https://doi.org/10.30775/KMES.55.3.425
음악교육연구 제55권 제3호 pp. 425-461 (2026년 8월) ISSN 1229-4179 (Print)
Korean Journal of Research in Music Education Vol.55, No.3, (August, 2026) ISSN 2713-3788 (Online)

전문 음악교육에서의 AI 기반 음악 연주 평가: 평가 역량 관점의 비판적 분석

AI-Based Music Performance Assessment in Professional Music Education: A Critical Analysis from an Assessment Literacy Perspective

자와시 헤리 리잘* · 삼푸르노 무하마드 바유 테조** · 사피안 압둘 라흐만***
Herry Rizal Djahwasi · Muchammad Bayu Tejo Sampurno · Abdul Rahman Safian

Abstract This study critically examines the adequacy of AI-based music performance assessment tools from an assessment literacy perspective and proposes a framework for their use in professional music education. Using Critical Interpretive Synthesis, 52 sources were analyzed through six dimensions of the Assessment Literacy Audit Protocol (ALAP): construct validity, reliability, fairness, formative utility, interpretive scope, and transparency. The findings indicate that AI tools are effective in consistently assessing technical elements such as pitch and rhythm and providing formative feedback for beginning learners, but remain limited in evaluating expressivity, interpretive individuality, and stylistic appropriateness. The study suggests integrating the bounded capabilities of AI assessment with human expert judgment.

Key words: assessment literacy, AI-based music performance assessment, music performance assessment, conservatoire education, assessment governance

초록 본 연구는 인공지능 기반 음악 연주 평가 도구의 평가적 적합성을 평가 역량 관점에서 비판적으로 검토하고, 전문 음악교육에서의 활용을 위한 이론적 · 방법론적 틀을 제안하는 데 목적이 있다. 이를 위해 비판적 해석 종합을 적용하여 음악 연주평가, AI 음악평가, 평가 역량 및 관련 철학 분야의 문헌 52편을 분석하였다. 분석 도구로는 해석학 · 미학, 후기현상학 · 비판적 기술론, 교육측정의 세 철학적 기반에 근거하여 구성타당도, 신뢰도, 공정성, 형성적 유용성, 해석 범위, 투명성의 여섯 차원으로 구성된 평가 역량 감사 프로토콜(ALAP)을 개발 · 적용하였다. 분석 결과, AI 연주 평가 도구는 음정과 리듬 등 기술적 요소의 측정 일관성과 초기 단계 학습자의 형성적 피드백에는 유용성을 보였으나, 표현성, 해석적 개별성, 양식적 적절성 등 전문 음악교육의 연주평가에서 중시되는 핵심 구성개념을 포괄하는 데에는 한계가 있는 것으로 나타났다. 이에 AI 평가 도구의 제한된 평가 기능을 명확히 인식하고 인간 전문가의 해석적 판단과 결합한 평가 체계를 구축할 필요가 있음을 제안하였다.

주제어: 평가 리터러시, 인공지능 기반 음악 연주 평가, 음악 연주 평가, 음악전문교육, 평가 거버넌스

* First author & Corresponding author, E-mail: herry@fmsp.upsi.edu.my
Senior Lecturer, Sultan Idris Education University (UPSI), Faculty of Music and Performing Arts, Tanjong Malim, Perak, Malaysia
** Co-author, E-mail: tejo@fmsp.upsi.edu.my
Senior Lecturer, Sultan Idris Education University (UPSI), Faculty of Music and Performing Arts, Tanjong Malim, Perak, Malaysia
*** Co-author, E-mail: rahmansafian@fmsp.upsi.edu.my
Senior Lecturer, Sultan Idris Education University (UPSI), Faculty of Music and Performing Arts,, Tanjong Malim, Perak, Malaysia

Received: 31 May 2026, Reviewed (Revised): 30 June 2026 (12 August 2026), Accepted: 21 August 2026
© 2026 Korean Music Education Society.

Ⅰ. Introduction

Although performance evaluation is only one component of instrumental music teaching, it remains central to assessing musical competence. It provides evidence of students' technical and artistic development and indicates how effectively the learning program supports that development (Eremenko et al., 2020; Wesolowski & Wind, 2019). Conservatoire performance assessment has long occupied contested epistemological terrain (Djahwasi et al., 2023; Ford, 2010), one in which the authority of the expert teacher (Walton, 2020), the interpretive latitude of the performer (Cook, 2013), and the institutional legitimacy of the examiners are entangled in ways that resist reduction to a single evaluative framework (Chmurzyńska, 2023). Within this contentious landscape, artificial intelligence performance assessment tools have been introduced with considerable confidence, asserting benefits such as objectivity, scalability, and reduced inter-rater variance, as if these were inherently desirable outcomes (Evin, 2024; Ou et al., 2025; Popham, 2018) rather than on theoretically loaded assumptions about the purpose of music performance assessment.

In conservatory contexts, music performance must be understood not only as an artistic practice but also as an assessment-informed practice. This makes assessment literacy central. In this sense, assessment literacy entails the capacity to design, interpret, and apply assessment instruments that are epistemologically grounded, pedagogically coherent, and ethically defensible. The central problem this article addresses is whether what AI measures constitutes a valid, reliable, fair, and educationally productive assessment (Popham, 2018; Stiggins, 1991).

Previous studies have examined the technical capabilities of AI tools (Djahwasi et al., 2025; Phanichraksaphong & Tsai, 2023) and educators' perceptions of AI-assisted feedback (Caruso, 2024; Li & Wang, 2024; Li et al., 2025). Nevertheless, none of these studies comprehensively evaluate assessment instruments from the perspective of assessment literacy. Additionally, they do not establish the philosophical underpinnings required to substantiate musical interpretation as a legitimate subject of assessment.

Assessment literacy scholarship offers criteria for evaluating assessment tools, but these are rarely applied rigorously to AI music tools. This creates a gap where tools are adopted based on efficiency rather than sound evaluation. This article addresses that by applying assessment literacy theory, supported by hermeneutic-aesthetic and post-phenomenological theories, as a framework to examine AI performance assessment tools in conservatory higher education. This study does not report primary empirical data from a single institutional site. Instead, it constructs a theoretically grounded critical interpretive synthesis of the existing empirical, technical, and pedagogical literature on

AI music assessment, organized through an assessment literacy matrix that permits systematic, criterion-referenced evaluation. The research question driving this article is: To what extent do current AI performance assessment tools satisfy the validity, reliability, fairness, formative utility, interpretive scope, and transparency criteria established by assessment literacy scholarship, and what philosophically grounded methodological frameworks are needed to evaluate their suitability for conservatoire practice?

Ⅱ. Literature Review

1. Assessment literacy: foundational positions

The concept of assessment literacy emerged in the early 1990s as a critical response to the gap between the technical complexity of educational assessment and educators' capacity to design and use assessment instruments effectively. Stiggins (1991) coined the term to identify the skill set required to evaluate the quality and accuracy of assessment results, arguing that most teachers lacked the foundational understanding needed to develop valid and reliable measures of student achievement. Popham (2018) subsequently expanded this conceptualization, distilling assessment literacy into six practical understandings: validity, reliability, fairness, score reporting, formative assessment, and affective assessment. These frameworks regard assessment literacy as a multidimensional evaluative orientation that encompasses technical, ethical, and pedagogical components (Yoon, 2020).

Crucially, assessment literacy scholarship holds that no instrument is inherently valid or reliable; validity and reliability are properties of how assessment results are interpreted and used in specific contexts (Messick, 1989). This constructivist understanding of validity, which Messick termed construct validity, requires that assessment instruments be evaluated not only for what they measure but also for whether that measure aligns with the construct as understood in the relevant domain. In conservatory music education, the construct of performance quality is irreducibly complex, encompassing technical accuracy, tonal control, rhythmic precision, expressive nuance, interpretive coherence, and stylistic appropriateness. These dimensions are not reducible to a common metric and carry different relative weights across instrument families, historical periods, and pedagogical traditions.

The application of assessment literacy theory to music education remains underdeveloped. Systematic reviews of the teacher-assessment-literacy literature indicate that empirical work has focused on

general classroom contexts, leaving specialist domains such as the performing arts largely unexamined (Evin, 2024; Kaufman et al., 2025; Moura et al., 2024; Pastore, 2023). The same reviews identify significant gaps in assessment literacy among practicing music teachers, particularly in the areas of validity and fairness, suggesting that the field is not well positioned to critically evaluate new instruments, whether human or computational.

