

Commentary

Beyond Measurable Performance: Reconsidering the Distinctive Nature of Music Education Assessment in the Age of Artificial Intelligence

Bolim Lee¹

1. LBL Institute of Culture & Arts Education and Engineering, South Korea

Assessment in music education has never been entirely equivalent to assessment in subjects in which knowledge can be more readily separated into discrete and reproducible answers. A musical performance may be technically accurate yet musically unconvincing; a composition may depart from an expected form yet demonstrate meaningful artistic intention; and two markedly different interpretations may both constitute educationally valuable musical responses. The distinctive nature of music therefore requires assessment to attend not only to correctness or measurable achievement but also to interpretation, contextual understanding, artistic judgement, and the processes through which musical decisions are made. The rapid introduction of artificial intelligence (AI) into music education makes this longstanding issue newly urgent. AI can increasingly generate musical products, analyse performance data, produce feedback, and participate in assessment. These capabilities may improve the efficiency and range of assessment, but they also expose a fundamental question: if an apparently successful musical product can be generated or substantially shaped by AI, what exactly should be assessed in the human learner?

This conceptual paper argues that AI does not eliminate the distinctive character of music assessment; rather, it makes the recognition of that character more necessary. It proposes moving away from excessive dependence on observable products towards closer attention to the learner's artistic discernment: the capacity to distinguish, select, interpret, justify, revise, reject, and take responsibility for musical decisions. This does not mean replacing musical performance or creative products with an abstract assessment of thought. Instead, the paper argues for reconnecting musical products with the human processes of judgement through which they acquire educational meaning. From this perspective, the central task of music assessment in the AI era is not simply to determine how

accurately AI can assess music, but to reconsider what remains educationally meaningful to assess in the person who learns music.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

1. Introduction

Assessment is necessary in almost every educational domain, but what is assessed and how it is interpreted cannot be entirely separated from the nature of the subject itself. This is particularly important in music education. Music involves knowledge and skill, but musical learning is not exhausted by either. Performing, composing, improvising, listening, and responding require learners to make decisions within particular musical, aesthetic, cultural, and educational contexts. Consequently, the quality of musical learning cannot always be reduced to the number of correct responses, the absence of technical errors, or the successful reproduction of predetermined criteria.

This is not a new problem. Scholarship in music education assessment has long addressed the tension between reliability and the preservation of qualities that make musical learning distinctive. Swanwick^[1], for example, questioned assessment approaches in which curriculum descriptors, complexity, and range could obscure the more fundamental question of musical quality. Brophy^[2] similarly argues that well-designed assessment of music learning must take account of validity, purpose, alignment, authenticity, and the particular contexts in which music learning occurs. The problem therefore precedes artificial intelligence. AI has not created the difficulty of assessing musical quality; it has made an existing difficulty considerably harder to ignore.

Contemporary AI systems can analyse musical information, generate compositions, support lesson planning, provide formative feedback, and increasingly participate in assessment processes. Recent studies have examined AI-supported music assessment from several directions. Generative AI has been used to develop and score written music-assessment items, with teacher judgement remaining important to the validity of the resulting assessment^[3]. Multimodal systems have been designed to combine audio, video, symbolic notation, and textual information in music teaching evaluation^[4], while other research has examined relationships among AI-assisted formative assessment, feedback literacy, and engagement in music learning^[5]. These developments are educationally significant and should not simply be resisted.

Yet a prior question remains. Before asking how effectively AI can evaluate music learning, music educators need to ask what music education intends to recognise as evidence of learning in the first place. This paper therefore begins from the premise that assessment within the music subject requires distinctive considerations because the object of assessment is itself multidimensional. It then considers how AI unsettles conventional relationships among musical process, product, authorship, and assessment. The central argument is not that musical performance should cease to be assessed, nor that education should retreat from technology. Rather, AI makes it increasingly necessary to identify the human musical judgement embodied within, behind, and sometimes in resistance to the resulting product. The concept of artistic discernment is used here to describe this evaluative emphasis.

2. The Distinctive Character of Assessment in Music Education

A difficulty arises when assessment practices organised primarily around measurable achievement are transferred to music without sufficient consideration of what musical achievement entails. Technical accuracy is unquestionably part of musical competence. Rhythm can be inaccurate, pitch can be unstable, and a learner can misunderstand musical terminology or fail to demonstrate an intended skill. Such dimensions can and should be assessed when they correspond to the purpose of learning. The difficulty begins when these dimensions become proxies for the whole of musical achievement simply because they are comparatively easy to specify and measure.

