explicit, auditable, and defensible, particularly in contexts where assessment outcomes carry institutional or professional consequences. In many disciplines, such formalization is treated as a straightforward improvement, assumed to yield fairer and more valid judgments.

In the arts and other judgment-based fields, however, this assumption has long been uneasy. Evaluation in these domains depends not on the mechanical application of rules but on trained perceptual and interpretive capacities developed through disciplinary practice. Whether assessing performance, studio work, creative artifacts, or perceptual understanding, evaluators routinely exercise forms of judgment that are contextual, comparative, and responsive to qualities that resist exhaustive

[1] Singapore. Email: info@eugeneseowmusic.com

specification. The attempt to translate such judgment into fixed criteria introduces a persistent tension between the epistemic nature of disciplinary knowledge and the administrative demands of assessment systems.

Importantly, this tension has proven resistant to reform. Repeated efforts to improve assessment through more precise criteria, more detailed rubrics, and enhanced calibration processes have not eliminated dissatisfaction in judgment-based disciplines. Instead, concerns about fairness, rigor, and validity tend to recur in modified form. This persistence suggests that the difficulty is not merely technical but structural: it arises from a mismatch between how assessment systems seek to secure validity and how evaluative knowledge is actually produced in disciplines in which judgment is central.

This tension is often framed as a problem of reliability or subjectivity, leading judgment to be treated as error to be constrained through increasingly precise criteria and rubrics (Sadler, 2009). Yet despite decades of refinement, dissatisfaction with arts assessment remains widespread, often resurfacing as concerns about construct validity, misrecognition of expertise, or the inadequacy of formal instruments to capture disciplinary quality. These recurring difficulties suggest that the problem may not lie in insufficient specification, but in a more fundamental misunderstanding of how judgment functions within assessment. These concerns intersect with wider debates in higher-education assessment about rubric-based evaluation, tacit disciplinary knowledge, and the role of expert judgment in evaluation.

This short comment argues that in judgment-based arts disciplines, increasing reliance on formalized assessment criteria produces a distinctive form of validity failure. By treating expert judgment as a methodological weakness rather than as an epistemic capacity, formal assessment instruments misrepresent what they seek to measure. The result is not simply imperfect evaluation, but a structural distortion in which assessment practices undermine their own validity claims. Drawing on assessment theory and humanities-informed accounts of judgment, the article reframes validity as an inferential problem rather than a procedural one. It clarifies why refinement of criteria alone cannot resolve the tensions at stake. In this article, *assessment formalism* refers to the assumption that increasing the explicitness and procedural specification of assessment criteria necessarily improves fairness, reliability, and validity.

2. Judgment and assessment in the Arts

Assessment in the arts depends fundamentally on judgment. Unlike domains in which performance can be exhaustively captured through discrete indicators or algorithmic scoring, arts evaluation requires the interpretation of complex, relational qualities whose significance emerges only in context. Such judgment is a learned capacity, developed through sustained disciplinary engagement, exposure to exemplars, and participation in shared evaluative cultures. Experienced assessors do not simply apply preferences; they recognize patterns, discern coherence, and evaluate adequacy relative to disciplinary norms that are themselves historically and socially constituted.

The development of such judgment is not incidental. It is cultivated through prolonged exposure to disciplinary exemplars, participation in evaluative dialogue, and repeated engagement with borderline or ambiguous cases. Through these processes, assessors learn not only what counts as quality, but how qualities relate, conflict, and trade off in particular contexts. Judgment, in this sense, is not the absence of standards, but the capacity to work with standards that cannot be exhaustively specified in advance.

Assessment scholarship has repeatedly emphasized the distinction between judgment and preference. Judgment involves comparison, calibration, and justification within a community of practice, whereas preference reflects individual taste or idiosyncratic response. Conflating the two collapses a crucial epistemic difference. When judgment is reduced to preference, the exercise of expertise appears illegitimate, and calls for objectivity are answered with demands for tighter criteria (Sadler, 2009). Yet this response overlooks the extent to which criteria themselves depend on judgment for their interpretation and application.

In the arts, judgment functions not as a residual category left over after criteria have done their work, but as the primary means through which evaluative meaning is produced. Assessors attend to timing, balance, coherence, responsiveness, and expressive direction, qualities that cannot be fully specified without losing the very relationships that make them intelligible. Attempts to externalize these judgments into lists of descriptors often result in abstraction and flattening, transforming dynamic

perceptual assessments into static representations. What is gained in apparent clarity is frequently lost in epistemic adequacy.