2. A three-layer philosophical foundation for evaluating AI assessment instruments

A purely instrumentalist application of assessment literacy criteria, while necessary, is insufficient for AI music assessment. Treating an algorithmic instrument as merely another assessment tool, structurally equivalent to a rubric, a multiple-choice test, or a viva voce examination, obscures the specific ways in which computational mediation transforms what is assessed and how. To evaluate AI assessment instruments rigorously, the assessment literacy framework must be grounded in two further philosophical resources: a hermeneutic-aesthetic theory of musical interpretation that establishes the ontology of the construct being measured, and a post-phenomenological and critical-technology theory of mediation that explains how AI shapes the perceptual situation in which assessment occurs.

The three layers below are not intended as a general philosophical account of AI in music education, nor as a comprehensive survey of the traditions from which each layer is drawn. They constitute the minimum theoretical apparatus required to conduct the six-dimensional evaluative work in Section 4. Each layer addresses a specific analytical question that the assessment-literacy criteria alone cannot address: what is being measured (hermeneutic-aesthetic), how the instrument mediates the measurement situation (post phenomenological and critical-technology), and against what normative standard the instrument is evaluated (educational measurement). Theorists cited within each layer are drawn upon selectively for the specific analytical work they perform on the six ALAP dimensions elaborated below, rather than for a full account of their positions.

1) The hermeneutic-aesthetic layer: ontology of the construct

What is musical interpretation, considered ontologically? Hans-Georg Gadamer's hermeneutics offers one of the most influential answers in twentieth-century philosophy: interpretation is not the transmission of fixed meaning from a stable source to a receptive listener, but a fusion of

horizons *(Horizontverschmelzung)* in which the work, the tradition, the performer, and the listener participate in a co-constitutive event of understanding (Gadamer, 2004; Levinson, 2015). Adorno & Leppert (2002) extend this position into an explicitly musical register, treating performance not as the realization of a notated score but as an act of critical reflection in which the work is rearticulated within a present the score itself cannot anticipate. Said (1991) reframes performance as a culturally elaborate act, embedded in histories of authority, repertoire, and institutional power that no acoustic measurement can register.

More recent scholarship in performance studies has reinforced this position. Cook (2013) argues that the score is properly understood as a script for performance rather than a text to be reproduced, and that the musical analysis is incomplete until it engages with the performances that constitute the music as a sounding event. Clarke's (2005) ecological theory of musical perception similarly insists that interpretation is an embodied, situated act of meaning-making whose proper unit of analysis is the performance event, not the abstracted acoustic signal. Bowman's (1998) philosophical work on embodied musicianship frames musical knowing as enactive and tacit, irreducible to formal description and, in principle, resistant to algorithmic capture.

The hermeneutic-aesthetic layer has direct implications for AI assessment. If musical interpretation is constitutively a fusion of horizons, that is, performer, work, tradition, listener, and situation, then any assessment instrument that measures only the acoustic signal has access to only one term of an irreducibly multi-term relation. The acoustic signal is a trace of interpretation, not interpretation itself. To assess interpretation by measuring the signal is to commit a category error, treating a footprint as the walk that left it.

2) The post-phenomenological and critical-technology layer: epistemology of mediation

Before the conceptual scaffolding is introduced, a recent empirical finding clarifies what is at stake when AI is positioned as an assessor. The phenomenon of *AI anxiety* (Ou et al., 2025) illustrates the point that in encounters with an algorithmic assessor, students report a structurally different perceptual situation from human assessment, in which the assessor is encountered as a person within a shared interpretive community. The difference is not reducible to the question of accuracy. Even when an AI assessment is highly accurate, its use may alter the ontology of assessment by transforming the relational structure through which performance is experienced, interpreted, and evaluated.

According to Ihde (1990), human-technology relations vary systematically along a phenomenological axis defined by how a technology presents itself in experience, and this variation determines the unit of analysis any evaluation must adopt. He distinguishes four such relations, ordered by the degree to which a technology withdraws from, or insists upon, focal awareness. In embodiment relations, the technology withdraws most completely and becomes the medium through which the world is perceived; spectacles are the paradigm case, transparent in use precisely because perception passes through them. In hermeneutic relations, the technology no longer withdraws but produces a representation that must be read, as when a thermometer's reading stands in for a thermal state the body cannot directly register; the mediator becomes an object of interpretation in its own right. The four are not a closed catalog but an analytical instrument. They locate a technology's perceptual position, and that position determines what evaluation has to account for.

Verbeek expands on Ihde's theory from description to constitution. His mediation theory holds that technologies do not merely serve as intermediaries between humans and the world. Instead, they actively influence how the world is formed for the user and shape the user's role in relation to it (Verbeek, 2005, 2011). Perception is rearranged. Moral situations are reconfigured. The user who reads a thermometer is not the same epistemic agent as the user who notices a chill, and the student assessed by an algorithm is not the same evaluative subject as the student assessed by a teacher. The mediator does not transmit a pre-existing situation. Instead, it produces a new one.

Feenberg (2002) extends the argument into political terrain. His critical theory of technology rejects the view that technologies are value-neutral instruments whose ethics depends only on their use. Every technology, on this account, bears a technical code, an embedded set of choices about what the technology is for, whose interests it serves, and what practices it renders thinkable or unthinkable. The technical code of an AI music assessment system is therefore an object of evaluation in its own right, independent of how any particular institution deploys it. Winner (2017) sharpens this position by reviving an older argument that has lost none of its force. Artifacts have politics. Technological designs encode social arrangements, and the question of whose interests a design serves cannot be deferred to the question of use. A bridge built too low to admit public buses is exclusionary regardless of what its designer intended. An AI assessment system trained on a demographically narrow corpus encodes a culturally specific definition of musical quality regardless of how its developers describe it.

Pasquale's (2015) account of the black box society shows how proprietary status, technical complexity, and contractual obscurity combine to render the operations of contemporary algorithmic systems unintelligible to the people whose lives their outputs affect. Under such conditions,

Feenberg's technical code is invisible, and Winner's politics of artifacts is undetectable from the outside. The question of whose values the system encodes can no longer be answered by inspection. It becomes a question of access and, therefore, of power. Crawford & Paglen (2021) push this point upstream to training data. Their analysis demonstrates that the datasets on which contemporary machine-learning systems are trained are not neutral records of the world but contestable taxonomies, freighted with specific acts of classification and decisions about what counts as a category and what counts as a member of it.

Read in sequence, these five theorists offer the conceptual vocabulary needed to treat AI music judgment as a distinct evaluative formation, rather than as a faster, more consistent, or more scalable imitation of human judgment. This approach serves four analytical purposes. It shifts the discussion of AI assessment from automation to mediation, showing how AI alters the perceptual conditions under which musical evaluation occurs (Ihde, 1990; Verbeek, 2005, 2011). It identifies the value choices embedded in any technical system as objects of scrutiny rather than by-products of design (Feenberg, 2002; Winner, 2017). It reframes algorithmic opacity as an evaluative problem rather than an engineering inconvenience (Pasquale, 2015). In addition, it traces the question of bias upstream from output to training data, where the most consequential design choices have already been made (Crawford & Paglen, 2021). The post-phenomenological-critical-technology layer has direct implications for AI assessment. AI is not a neutral measurement instrument that returns a value; it is a mediating technology that reconfigures the assessment situation, encodes value choices, and mediates power. Evaluating it requires the philosophical resources to make these mediations and value choices visible.

3) The educational measurement layer and the coherence of the three layers

To clarify how each philosophical layer performs analytical work within the six-dimensional Assessment Literacy Audit Protocol (ALAP), the following mapping identifies which theoretical resources ground each dimension of the evaluation developed in Section 4. The hermeneutic-aesthetic layer, drawing principally on Gadamer, Adorno, Bowman, and Cook, grounds the construct-validity critique in Section Ⅳ.1. It explains what conservatoire assessment seeks to evaluate and why acoustic feature vectors cannot fully represent that construct. The same layer also grounds the interpretive-scope critique in Section Ⅳ.5 by distinguishing corpus typicality from the critical judgment of artistic individuality.