Accuracy alone cannot account for the quality of a musical response. Two students may perform the same work without substantial technical errors and nevertheless produce musically different results. Differences in phrasing, articulation, timing, balance, tone, dynamic direction, and expressive intention may alter the meaning of the performance without producing a simple distinction between “correct” and “incorrect.” Similarly, a composition that satisfies every formal requirement of an assignment may reveal little evidence of musical imagination, while a less conventional product may demonstrate coherent and purposeful musical thinking. Assessment in music therefore involves a difficulty that is not adequately represented by the accumulation of objectively measurable components.

This issue is connected with the nature of music as human activity rather than merely as an object produced at the end of instruction. Elliott and Silverman^[6] conceptualise music and musicianship through multidimensional musical practices, emphasising forms of musical knowing that are enacted through performing, listening, composing, improvising, arranging, and reflective musical action. Such a position has direct implications for assessment. If musical understanding is enacted through these

practices, assessment needs to remain sensitive not only to what has been produced but also to the quality of the musical activity and understanding embodied in that production.

The long-standing problem of standardisation becomes particularly acute at this point. Assessment requires criteria, and educational decisions require sufficient consistency to be defensible. Yet criteria can become reductive when what can most easily be specified gradually becomes what is most strongly valued. In such a situation, the assessable begins to define the educationally valuable rather than the educationally valuable determining what ought to be assessed.

Eisner's^[7] discussion of educational connoisseurship and criticism is useful in understanding this tension. His work challenged the assumption that educational quality can always be represented adequately through predetermined quantitative indicators. Discernment involves perceiving qualities, relationships, differences, and significance that require informed judgement. This does not imply that such judgement should be arbitrary or merely intuitive. On the contrary, informed qualitative judgement requires knowledge, attention, experience, and the capacity to make an evaluation intelligible to others.

Music education has particular need for this distinction. Assessment can be systematic without pretending that every meaningful musical quality is reducible to a numerical variable, just as professional judgement does not need to become impressionistic judgement without criteria. A basic principle can therefore be established before AI is considered at all: music assessment must preserve a defensible relationship between what is measured and what is musically and educationally meaningful.

3. What AI Changes—and What It Does Not

AI changes the conditions under which this principle now operates. The capacity of AI to participate in educational assessment is expanding rapidly, and music education is no exception. Recent research has examined AI-supported item development and scoring^[3], multimodal evaluation systems^[4], AI-mediated rubric and feedback practices^[8], AI-assisted formative assessment^[5], and assessment rubrics designed specifically for generative-AI-based music creation^[9]. Taken together, these developments suggest that functions once regarded as inseparable from human evaluation may increasingly be shared with computational systems.

There are legitimate advantages. AI may provide rapid feedback, process large quantities of information, identify recurrent patterns, and support differentiated learning. In written music assessment, You^[3] found that generative AI could support both item development and scoring, while also showing

that incorporating teacher scoring information increased agreement between AI and human evaluation. In generative-AI-based music creation, Lee^[9] developed a process-oriented assessment rubric that explicitly takes account of interaction, modification, reflection, and human-AI collaboration. Such studies indicate that the important educational question is no longer simply whether AI can enter assessment. It already has.

Greater technical capacity, however, does not resolve the educational question of what deserves assessment. In some respects, generative AI makes that question more difficult. A student may now submit a musical or written product that has been partially or substantially shaped by an AI system. Shaw^[10] identifies changes in the relationship between student work, assessment, and grading as a central implication of AI for music educators. The difficulty cannot be reduced to academic dishonesty. Even when AI use is explicitly permitted, the relationship between the quality of a product and the musical competence of the learner becomes less transparent.

Music education has long used observable products—performances, compositions, arrangements, analyses, and other artefacts—as evidence of learning. Such products remain important and should remain assessable. Yet a polished product can no longer automatically function as sufficient evidence of the human process through which it was produced. When sophisticated external

systems can contribute materially to composition, revision, interpretation, explanation, or feedback, additional evidence may be necessary to determine where learning is actually located.

AI therefore does not create the fundamental assessment problem in music education. Instead, it exposes an assumption that was always open to question: that complex musical understanding can be inferred with sufficient confidence from the quality of an observable product. The introduction of AI makes the distance between product quality and learner understanding potentially greater, and this distance requires closer educational attention.

4. From Product to Judgement

One possible response is to prohibit AI in assessment and attempt to restore a more direct relationship between student and product. In certain situations, restrictions on AI will remain both reasonable and necessary. Nevertheless, prohibition alone is unlikely to constitute a sufficient educational response. AI is becoming part of the environment in which learners search, listen, practise, create, revise, compare, and communicate. UNESCO's guidance on generative AI similarly argues for a human-centred approach in

which technological development remains subordinate to human agency and educational purpose^[11]. The educational task, then, is not simply to decide whether AI should be present, but to determine what its presence changes about the interpretation of evidence.