Within assessment systems, this narrowing often proceeds unnoticed. Criteria are treated as neutral tools rather than as epistemic commitments, and their limitations are attributed to implementation rather than to design. As a result, dissatisfaction with judgment-based assessment is routinely addressed by further refining criteria, adding descriptors, or developing more elaborate rubrics, responses that leave the underlying epistemic tension intact (Sadler, 2009). Recognizing judgment as an epistemic resource rather than a contaminant is, therefore, a prerequisite for addressing validity in arts assessment.

3. The limits of formal criteria

Formal assessment criteria are often introduced as safeguards against subjectivity. By specifying what counts and how it should be evaluated, criteria promise to stabilize judgment, reduce inconsistency, and make evaluative decisions transparent. In judgment-based arts disciplines, however, this promise conceals a deeper problem: criteria do not replace judgment; they reconfigure it. When formalized criteria are treated as substitutes for expertise rather than as aids to it, they generate a distinctive form of validity failure.

Assessment criteria are representations. They abstract selected features of disciplinary practice into codifiable, shareable, and auditable language. This representational move is not inherently flawed; criteria can support communication and calibration when used with an awareness of their limits. Problems arise when representations are mistaken for the phenomena they describe. In arts assessment, the qualities that matter most (coherence, balance, responsiveness, and expressive direction) are relational and emergent. They acquire meaning through context and comparison rather than through isolated descriptors. No matter how detailed, criteria remain partial translations of these qualities, not their equivalents (Sadler, 2009).

As criteria proliferate, the gap between representation and practice often widens. Increased specificity is assumed to reduce ambiguity, yet it usually narrows the construct under study. Qualities that resist precise description are excluded or marginalized, while those that can be articulated in advance are privileged. Over time, assessment shifts from evaluating disciplinary judgment to evaluating alignment with its representations. Disagreement between expert assessors and rubric outputs is interpreted as inconsistency or bias, prompting further refinement of criteria that intensifies the very problem it seeks to solve (Sadler, 2009).

In practice, this dynamic is often experienced as a conflict between professional judgment and formal assessment outputs. Assessors may recognize qualities of coherence or responsiveness that are difficult to reconcile with rubric descriptors, while rubric scores may suggest deficiencies that expert judgment does not support. Rather than prompting reflection on the limits of representation, such conflicts are frequently interpreted as assessor inconsistency. The result is a gradual displacement of judgment toward procedural compliance, even as assessment outcomes become less aligned with disciplinary understanding.

By misrecognizing judgment as error, formal assessment instruments invert the epistemic logic of arts evaluation. What assessors know how to do (discern quality through trained comparison and contextual understanding) is rendered suspect, while what can be specified in advance is elevated as evidence. The resulting assessments may appear transparent and consistent, yet they systematically misrepresent the expertise they claim to evaluate. This misrepresentation is not incidental; it is the predictable outcome of treating formalization as a solution to an epistemic problem it cannot resolve.

This displacement has cumulative effects. As assessment practices increasingly reward alignment with formal criteria, they subtly reshape what is recognized as competence. Qualities that are difficult to specify are treated as marginal or discretionary, while those that can be articulated in advance gain disproportionate weight. Over time, assessment comes to reflect the logic of its instruments rather than the epistemic priorities of the discipline it claims to evaluate.

4. Validity without formalism

Validity in assessment is often treated as a property of instruments. Well-designed criteria, clearly articulated descriptors, and reliable scoring procedures are assumed to secure the legitimacy of evaluative outcomes. In judgment-based arts assessment, this assumption is particularly persistent, as concerns about subjectivity are routinely addressed through further formalization. Yet assessment theory offers a different starting point. Validity does not reside in instruments or procedures themselves, but in the inferences drawn from assessment evidence in relation to stated constructs (Messick, 1995).

From this perspective, the central question is not whether criteria are explicit, but whether assessment practices support warranted interpretations of performance or work. In arts contexts, assessors must determine which features are salient, how they relate to one another, and what they signify in context. These determinations are inferential acts, grounded in disciplinary knowledge rather than mechanical application. Treating validity as an inferential problem foregrounds the role of judgment rather than attempting to eliminate it (Messick, 1995).