The post-phenomenological and critical-technology layer, drawing principally on Ihde, Verbeek,

Feenberg, Winner, Pasquale, Crawford, and Paglen, grounds three dimensions of the audit. First, it supports the reliability critique in Section Ⅳ.2 by showing how technical code decisions and variability in training data can produce measurement instability. Second, it supports the fairness critique in Section Ⅳ.3 by providing a vocabulary to identify and contest encoded value choices. Third, it supports the transparency critique in Section Ⅳ.6 by naming the algorithmic and corpus opacities that shape the commercial AI assessment market.

The educational-measurement layer, drawing primarily on Messick, Stiggins, Popham, and Black and Wiliam, grounds the formative-utility critique in Section Ⅳ.4. It provides criteria for evaluating the quality of formative feedback and a shared assessment-literacy vocabulary that integrates the six dimensions into a coherent audit rather than a loose checklist. Each dimension of the ALAP is therefore doubly anchored. It is grounded, first, in a philosophical layer that clarifies what the dimension addresses and, second, in assessment-literacy theory that specifies how to evaluate it. This scaffolding distinguishes an assessment-literacy critique of AI tools from a purely philosophical objection to their use. It also provides the theoretical foundation that Section Ⅳ applies to each audit dimension in turn.

3. Music performance assessment: an epistemologically contested tradition

Music performance assessment has generated a substantial body of empirical research spanning more than six decades, much of it preoccupied with the reliability problem. The persistent difficulty of achieving consistent ratings across adjudicators, across time, and across assessment contexts. Early experiments at the 1960 International Chopin Competition, revisited and synthesized by Chmurzyńska (2023), established what would become a defining finding in the field, namely that expert music judges exhibit both low inter-rater and low intra-rater reliability, rating the same performances differently when heard on separate occasions and often diverging markedly from one another even on ostensibly technical criteria. More than six decades of subsequent research have not resolved this problem.

A study by Pastore (2023) confirmed that noise in judgment, that is, systematic and random error in adjudicators' ratings, remains a pervasive feature of music performance assessment, even when rubrics and structured criteria are employed. A systematic review of solo music performance assessment criteria by Moura et al. (2024) analyzed 26 assessment systems drawn from 20 empirical studies and found that, while systems typically structured evaluation around technical, interpretative, and expressive categories, the operationalization of these categories was highly inconsistent, and no

consensus has emerged on how expressive and interpretative qualities should be weighted relative to technical accuracy. The review also found that no differentiation was made in assessment criteria based on students' learning stages, pointing to a structural failure in the field's approach to formative versus summative assessment design.

These empirical findings carry a sharp methodological implication. Music performance assessment suffers from a persistent validity-reliability trade-off. Instruments that optimize for reliability by narrowing the construct do so at the expense of validity; instruments that attempt comprehensive construct coverage produce unreliable ratings. Rubrics have been proposed as a partial solution. Álvarez-Díaz et al. (2021) demonstrated that well-designed rubrics, combined with trained adjudicators, improved inter-rater reliability in competition settings, but even rubric-based systems require training and contextual calibration to function effectively. Wind et al. (2026) further complicated the picture by demonstrating that rater connectedness, defined as the social and professional relationships among adjudicators, significantly affects student achievement estimates, introducing a relational variable into reliability that technical solutions alone cannot address.

From an equity standpoint, Kaufman et al. (2025) documented systematic patterns of adjudicative inequity in music performance assessment, finding that factors such as race, gender, and institutional affiliation influence assessment outcomes in ways rarely acknowledged in the design of assessment instruments. Their literature review argued that adjudicating equity requires explicit methodological attention, not merely the declaration of neutral criteria, and that the field's tendency to treat reliability as the primary indicator of quality has obscured persistent fairness deficits.

Ⅲ. Method

1. Research design

This article employs a critical interpretive synthesis as its methodological design (Dixon-Woods et al., 2006). Critical interpretive synthesis goes beyond a systematic narrative review by explicitly deploying a theoretical framework as an analytical lens, using it to interrogate, organize, and evaluate the findings of the reviewed literature rather than merely summarizing them. This approach is epistemologically appropriate for an inquiry whose goal is not to aggregate empirical findings but to assess the evaluative adequacy of a class of instruments, AI performance assessment tools, against criteria that have not previously been applied to them in a structured way.

The paradigmatic positioning of this inquiry is critically interpretivist. Drawing on the tradition of critical educational research (Apple & Apple, 2004; Biesta, 2009) and the post-phenomenological theory of technological mediation (Ihde, 1990; Verbeek, 2011), this study positions AI music assessment tools as objects of normative critique rather than as neutral instruments to be benchmarked.

2. Sources and inclusion criteria

Database searches were conducted from August 2025 to March 2026, with a repeat run in April 2026 to capture publications that appeared during manuscript preparation. In Scopus, the primary Boolean combination was TITLE-ABS-KEY (‘artificial intelligence’ OR ‘machine learning’ OR ‘automated assessment’ OR ‘algorithmic assessment’) AND (‘music performance’ OR ‘instrumental performance’ OR ‘piano’ OR ‘voice’ OR ‘conservatoire’ OR ‘music education’) AND (assessment OR evaluation OR literacy OR validity OR reliability OR fairness). Equivalent combinations were adapted for Web of Science Core Collection using the TS = topic operator across the SSCI and AHCI indices; for ERIC using descriptor-based search with the DE = operator; and for Google Scholar using truncated Boolean strings, limiting results to the first 200 relevance-ranked results per query to manage the higher noise-to-signal ratio typical of that database. The assessment-literacy and philosophy-of-technology strings were run as separate but overlapping searches to identify theoretical and empirical literature independently. Records were exported to Zotero for deduplication, performed at the title-and-author level.

Inclusion criteria were specified separately for each of the three source categories and are described in detail below. Exclusion criteria that applied across all categories were: non-scholarly commentary or opinion pieces lacking an argumentative structure; conference abstracts without accompanying full papers; duplicate publication of material already included; publications in languages other than English or Malay for which no authoritative translation could be verified; and, for AI-adjacent sources, an exclusive focus on music generation or composition rather than assessment (which is a construct-non-equivalent enterprise and lies outside the article’s scope). Sources whose full text could not be retrieved through institutional access, interlibrary loan, or open repositories within the search window were excluded, and their exclusion was recorded.

The corpus construction followed this flow. Database searches identified approximately 1,140 records across the four databases before deduplication, which were reduced to approximately 830 after duplicate removal. Title-and-abstract screening against the category-specific inclusion criteria yielded approximately 210 records for full-text assessment. Full-text review, backward

and forward citation tracing, and hand-searching of the six named journals yielded a final analytical corpus of 52 sources: 18 in Category A (empirical music performance assessment), 23 in Category B (empirical and technical AI music assessment literature, including four commercial-vendor technical documents cited alongside peer-reviewed sources), and 11 in Category C (theoretical and philosophical scholarship anchoring the three-layer framework). The distribution of sources across categories reflects the article's analytical emphasis on Category B, where the most consequential recent literature clusters and the reviewed evidence base is thinnest against assessment-literacy criteria. These numbers are reported for transparency; the study is a critical interpretive synthesis in the sense of Dixon-Woods et al. (2006) rather than a systematic review, and the numbers are intended to indicate corpus scope rather than to warrant a claim to exhaustive coverage of the field.

The analytical procedure proceeded in four discrete steps. First, each source was read in full and coded against the six ALAP dimensions using a structured coding template that captured the source's central claim(s), methodological approach, principal evidentiary basis, and stance on each dimension. Coding was recorded in a spreadsheet with one row per source and one column per dimension. Second, the coded entries were aggregated into a dimension-by-dimension analytical matrix, allowing the corpus to be inspected column-wise for each dimension. Third, the matrix was analyzed for convergence, divergence, and gap patterns in each dimension, with attention to whether an apparent consensus rested on a common evidence base or on parallel citations to the same limited primary source. Fourth, the dimensional verdicts that appear in Tables 1, 3, and 4 and are elaborated in Sections Ⅳ.1 through Ⅳ.6 were derived from the matrix analysis by identifying the modal position across the corpus at each dimension, its principal evidentiary basis, its methodological limitations, and its consequences for the assessment-literacy evaluation. Coding was conducted by the first author, with the second and third authors independently reviewing the coded matrix for interpretive validity across the six dimensions. Four coding discrepancies were identified during this cross-check, all of which concerned the classification of dimensional stance rather than the extraction of source content; they were resolved through rereading the disputed sources and discussion, with all four resolved without residual disagreement.