For music assessment, this requires reconsidering the evidential relationship between musical product and musical learning. The product should not disappear. Music is heard, performed, made, revised, embodied, and shared, and an educational approach that assessed only verbal explanations of musical thinking would be as reductive as one that assessed only technical accuracy. What needs to change is not the existence of the product in assessment but the assumption that the product can always speak adequately for the learner.

The evaluator may therefore need to ask what the learner noticed, which alternatives were considered, and why one musical possibility was selected rather than another. Where AI has been used, the evaluator might also ask which suggestions were accepted, which were rejected, what was revised after further listening, and whether the learner can recognise a technically convincing but musically inappropriate result. A learner's ability to explain why a musical decision works in one context but not another may reveal something that the final product alone cannot show.

These questions shift attention from possession of a successful artefact towards evidence of musical judgement. This becomes increasingly important because AI can itself generate plausible musical alternatives. Producing an option is therefore not necessarily the most demanding part of a creative task. In some circumstances, the educational work lies in distinguishing among alternatives, perceiving inadequacy within superficially convincing results, making purposeful selections, revising them, and taking responsibility for the musical consequences of those decisions.

The distinction is subtle but important. Process should not be valued simply because it is process, nor should every recorded decision automatically become evidence of learning. What matters is the quality of the learner's judgement within that process. This is where artistic discernment becomes particularly relevant to assessment in the AI era.

5. Artistic Discernment as an Assessment Concern

Artistic discernment, as used in this paper, does not refer simply to personal taste, preference, or an assumed ability to identify superior art. Nor is the term presented here as an entirely new educational theory. It is used more cautiously to describe a capacity that becomes particularly visible under AI-

mediated conditions: the learner's ability to distinguish, select, interpret, justify, revise, reject, and take responsibility for musical choices in relation to musical, aesthetic, contextual, and educational purposes.

The concept has affinities with Eisner's^[7] discussion of connoisseurship, especially the capacity to perceive and differentiate qualities that might escape less informed attention. Its present use, however, is located specifically within contemporary music learning environments in which students may encounter machine-generated alternatives, algorithmic recommendations, and automated feedback. Artistic discernment in this context is therefore not simply a cultivated sensitivity towards musical quality. It includes the learner's capacity to exercise

that sensitivity when musical possibilities are externally generated and when the apparent quality of an output does not necessarily correspond to the learner's own level of musical understanding.

The distinction between generating an answer and judging an answer becomes important here. A learner may ask a generative system to produce a chord progression, melodic continuation, arrangement, programme note, interpretation, or evaluative commentary. The generated response, however sophisticated, provides little evidence by itself of the learner's musical understanding. Educationally informative evidence begins to emerge in what happens afterwards. Does the learner accept the first output without question, or recognise a stylistic inconsistency? Can the learner perceive an expressive weakness that is not technically erroneous? Can the material be altered for a particular performer, audience, cultural context, or musical intention? Can the learner disagree with automated feedback and provide a musically defensible reason?

These actions cannot be located simply in the final artefact. They concern the relationship between the learner and the musical possibilities available to them. In this sense, the question of agency becomes closely connected with assessment. Peng et al.^[12], in their systematic review of intelligent assessment and generative AI in music education, identify a tendency for existing research to emphasise performance outcomes while paying less attention to learners' cognitive agency and possible forms of dependence. This is consequential for assessment because improvement in an AI-supported product does not necessarily demonstrate corresponding development in the learner's capacity to make independent musical judgements.

This argument should not be misunderstood as an assumption of human superiority. AI systems may identify certain technical features more consistently or rapidly than human evaluators, and human judgement is itself fallible. Nor should a decision be regarded as educationally valuable simply because a

human made it. What matters is whether the learner can exercise informed musical judgement and provide evidence of that judgement. In some cases, the most revealing evidence of learning may lie precisely in the distance between what AI offers and what the learner ultimately chooses.

6. Implications for Music Education Assessment

Reconsidering music assessment through artistic discernment does not require the immediate construction of another universal rubric. Indeed, imposing a single instrument across highly different musical practices could reproduce the very problem this paper seeks to address. Music assessment needs sufficient structure to be credible, but it also needs sufficient responsiveness to recognise differences among performance, composition, improvisation, listening, creative work, and the varied contexts in which these activities occur.