Formal criteria do not remove the need for inference; they structure it. By foregrounding particular descriptors, criteria signal which features should be attended to and which may be ignored. This signaling effect shapes the inferential pathway through which assessment claims are constructed. When criteria align closely with the underlying construct, they can support valid inference. When they do not, they introduce systematic distortion. Consistent application of inadequate criteria produces consistent misrepresentation rather than a valid assessment (Messick, 1995).

Understanding judgment as an assessment instrument rather than as a contaminant alters the terms of the debate. Judgment is not an obstacle to validity; it is the means by which evidence becomes meaningful. The question is therefore not how to constrain judgment, but how to recognize its epistemic role and its limits. Criteria can support evaluation, but they cannot substitute for the interpretive work through which meaning is made.

5. Implications for arts assessment

Seen in this light, disputes about assessment in the arts are less about the presence or absence of criteria than about how evaluative authority is conceptualized. When validity is secured through formal representation alone, judgment appears to threaten fairness. When validity is understood inferentially, judgment becomes the means through which assessment claims are warranted. In practical terms, this implies that disagreement between expert assessors and rubric outcomes should not automatically be interpreted as inconsistency or bias. Instead, such disagreement may signal a misalignment between the formal assessment instrument and the disciplinary construct it seeks to represent. This distinction helps explain why technical adjustments to assessment instruments often fail to address deeper concerns about legitimacy and rigor.

The analysis developed in this comment has implications for how assessment in the arts is understood, debated, and reformed. These implications are conceptual rather than prescriptive. They do not point toward a particular assessment model or technique, but toward a reframing of the problem underlying many recurring disputes about fairness, rigor, and objectivity in arts evaluation.

First, the persistence of dissatisfaction with criteria-based assessment can be understood as a symptom of epistemic mismatch rather than implementation failure. When formal criteria are applied in judgment-based contexts, tensions are often attributed to insufficient training, inconsistent application, or inadequate calibration. While such factors may play a role, the analysis here suggests that deeper problems arise when assessment instruments are misaligned with the nature of the constructs they seek to measure.

Second, reframing judgment as an epistemic resource clarifies why debates about subjectivity in arts assessment are so often unproductive. Treating judgment as bias to be eliminated invites procedural solutions that obscure rather than address the inferential work involved in evaluation. By contrast, understanding judgment as a trained capacity foregrounds questions of disciplinary expertise, shared standards, and evaluative reasoning. This shift does not remove disagreement, but it renders disagreement intelligible as part of legitimate evaluative practice rather than as evidence of methodological failure.

Third, this reframing challenges assumptions about transparency. Formal criteria are commonly justified because they make assessment decisions visible and accountable. However, transparency achieved through representation may come at the expense of obscuring the interpretive judgments that underlie evaluation. A focus on inferential validity suggests that transparency should be understood not only as access to criteria, but as clarity about how judgments are formed and warranted.

6. Conclusion

This short comment has argued that increasing reliance on formalized assessment criteria in judgment-based arts disciplines produces a distinctive form of validity failure. By treating expert judgment as a methodological weakness rather than as an epistemic capacity, formal assessment instruments misrepresent what they claim to measure. The resulting assessments may appear transparent and defensible, yet they systematically distort disciplinary expertise by privileging what can be specified in advance over what must be discerned in context.

Reframing validity as an inferential problem clarifies why refinement of criteria alone cannot resolve the tensions at stake. Criteria can support evaluation, but they cannot substitute for the interpretive work through which meaning is made. When this distinction is overlooked, formalization becomes a source of mismeasurement rather than a safeguard against it.

The contribution of this comment is conceptual. It does not propose an assessment framework, pedagogical intervention, or institutional policy. Instead, it offers an account of assessment formalism that helps explain why dissatisfaction with arts evaluation persists despite repeated reform efforts. Recognizing judgment as central rather than residual does not eliminate the challenges of arts assessment, but it alters the terms on which those challenges are addressed. Although developed here in relation to the arts, similar tensions arise in other judgment-dependent fields such as architecture, design, literary studies, and clinical evaluation. Rather than seeking certainty through ever-greater specification, assessment practices may need to reckon more explicitly with the inferential conditions under which evaluative claims are warranted.

References

Sadler, D. R. (2009). *Indeterminacy in the use of preset criteria for assessment and grading*. Assessment & Evaluation in Higher Education, 34(2), 159–179. https://doi.org/10.1080/02602930801956059

Messick, S. (1995). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as a scientific inquiry into score meaning*. American Psychologist, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Kane, M. T. (2006). *Validation*. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 17–64). Praeger.