Category A — Empirical studies on music performance assessment. Peer-reviewed studies that examine the reliability, validity, fairness, and criteria of music performance assessment in formal educational or competition contexts, published in indexed music education, music psychology, and music analysis journals between 2000 and 2026. Inclusion criterion: clear methodological design in Scopus- or WoS-indexed journals.

Category B — Empirical and technical literature on AI music assessment. Peer-reviewed studies reporting on AI or computational tools for music performance assessment, including technical architecture, validation procedures, and pedagogical effects, published between 2018 and 2026. Inclusion criterion: studies addressing conservatory-relevant performance parameters or pedagogical effects in postsecondary contexts. Adjacent literature on algorithmic bias in audio-domain machine learning (Koenecke et al., 2020) and on automated educational scoring more broadly (Madnani & Cahill, 2018) was included where it directly informed the analysis of fairness or reliability.

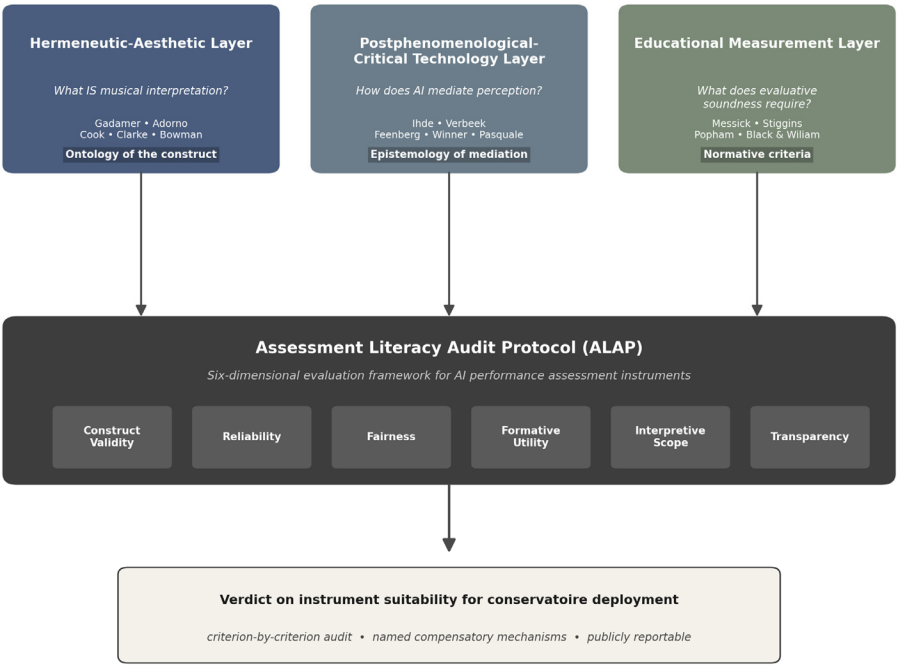
Category C — Assessment literacy, philosophy of music, philosophy of technology, and educational measurement scholarship. Foundational and contemporary scholarship anchoring the three philosophical layers of the framework: hermeneutic-aesthetic theory (Adorno & Leppert, 2002; Bowman, 1998; Clarke, 2005; Gadamer, 2004; Said, 1991); post-phenomenological and critical-technology theory (Crawford & Paglen, 2021; Feenberg, 2002; Ihde, 1990; Pasquale, 2015; Verbeek, 2005, 2011; Winner, 2017); and educational measurement (Black & Wiliam, 1998; Messick, 1989; Popham, 2018; Stiggins, 1991).

3. Analytical framework: the assessment literacy audit matrix

The analytical matrix developed in this synthesis draws on Popham's (2018) six understandings of assessment literacy and is further informed by Stiggins's (1991) seven criteria: clear purpose, clear achievement targets, appropriate assessment methods, adequate sampling, control of bias, effective communication, and student involvement. These assessment-literacy principles are situated within a three-layer philosophical foundation comprising the hermeneutic-aesthetic, post-phenomenological and critical-technology, and educational-measurement layers.

Figure 1 presents the overall architecture of the framework, showing how the three philosophical layers inform the six-dimensional Assessment Literacy Audit Protocol (ALAP) used to evaluate the suitability of AI assessment tools for conservatoire performance education. Table 1 specifies the criteria associated with each ALAP dimension, the evidence identified in the literature, and the corresponding evaluative verdict concerning current AI-based music assessment systems. Table 2 explains the conceptual basis of each philosophical layer and its particular analytical function within the framework. Rather than merely summarising previous findings, each ALAP dimension treats the reviewed literature as evidence to be critically evaluated. The framework therefore moves beyond descriptive synthesis by examining the extent to which existing AI music assessment

systems satisfy broader requirements of validity, reliability, fairness, formative utility, interpretative scope, and transparency.



[Figure 1] Three-layer philosophical architecture of the assessment literacy audit protocol (ALAP)

<Table 1> The assessment literacy audit protocol (ALAP): Six-dimensional matrix for evaluating AI music performance assessment instrument

Dimension	Evaluative criterion	Evidentiary basis from reviewed literature	Verdict for current AI tools
Construct validity	Does the instrument measure the conservatoire performance-quality construct, or only a proxy?	Moura et al. (2024); Phanichraksaphong & Tsai (2023); Messick (1989); Cook (2013); Clarke (2005)	Construct under-representation; measures notation-fidelity rather than interpretive coherence.
Reliability	Are scores consistent across raters, occasions, and conditions, within validity limits?	Chmurzyńska (2023); Wind et al. (2026); Madnani & Cahill (2018)	High consistency for pitch/rhythm; sensitive to acoustic conditions; model-drift untested.

<Table 1> Continued

Dimension	Evaluative criterion	Evidentiary basis from reviewed literature	Verdict for current AI tools
Fairness	Does the instrument avoid systematic disadvantage of groups on construct-irrelevant grounds?	Kaufman et al. (2025); Holmes (2012); Mangal & Pardos (2024); Koenecke et al. (2020)	Inherits training-corpus inequities; AI-anxiety threats to ecological validity.
Formative utility	Is feedback specific, timely, actionable, and aligned with learning goals?	Black & Wiliam (1998); Ou et al. (2025); W. Li et al. (2025); M. Zhang (2025)	Effective for early technical skills; ineffective for advanced expressive learning.
Interpretive scope	Does the instrument access expressive, interpretive, and stylistic dimensions of performance?	M. Zhang (2025); Cook (2013); Bowman (1998)	Expressive, timbral, and stylistic dimensions remain outside the model.
Transparency	Are scoring procedures and training data sufficiently visible to be examined and contested?	Stiggins (1991); Buck (2008); Holmes (2012); Pasquale (2015); Nichols (2014); Crawford & Paglen (2021)	Proprietary algorithms; undisclosed training data; structural opacity.

<Table 2> Three philosophical layers grounding the assessment literacy audit protocol

Layer	Ontological / epistemological / normative question	Analytical work performed in this article	Principal theorists
Hermeneutic-aesthetic	What is musical interpretation, ontologically?	Grounds the construct critique which distinguishes expressive from erroneous deviation; argues that interpretation is a fusion of horizons not reducible to acoustic features.	Gadamer (2004); Adorno & Leppert (2002); Said (1991); Cook (2013); Clarke (2005); Bowman (1998)
Post-phenomenological & critical-technology	How does AI mediate the perceptual and political situation in which assessment occurs?	Grounds the AI-anxiety / ecological validity analysis and the transparency / power critique; names the alterity relation and the technical code.	Ihde (1990); Verbeek (2005); Feenberg (2002); Winner (2017); Pasquale (2015); Crawford & Paglen (2021)
Educational measurement	What does evaluative soundness require of any assessment instrument?	Provides the six-dimensional normative framework; operationalises validity, reliability, fairness, formative utility, interpretive scope, transparency.	Messick (1989); Stiggins (1991); Popham (2018); Black & Wiliam (1998)

4. Limitations of the design

As a conceptual review rather than a primary empirical study, the article is constrained by the limits of its source corpus, particularly restricted access to unpublished pilots, proprietary documentation, and developer perspectives. The three-layer philosophical foundation offers a coherent, defensible basis for analysis, though it is just one of several evaluative frameworks for AI music tools. An explicitly socio-material perspective (Fenwick & Edwards, 2010) or an ecological dynamics perspective (Araújo & Davids, 2011), for example, may yield different yet potentially complementary insights. The present positioning is therefore intentional and forms part of the article's analytical contribution, while remaining evaluative rather than purely descriptive.