One implication is that assessment may need to make aspects of decision-making more visible. Short oral explanations, annotations, successive versions of a composition, reflective accounts of revision, comparisons between alternative interpretations, or selected records of interaction with AI may provide useful evidence. The aim should not be to increase documentation for its own sake or to turn every creative action into an administrative record. Evidence is worthwhile only where it enables the teacher to understand dimensions of musical learning that the finished product cannot adequately reveal.

A second implication is that AI use itself should not become either a positive or negative assessment criterion. A learner who makes extensive use of AI may demonstrate sophisticated musical judgement, while another may simply accept generated material with little reflection. Conversely, a student who avoids AI has not thereby demonstrated deeper musicianship. The more relevant question concerns what the learner actually does musically and intellectually with the resources available. Recent work on generative-AI-based music creation similarly suggests the value of examining planning, generation, modification, and reflection rather than treating the resulting artefact as the sole object of assessment^[9].

Disagreement may also become meaningful evidence. If an AI system proposes an evaluation and the learner rejects it, the disagreement can reveal learning when the learner offers a musically defensible justification. Critical refusal should therefore not automatically be interpreted as unsuccessful use of technology. In some circumstances, the capacity not to follow an apparently plausible recommendation may demonstrate stronger musical understanding than compliant use.

Teachers' professional judgement consequently becomes more, rather than less, important. Research on AI-supported written music assessment has already indicated the significance of teacher review and human scoring information^[3], while reflective work on GPT-mediated assessment likewise emphasises the preservation of human judgement, student authorship, and assessment understanding^[8]. Automated assessment should therefore not be understood simply as a gradual transfer of evaluation from teacher to machine. It may instead require stronger assessment literacy so that teachers can decide which dimensions can appropriately be supported by automation and which require contextual, musical, and pedagogical interpretation.

A final implication concerns the distinction between assistance and achievement. AI may help a learner produce a stronger musical outcome, and such assistance may itself contribute to learning. Yet assessment still needs to ask what has changed in the learner. If a quality observed in an AI-supported product disappears completely when the support is removed, the educator has reason to consider whether the achievement belongs primarily to the learner, the tool, their interaction, or some combination of the three. This is not an entirely new educational question: teachers, peers, recordings, notation software, instruments, and other resources have always mediated musical learning. AI intensifies it because its contribution can extend beyond support towards the generation and evaluation of substantive musical content.

7. Discussion: What Should Remain Human in Assessment?

The future of music assessment could easily be framed as a competition between human and artificial intelligence: Which evaluates more accurately? Which provides more reliable feedback? Which recognises musical quality more consistently? These are legitimate empirical questions, and developments in multimodal AI suggest that the technological capacity to analyse increasingly complex forms of music teaching and learning will continue to expand^[4]. They are not, however, sufficient educational questions.

Music education is concerned not only with obtaining better musical products but with developing people who are increasingly capable of engaging with music.

This distinction matters because it is possible, at least theoretically, for an educational environment to produce increasingly polished outcomes while learners become less capable of explaining, evaluating, revising, or independently directing the musical decisions embodied in those outcomes. Recent findings concerning AI-assisted formative assessment indicate that learners' capacity to understand and use

feedback remains consequential even when feedback itself is technologically enhanced^[5]. Similarly, concerns regarding cognitive agency and dependence suggest that better technological support cannot automatically be equated with stronger human learning^[12].

The concern, therefore, is not that AI might become too musical. A more important concern is that education might gradually become less interested in whether the learner is becoming musically discerning. If assessment increasingly privileges what can be automated, counted, or generated, this could reverse the relationship between educational purpose and measurement. Music education might begin with the capacities of the available technology and subsequently redefine musical achievement around what those capacities can recognise.

The sequence should be the opposite. Music education should begin with a defensible conception of musical learning and then decide what evidence can meaningfully represent that learning, where human judgement is necessary, and where AI can contribute. The distinctive nature of music assessment is therefore not an obstacle to technological innovation. It is a condition for determining whether technological innovation remains educational.

Seen in this way, artistic discernment may become more visible as an educational achievement precisely because production is becoming easier to outsource. To perceive a difference, make a choice, provide a reason, revise after listening, resist an attractive but inappropriate suggestion, and accept responsibility for a musical decision are not peripheral activities surrounding music making. They have always been part of musical learning. AI makes their educational significance more difficult to overlook.

8. Conclusion

Artificial intelligence does not create the need for distinctive assessment in music education. That need already exists because musical learning contains dimensions of technical competence, interpretation, experience, context, judgement, and artistic decision-making that cannot always be represented adequately through uniform measures. The history of music assessment has already demonstrated the difficulty of balancing reliability and specification with musical quality and professional judgement^{[2][1]}. AI places this unresolved tension in a new environment.