5. Ethical considerations, conflicts of interest, and AI-assistance disclosure

As a conceptual review of peer-reviewed scholarship and openly accessible technical documentation, the study does not involve human or animal subjects, primary data collection, identifiable personal data, or institutional records. Therefore, no ethics committee approval was required under the standards of the authors' home institution. The authors declare no financial, professional, or personal conflicts of interest with any of the commercial platforms named in the synthesis (*SmartMusic / MakeMusic Cloud, Yousician, Simply Piano, Skoove, Flowkey, Tonara, MatchMySound, Musiah*), have received no funding, equipment, or in-kind support from their vendors, and have no consulting, advisory, or editorial relationships with the cited research-stage AI assessment systems. AI assistance was used during manuscript preparation in two limited ways: (i) consultation with a large language model for literature scoping and proofreading suggestions, and (ii) the use of a reference manager (Mendeley) for citation insertion and bibliography formatting. No AI system generated the article's analytical claims, theoretical positions, or evaluative judgments; these are the responsibility of the named authors.

Ⅳ. Results

1. Construct validity: What do AI tools actually measure?

The AI music-assessment market offers a finite set of products against which this construct

critique can be made specific. Eight platforms dominate institutional and consumer use: *SmartMusic (MakeMusic Cloud), Yousician, Simply Piano, Skoove, Flowkey, Tonara, MatchMySound,* and *Musiah* (Chen, 2025; Cui, 2023; Kão & Niitsoo, 2014; Lovett, 2024; Merchán Sánchez-Jara et al., 2024; Preda-Uliță, 2025; Rizal & Milyartini, 2024; Tao, 2026; Tian, 2025; Wei et al., 2025) Each, in its own technical documentation, discloses a measurement pipeline organised around two acoustic dimensions: pitch deviation from a notated reference, and onset or duration deviation from a notated metric grid.

The studies by Chen (2025), Wei et al. (2025), and Buck (2008) examine the effect of *SmartMusic*® assessment on students' music performance and the integration of research-based teaching and learning components. *SmartMusic®,* now known as *MakeMusic Cloud,* is a computer program that uses a combination of software and hardware to assist a student's musical performance. The data suggest that the *SmartMusic*® assessment program positively affected music performance skills, particularly in technically demanding passages.

Eremenko et al. (2020) carried out a case study applying the system, which is based on *Essentia* and *MusicCritic,* and developed a complete system for the formative assessment of guitar exercises. *Essentia* and *MusicCritic* provide a way to create and assess tailored exercises that support personalized rubrics. Using a Web API, *MusicCritic* acts as an external music assessment tool via the *Learning Tools Interoperability (LTI)* standard, thus able to communicate with any *Learning Management System (LMS)*, for example *MOOC* platforms (Bozkurt et al., 2018).

Phanichraksaphong & Tsai (2023) delineate an assessment structured as a classification problem that categorizes piano performances into 'Good', 'Fair', or 'Poor'. For timbre-based methodologies, they introduce a *timbre-based WaveNet*, a *timbre-based MLNet*, a *timbre-based CNN*, and a *timbre-based CNN Transformer*. For pitch-based methodologies, this approach presents a *pitch-based CNN* and a *pitch-based CNN Transformer*. Each platform offers *timbre-based WaveNet, timbre-based MLNet, timbre-based CNN,* and *timbre-based CNN Transformers*. The study offers *Pitch-based CNN* and *Pitch-based CNN Transformers* for pitch-based methodologies.

Caruso (2024) focuses on biofeedback technologies and artificial intelligence, exploring their applications in music performance analysis and creation through the lens of his artistic practice and research path. This is achieved by integrating qualitative and quantitative data using the multimodal annotation software *ELAN*. The software allows the import of video recordings, audio tracks, and motion capture data (displacement, velocity, and acceleration), enabling their integration as streams. The system thereafter allows the annotation of multimodal audio and video data.

Ou et al. (2025) conducted a study involving students in an experimental group who used

an AI-assisted practice application called *Violy*, installed on their personal devices (smartphones or tablets) compatible with both Apple iOS and Android. *Violy* provides practice sessions for multiple instruments, including piano, violin, saxophone, flute, and trumpet. The application incorporates Score Following technology, using AI and machine learning algorithms to assess players' intonation, rhythm, and performance approaches, augmented by the *OLTW* and *GMM algorithms*. *Violy* offers a range of tools for music learners, including Daily Challenge, Music Cloud, Video Demonstrations, Accompaniment, Audition Report, and Note-by-Note.

Li et al. (2025) explore the impact of integrating artificial intelligence (AI) and e-learning tools, including the ability to upload multiple audio files, real-time playback, recording, and the visualization of audio differences. To ensure the tool would efficiently implement the audio comparison function, relevant code was generated using the generative AI software *ChatGPT-4o*. Using *JavaScrip*t and the *Web Audio API*, the code used the *Wavesurfer.js library* to visualize audio waveforms and compare two recordings. Code editing and testing were performed online via *CodePen*, which provided a visual editor and real-time previews of audio playback and comparison results.

Ouyang (2025) investigates how AI-enhanced learning environments influence creative expression in vocal performance. The *Vienna Integrated Model of Art Perception (VIMAP)* helps explain students' aesthetic engagement at the cognitive level. It features five outcome types, namely *Facile, Novelty, Harmonious, Negative*, and *Transformative,* which are tailored to represent learners' interpretative responses to vocal performance tasks.

Arora (2025) conducts an empirical analysis of self-recorded violin performances, employing a deconstructive system that enables users to compare their audio with original reference tracks using signal processing techniques, including *dynamic time warping (DTW), chroma feature extraction,* and *waveform deviation analysis*. The evaluation results were consolidated into individualized, summarized feedback using *large language models (LLMs)*, enabling real-time feedback delivery via the application *SoundRubric.*

Zhang (2025) conducts a comparative examination of three prominent deep learning architectures, including *Long Short-Term Memory (LSTM) networks, Transformer models,* and *Generative Adversarial Networks (GANs*), focusing on AI-generated music composition and transcription using the *MAESTRO dataset*. The principal innovation is a dual evaluation system that combines objective measures (perplexity, harmonic consistency, and rhythmic entropy) with subjective human assessments via a *Mean Opinion Score (MOS)* study with 50 listeners.

The pattern visible across Table 3 supports the structural claim that AI music assessment

systems perform well on the dimensions conservatoire pedagogy treats as preconditions for evaluation, and weakly or not at all on the dimensions conservatoire pedagogy treats as the criteria of evaluation proper.

<Table 3> Construct validity audit: Conservatoire performance dimensions against AI measurement capability

Dimension	Conservatoire weighting	AI measurement	Omitted construct component	Validity verdict (source)
Technical accuracy (pitch, intonation)	Foundational; treated as a precondition rather than a primary criterion of artistic distinction.	Cent-level deviation from notated reference frequencies; reliable onset and offset detection.	Minimal within bounded domain; the assumption that fidelity to notation indexes quality holds at early stages but weakens at advanced ones.	Adequate within domain (Phanichraksaphong & Tsai, 2023) .
Rhythmic precision	Foundational; rubato and metric flexibility are valued, not penalised, at advanced level.	Timing deviation from a notated metric grid.	The pedagogically central distinction between metronomic error and rubato as expressive choice; current AI systems typically classify both as timing deviation.	Partial; systematic conflation of expressive and erroneous timing (Phanichraksaphong & Tsai, 2023).
Tonal control / timbre	High-weighted across instrument families; treated as a primary index of artistic development.	Limited spectral feature extraction (centroid, MFCCs, brightness); strongly sensitive to recording conditions.	Tonal choices made in relation to phrase, register, period, and ensemble context; AI extracts features without access to that situatedness.	Low validity beyond gross categorisation (Moura et al., 2024)
Expressive quality (dynamics, articulation, agogics)	Rated equally or more important than technical accuracy in advanced evaluation.	Statistical distance from a modelled expressiveness profile derived from a training corpus; partial correlation with listener preference scores.	The performer's situated interpretive choices, grounded in style, training, and announced intent; corpus correlation is not the same construct as conservatoire faculty's evaluative criteria.	Construct mismatch (Zhang, 2025).
Interpretive coherence	Criterion of artistic maturity; weighted heavily in advanced examination.	Not measured at the level of long-arc phrase, formal section, or whole-work trajectory.	The entire dimension; AI has no representation of large-scale interpretive architecture.	Construct under-representatio n Cook (2013).

<Table 3> Continued

Dimension	Conservatoire weighting	AI measurement	Omitted construct component	Validity verdict (source)
Stylistic appropriateness	Instrument-, period-, and tradition-specific; central to repertoire-grounded pedagogy.	Not measured beyond superficial corpus matching where attempted.	Style as a hermeneutic relation between performer, work, and performance tradition; not a feature of the acoustic signal.	Construct under-representation (Cook, 2013; Moura et al., 2024).
Embodied communicativity	Recognised constituent of performance quality, particularly in live evaluation.	Not accessible to acoustic signal in isolation.	The performer's bodily, intersubjective enactment of the work; resistant in principle to capture by signal-based methods.	Theoretically resistant to acoustic quantification (Bowman, 1998)

2. Reliability: Consistency, reproducibility, and the illusion of objectivity

AI performance assessment tools are often promoted for their reliability advantage over human adjudicators. In this respect, the same recording, processed by the same algorithm, will yield the same score regardless of who administers the assessment or when. This is a real and nontrivial advantage in a field where intra-rater and inter-rater reliability problems have been documented for more than sixty years (Chmurzyńska, 2023; Wind et al., 2026). For the technically bounded parameters that AI systems measure (pitch deviation from a reference frequency, onset and offset timing relative to a notated score), consistency is near-perfect, and this represents a genuine contribution to the reliability landscape of music performance assessment.

However, assessment literacy distinguishes between measurement consistency and reliability in the full evaluative sense. Popham (2018) situates reliability within the overarching framework of validity, arguing that a measurement may produce highly consistent results while nevertheless failing to capture the intended construct. In this case, an AI system that evaluates every performance of a Beethoven sonata against a MIDI reference may yield highly consistent scores for deviations from the notated score. However, if the construct of interest also includes interpretive deviation as a musically meaningful quality, such consistency may not provide valid evidence of performance quality. This distinction helps explain why reliability, considered in isolation, may not provide a sufficient foundation for an assessment instrument.

The history of educational AI offers an instructive precedent. Automated essay-scoring systems,

widely used in standardized testing and writing instruction since the early 2000s, have demonstrated high scoring consistency, with the same essay typically receiving the same result when processed repeatedly by a given algorithm. At the same time, these systems have been criticized for persistent underrepresentation of the construct. Their evaluations often rely heavily on surface features, including text length, lexical diversity, and syntactic complexity, which may correlate with human ratings without fully representing the broader construct of writing quality (Madnani & Cahill, 2018). A comparable tension may be observed in AI-based music assessment. Gains in consistency achieved through narrowly defined proxies can be presented as evidence of assessment quality even when the broader construct remains only partially represented by the instrument. The precedent of automated essay scoring therefore suggests that measurement consistency, although valuable, may not by itself provide sufficient justification for adoption.

Beyond this conceptual point, the reliability of AI systems for parameters beyond pitch and rhythm is demonstrably lower than the available information suggests. Research on timbre assessment illustrates this clearly. Timbre, that is, the spectral and temporal envelope characteristics that define tonal quality and are central to conservatory instrument teaching, is highly sensitive to recording conditions, microphone placement, room acoustics, and playback equalization (Zhang, 2025). Studies on *SmartMusic* have documented that microphone placement significantly affects pitch detection accuracy (Buck, 2008; Nichols, 2014), introducing an environmental reliability variable that has no analog in human adjudication. The implication is that AI reliability for technical parameters is conditional on controlled acoustic conditions that are rarely available in conservatory teaching environment

A further reliability concern specific to AI systems is model drift, whereby systems trained on particular repertoire corpora, recording conditions, or performer populations may yield different results when applied beyond those conditions. Because the provenance and composition of the training data used in commercial AI music-assessment tools are seldom disclosed, it remains difficult to determine whether reported reliability extends to unfamiliar repertoire, performance contexts, recording environments, or performer groups.

A separate and equally consequential reliability concern is the near-total absence of published blinded comparisons of AI scoring with conservatoire-level examiner panels on graded repertoire. The empirical literature on SmartMusic, the most widely deployed classroom platform, consists predominantly of vendor-supported dissertations: the principal doctoral study openly acknowledges that *MakeMusic p*rovided 60 subscriptions and 60 microphones for this study through a successful *SmartMusic* Research Grant (Buck, 2008).

3. Fairness: Algorithmic bias and structural inequity

The newly developed fairness analysis identifies six distinct mechanisms by which AI music assessment instruments may produce, or fail to produce, equitable evaluation. These mechanisms operate at different levels of the assessment pipeline (the human comparator, the training data, the feature-extraction architecture, the source frameworks, the methodological literature, and the perceptual situation of the student being assessed), and they call for different evaluative responses, yet they are typically discussed in the algorithmic-fairness literature in isolation from one another. Table 4 presents the six mechanisms in a single view.

<Table 4> Fairness mechanism decomposition: Six distinct sources of equity deficit in AI music performance assessment

Fairness threat	How it operates in AI music assessment	Evidence base	Diagnostic status and corrigibility
Human adjudicator baseline bias	Race, gender, and institutional affiliation may systematically influence human-issued assessment ratings even under conditions of equivalent technical performance. This is the comparator AI is expected to improve on; it is the bar AI must clear rather than a property of AI itself.	Kaufman et al. (2025)	Comparator condition. Establishes the fairness baseline; AI tools that merely match this profile have produced no equity gain.
Training-data inheritance	Machine-learning models are trained to predict the output distribution of their training data. Where the data over-represent White, Western-trained, conservatoire-educated performers, the demographic norms of that population become the implicit reference against which subsequent students are evaluated.	Holmes (2012); Mangal & Pardos (2024) on general algorithmic bias in education.	Structurally inherent to current ML methodology. Partially mitigable through corpus diversification, but no procurement-stage method exists to verify diversification claims without training-data disclosure.
Audio-domain pipeline inheritance	AI music assessment systems use acoustic feature-extraction pipelines that are technically continuous with commercial automatic speech recognition. In the adjacent speech domain, comparable pipelines exhibit approximately twofold disparities in error rate for minoritised speakers.	Koenecke et al. (2020): five major commercial ASR systems show approximately 2× word error rate for Black speakers compared with White speakers in a controlled study.	Documented in adjacent domain; not reversed by music-domain evidence. The burden of proof rests on developers to demonstrate that music-domain implementations behave differently.

<Table 4> Continued

Fairness threat	How it operates in AI music assessment	Evidence base	Diagnostic status and corrigibility
Cultural specificity of source frameworks	Most published music performance assessment systems are designed for Western classical repertoire and for specific instrument families, with little differentiation by cultural background, learning history, or stylistic tradition. AI systems trained on these frameworks inherit their cultural specificity as a built-in property.	Moura et al. (2024): systematic review of 26 assessment systems drawn from 20 empirical studies.	Inherited from the music assessment literature itself. Partially mitigable through deliberate framework adaptation, but no industry standard for such adaptation currently exists.
Methodological gap: intersectionality-aware design	Intersectionality-aware machine learning has been proposed as a methodological response to compounded bias in educational AI more broadly, but no music-specific implementation has been reported in the published literature at the time of writing.		Identified gap. No current procurement-stage option; addressable only through research and development the field has not yet undertaken.
Ecological validity threat: AI anxiety / alterity	Students performing under AI assessment may experience heightened anxiety and reduced expressive risk-taking, particularly those with strong artistic identities. The performance heard by the AI is therefore not necessarily the same performance the student would give to a teacher in a room; the assessment situation has been transformed by the AI's evaluative gaze.	Nichols (2014); Ou et al. (2025); post-phenomenological reading based on Ihde (1990) and Verbeek (2005).	Not corrigible through algorithmic improvement. The threat arises from the alterity relation itself, meaning that the AI is experienced as a quasi-other. This persists even if model accuracy improves.