When machines can contribute to the creation, revision, analysis, and evaluation of musical products, the quality of the final product becomes a less transparent indicator of what a learner knows and can do. The educational response should not be to abandon products, performance, or technology. Nor should process

automatically replace product as the new object of assessment. Rather, music education needs to examine more carefully the human musical judgement through which products are selected, shaped, revised, understood, and ultimately claimed as evidence of learning.

Artistic discernment offers one way of naming this concern. It directs attention towards the learner's capacity to distinguish, select, interpret, justify, revise, reject, and take responsibility for musical choices. These capacities have always mattered in music. AI simply makes it more difficult to mistake the quality of a product for the whole of musical learning.

The central question for music education assessment in the AI era, therefore, may not be whether artificial intelligence can evaluate music as effectively as a human evaluator. A more consequential question precedes it: What qualities of the human learner does music education consider sufficiently important to notice, cultivate, and assess? The answer to that question should determine the role of AI in assessment, rather than allowing the capacities of AI to determine what music education comes to value.

References

1. ^a ^bSwanwick K (1997). "Assessing Musical Quality in the National Curriculum." *Brit J Music Ed.* 14(3):205–215. doi:[10.1017/s0265051700001200](https://doi.org/10.1017/s0265051700001200).
2. ^a ^bBrophy TS (2019). "Assessment in Music Education." In: Brophy TS, editor. *The Oxford Handbook of Assessment Policy and Practice in Music Education*. Vol. 2. Oxford: Oxford University Press. pp. 902–932. doi:[10.1093/oxfordhb/9780190248130.013.75](https://doi.org/10.1093/oxfordhb/9780190248130.013.75).
3. ^a ^b ^c ^dYou H (2026). "Exploring AI-Supported Restricted and Extended Essay Assessment in Music Education: The Potential of ChatGPT for Item Development and Scoring." *Korean Music Educ Soc [Korean Journal of Research in Music Education]*. 55(1):231–255. doi:[10.30775/kmes.55.1.231](https://doi.org/10.30775/kmes.55.1.231).
4. ^a ^b ^cWang Z (2026). "Music Teaching Evaluation Using Multimodal Deep Reinforcement Learning." *Discov Artif Intell.* 6:Article 151. doi:[10.1007/s44163-026-00832-0](https://doi.org/10.1007/s44163-026-00832-0).
5. ^a ^b ^cSun H (2026). "The Dynamic Relationships Among AI-Assisted Formative Assessment in Music Teaching, Feedback Literacy, and Music Learning Engagement." *Front Psychol.* 17:1919875. doi:[10.3389/fpsyg.2026.1919875](https://doi.org/10.3389/fpsyg.2026.1919875).
6. ^aElliott DJ, Silverman M (2015). *Music Matters: A Philosophy of Music Education*. 2nd ed. Oxford: Oxford University Press.

7. ^a ^bEisner EW (2002). *The Educational Imagination: On the Design and Evaluation of School Programs*. 3rd ed. Upper Saddle River: Prentice Hall.
8. ^a ^bImam Koning S, Mazlan CAN (2026). "AI as Rubric Mediator: A Reflective Practice Brief on GPT-Supported Assessment for Learning in Music Education." *Asian J Assess Teach Learn*. 16(1):68–79. doi:[10.37134/ajatel.vol16.1.6.2026](https://doi.org/10.37134/ajatel.vol16.1.6.2026).
9. ^a ^b ^cLee KM (2025). "Development of an Assessment Rubric for Generative AI-Based Music Creation Projects." *Korean J Arts Educ [Korean Journal of Arts Education]*. 23(4):451–474. doi:[10.23317/kaae.2025.23.4.024](https://doi.org/10.23317/kaae.2025.23.4.024).
10. ^aShaw BP (2024). "Artificial Intelligence and Assessment: Three Implications for Music Educators." *Music Educators J*. 111(2):19–25. doi:[10.1177/00274321241296118](https://doi.org/10.1177/00274321241296118).
11. ^aMiao F, Holmes W (2023). "Guidance for Generative AI in Education and Research." UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>.
12. ^a ^bPeng X, Sun K, Shan X, Zhang J (2026). "Empowerment or Dependency? A Systematic Review of the Impacts of Intelligent Assessment and Generative AI on Learners' Self-Beliefs and Cognitive Agency in Music Education." *Front Psychol*. 17:1776445. doi:[10.3389/fpsyg.2026.1776445](https://doi.org/10.3389/fpsyg.2026.1776445).

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.