4. Formative utility: Feedback quality and pedagogical affordance

Black & Wiliam's (1998) foundational synthesis of formative assessment research established that effective feedback must be specific, timely, actionable, and oriented toward the learning goal rather than the performance score. Popham (2018) identifies formative assessment understanding as a distinct literacy dimension precisely because feedback that does not meet these criteria, however reliably produced, fails as an educational instrument.

AI music assessment tools have demonstrated genuine formative strengths for technical parameters

in early learning contexts. Immediate, note-level feedback on pitch and rhythm deviations, delivered visually in real time, provides the specificity and timeliness that effective formative feedback requires. Research on AI-assisted practice found that students using AI tools showed improved self-monitoring of technical parameters and increased practice duration (Ou et al., 2025), and the study on AI feedback in vocal training found metacognitive improvements in breath management and pitch accuracy (Li et al., 2025). These are educationally meaningful gains, and they establish a genuine formative affordance for AI tools within the bounded construct domain of technical accuracy. However, both studies consistently found that AI tools' formative utility did not extend to expressive, interpretive, or artistic development.

The most comprehensive empirical synthesis of AI in music education to date is Liu (2026) PRISMA-protocol systematic review and meta-analysis of AI in piano education from 2020 to 2025, aggregating 62 studies and 2,812 participants and reporting a moderate pooled effect size of d = 0.442 (95% CI [0.398, 0.486]) for AI-assisted instruction over conventional methods. Disaggregated, the effect was d = 0.462 for technical-skill outcomes, d = 0.432 for musical expression, and d = 0.434 for motivation. The headline finding appears to support AI deployment. However, this review explicitly criticizes its own evidence base, noting that the included studies 'equate improved learning outcomes with higher scores on automated pitch detection or rhythm accuracy tests' without interrogating whether those proxies correspond to musicianship, and that "evidence on long-term learning trajectories and higher-order artistic judgment remains sparse". The construct-confound that this article has identified at the level of individual studies is therefore also present at the level of the meta-analytic synthesis: the field is measuring a moderate gain on operationalizable proxies and reporting it as a moderate gain on musicianship.

Zhang et al. (2026) scoping review of formative assessment in higher music education found that conservatoire practice is already weakly configured for formative assessment, with a summative examination culture dominating. The arrival of AI tools in this context creates a risk that their genuine but narrow formative capability is overextended, with tools appropriate for beginners' technical development being deployed as formative assessment instruments for advanced students whose learning needs lie primarily in the expressive and interpretive domains.

5. Interpretive scope: The aesthetic dimension of performance quality

The dimensions of musical performance that conservatoires value most highly (expressive communication, interpretive individuality, stylistic authenticity, embodied musicianship) constitute

what might be termed the aesthetic dimension of performance quality. Assessment literacy requires that any assessment instrument's construct be clearly specified and that its boundaries be honestly communicated. The critical problem with AI performance assessment tools in this dimension is not merely that they underrepresent expressive and interpretive qualities in their current construct model, but that they rarely make this limitation explicit in ways that allow users to calibrate their interpretations accordingly.

The interpretive dimension encompasses not only what current AI systems do not yet measure but also what may prove structurally difficult to measure through acoustic feature extraction alone. Musical interpretation, on the hermeneutic-aesthetic level, is not merely a complex pattern of acoustic features. Instead, it is a relational act of communication among performer, work, tradition, and listener within a specific cultural and temporal context (Clarke, 2005; Cook, 2013). An AI assessment instrument that reduces interpretation to a feature vector, however sophisticated, operationalizes a different construct than the one conservatory pedagogy targets. Assessment literacy requires that this category distinction be made explicit and that its consequences for construct validity be made clear in any deployment of AI assessment tools.

The interpretive-scope deficit may not be fully resolved by improved engineering alone. Current research on computational modeling of musical expressiveness, including recent deep-learning approaches, commonly predicts statistical patterns in timing, dynamics, and articulation by learning from large corpora of recorded performances. This constitutes a methodologically legitimate form of analysis, but its relevance to conservatory assessment may remain limited. Such systems primarily model typicality: they estimate how expressive a given work's performance tends to be across a training corpus. By contrast, Conservatoire assessment often focuses on the artistic distinctiveness of a particular performer's interpretation of a specific work within a particular performance context. Such judgments are informed by critical frameworks developed through sustained engagement with repertoire, stylistic traditions, performance practice, and pedagogical experience. Corpus typicality and artistic individuality may therefore be related, but non-equivalent constructs, and the distance between them may not necessarily diminish with additional training data. In evaluating a graduating recital, faculty members are unlikely to compare the performance solely to a statistical norm. Their judgment may also attend to the performer's relationship with the work, the relevant interpretive tradition, and the communicative demands of the performance situation. These relational dimensions are not readily reducible to acoustic features alone. Expanding the training corpus may improve the representation of observable performance patterns, but it may not fully capture the interpretive relationships through which artistic individuality is recognized

and evaluated.

6. Transparency: Scoring procedures and the right to contest

Stiggins (1991) identifies clear communication as a fundamental criterion of assessment literacy, requiring that students, teachers, and institutions understand how assessments work and be able to contest results that appear unjust. This transparency requirement has a specific application in AI assessment, in which algorithmic procedures by which scores are generated must be sufficiently visible so that their assumptions can be examined and their outputs challenged.

The transparency problem has both an algorithmic and a corpus dimension. At the algorithmic level, Pasquale's (2015) account of the black box society identifies the strategic opacity of commercial algorithmic systems as a structural feature of contemporary computation, in which the proprietary status of code becomes a barrier to democratic accountability. At the corpus level, Crawford & Paglen (2021) excavating AI demonstrated that the training datasets of widely deployed visual recognition systems encoded specific, contestable value choices about classification, taxonomy, and human categorization, choices that became invisible once the systems were in operation. AI music assessment tools are subject to both forms of opacity. The algorithmic logic by which scores are generated is rarely disclosed, and the demographic, stylistic, and repertoire composition of training corpora is virtually never documented. As a result, the score returned by an AI music assessment tool is, in practice, the output of a process that no one outside the developing organization can reconstruct, evaluate, or contest.

In the music assessment domain, this dual opacity has direct practical consequences. If a student's performance of a Chopin nocturne receives a low score from an AI system, the student and teacher need to know what features drove the score. Against what reference was the performance compared? What was the training corpus from which the model's norms were derived? What stylistic, demographic, and repertoire assumptions are encoded in the corpus? Without access to such information, the score may remain difficult to interpret, challenge, or translate into meaningful pedagogical action.

Reading Feenberg's critical theory of technology, this opacity should not be understood as a mere accident of commercial practice. Rather, it is a structural feature of the technical code embodied in these assessment systems. Assessment literacy, therefore, requires naming this asymmetry as an institutional problem, not treating it as a technical inconvenience. It also requires conservatoires and higher-education institutions to develop procurement criteria in which

transparency is a condition of adoption.

This demand for transparency now has regulatory force in at least one major jurisdiction. When used in a European conservatory or higher-education context, every commercial AI assessment platform identified earlier in this section falls under this regulatory language. This classification imposes obligations on both the provider and the deploying institution, including risk management, transparency, human oversight, and post-market monitoring. Yet no commercial music-assessment vendor included in this synthesis publishes a model card, a demographic disclosure of training data, or a fairness audit consistent with these obligations. This gap is reinforced by Manganello et al. (2025) governance-checklist study of Italian universities, which found that only 23 percent of surveyed higher-education institutions had AI acceptable-use policies. An earlier six-country European survey likewise reported that approximately 80 percent of institutions lacked AI policies altogether.

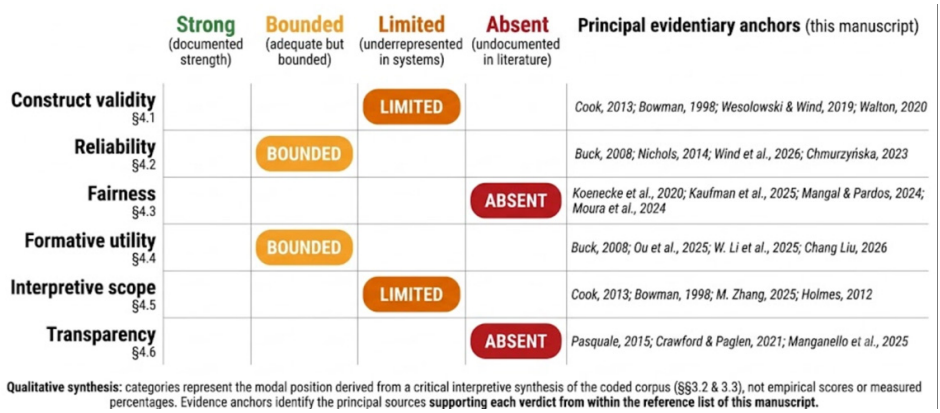
The problem, then, is not anecdotal but institutional. The Assessment Literacy Audit Protocol developed in this article should therefore be understood not only as a pedagogical instrument but also as a partial compliance instrument. By applying it, institutions generate much of the documentation required by the emerging regulatory regime and strengthen the epistemic and ethical defensibility of AI-supported music assessment.

Ⅴ. Discussion

1. The central finding

The six-dimensional analysis reveals a consistent pattern, summarized as a qualitative synthesis matrix in Figure 2. The asymmetry is not incidental; it reflects the fundamental design logic of AI music assessment systems, which are engineered to optimize measurable proxies for performance quality while leaving the construct dimensions that conservatoire pedagogy privileges most highly (expressive coherence, interpretive individuality, timbral quality, stylistic authenticity) outside the measurement model.

Figure 2 presents a qualitative synthesis rather than a quantitative measurement. The four category positions (documented strength, adequate within a bounded construct, limited or underrepresented in current systems, absent or undocumented) represent the modal position across the coded corpus for each of the six ALAP dimensions. These positions were derived through the four-step



[Figure 2] Qualitative synthesis of the six-dimensional ALAP analysis: Modal verdicts and principal evidentiary anchors across peer-reviewed and technical documentation sources

analytical procedure described in Section Ⅲ.2 and applied to the source categories specified in Section Ⅲ.3. Category assignment for each dimension followed this procedure. First, each of the 52 sources in the analytical corpus was coded for its stance on the dimension in question, where a stance was expressed. Second, the coded stances were aggregated at the dimension level and inspected for convergence and divergence. Third, the modal position (i.e., the stance shared by the largest proportion of coding sources for that dimension) was identified. Fourth, the modal position was assigned to one of the four categorical labels according to the following rules: Strong, where the modal position was supported by multiple peer-reviewed empirical or technical sources and no contradictory peer-reviewed source was located; Bounded, where the modal position supported the dimensional claim for a narrow construct (e.g., pitch and rhythm accuracy) but not for the full conservatoire construct; Limited, where multiple sources documented specific gaps or underrepresentation of the dimension in current systems without excluding future development; and Absent, where no located peer-reviewed source documented the dimension for the commercial platforms surveyed. The principal evidentiary anchor reported for each row identifies the sources whose evidence was most consequential for the categorical assignment. The verdict columns in Tables 1, 3, and 4 were derived through the same procedure, and the reader should treat them as qualitative categorical judgments grounded in the coded literature rather than as measurements on an interval scale.

2. Theoretical interpretation: assessment-illiterate instruments

Scaffolded by the three-layer philosophical foundation, the assessment literacy framework enables a more precise interpretation of this finding. AI performance assessment tools, in their current design and application, may be understood as assessment-illiterate in the broader evaluative sense of the term. This characterization does not imply an absence of measurement capability. For technically bounded parameters, their procedures may be sophisticated, consistent, and operationally effective. The limitation instead concerns their capacity to account for the purposes of assessment, the constructs being evaluated, the interests served, and the relationship between algorithmic outputs and the learning objectives of conservatoire performance education.

From this perspective, the problem may lie less in the tools as software than in the assessment-design processes through which they are developed and the institutional procedures through which they are adopted. Where these processes do not sufficiently examine construct validity, educational purpose, interpretative scope, fairness, and the meaning of generated scores, the resulting system may remain technically competent while being evaluatively underdeveloped.

This finding extends the theoretical framework in a specific direction. Assessment literacy scholarship has primarily focused on human assessors, including teachers, adjudicators, and institutional policy-makers, as both the objects of analysis and the targets of capacity-building. The present study proposes extending assessment literacy from the evaluation of assessment practitioners to the evaluation of assessment instruments themselves.

From this perspective, any instrument used within an educational assessment context, particularly in conservatoire performance assessment, whether human-mediated or computational, may be examined against comparable criteria of construct validity, educational purpose, fairness, interpretative adequacy, and consequential use. The automated, scalable, and institutionally authoritative character of AI-based instruments may make such scrutiny particularly important, because their outputs can circulate rapidly and acquire legitimacy across assessment systems.

The hermeneutic-aesthetic, post-phenomenological, and critical-technology layers help explain why this extension is not merely procedural but also ontologically and epistemologically significant. AI instruments may reshape the construct being assessed, mediate the perceptual conditions under which performance is interpreted, and alter the institutional and political economy of assessment. These transformations point toward the need for evaluative criteria that extend beyond technical measurement and address the meanings, purposes, and consequences of assessment practice within conservatoire education.

3. Engagement with prior literature

The interpretation of this study contrasts with a strand of recent AI music education literature that presents AI assessment tools as relatively straightforward improvements over human adjudication because of their apparent reliability advantage (Phanichraksaphong & Tsai, 2023). Such studies are justified in identifying the consistent measurement of technically bounded parameters as a meaningful benefit. However, the study fails to treat this gain as sufficient evidence of assessment quality without examining the validity, fairness, formative utility, and transparency dimensions.

A highly reliable instrument may still provide limited educational value when it measures only a partial or misaligned representation of the intended construct. Reliability therefore cannot be treated as a substitute for validity. An instrument that consistently measures the wrong construct is not necessarily superior to a less consistent instrument that engages more directly with the intended object of assessment; rather, it may reproduce its limitations with greater confidence and institutional authority

The finding is consonant, however, with the growing body of literature that identifies expressive and interpretive performance dimensions as beyond current AI reach (Ou et al., 2025; Zhang, 2025), and that documents the pedagogical risks of AI feedback for advanced students whose artistic development requires engagement with precisely the dimensions that current AI systems tend to underrepresent. It is also consistent with the equity literature's findings on algorithmic bias (Holmes, 2012; Kaufman et al., 2025; Koenecke et al., 2020; Mangal & Pardos, 2024), which collectively establish that the fairness deficit of AI assessment is not a peripheral concern but a structural feature of instruments trained on historically inequitable data. The convergence of these findings across general algorithmic bias research, audio-domain machine learning, and music education is striking, and it forms the empirical foundation on which any responsible deployment governance must be built.

4. The assessment literacy audit protocol in practice

The principal practical implication of this analysis is the need for a systematic instrument to evaluate AI performance assessment tools against assessment literacy criteria before institutional deployment. The Assessment Literacy Audit Protocol (ALAP), presented in Table 1, is offered as such an instrument. It is designed not as a rejection criterion (tools that fail some dimensions may still have legitimate, bounded uses) but as a deployment governance instrument that requires

institutions to be explicit about which assessment literacy criteria AI tools meet, which they do not, and what compensatory mechanisms are in place for dimensions the tools currently underrepresent. The architecture in Figure 1 makes the philosophical scaffolding visible: the ALAP is not a checklist but the operational outcome of a coherent foundation in the philosophy of musical interpretation, the philosophy of technology, and the philosophy of educational measurement.

5. Limitations

This analysis is bounded by its source corpus. The three-layer philosophical foundation, while internally coherent, is not the only philosophical foundation through which AI music tools could be evaluated; socio-material or ecological-dynamics perspectives would generate different but potentially complementary findings. The article does not compare specific commercial tools in head-to-head empirical analysis, which would require primary data and is therefore a task the future research agenda is designed to address. The ALAP itself is offered as a methodological proposal grounded in critical interpretive synthesis, not as an empirically validated instrument; institutional pilots will be required to test its operational adequacy.

Ⅵ. Conclusion

This article set out to examine the extent to which AI performance assessment tools satisfy the validity, reliability, fairness, formative utility, interpretive scope, and transparency criteria established by assessment literacy scholarship, and to propose a philosophically grounded methodological framework for evaluating their suitability for conservatoire practice. The analysis, conducted through a critical interpretive synthesis of peer-reviewed empirical, technical, and theoretical sources, demonstrates that current AI tools exhibit genuine strengths in measurement consistency for technically bounded parameters and in formative utility for beginner and early-intermediate learners, but tend to underrepresent construct validity for the full conservatoire construct, fairness disclosure, formative utility for advanced learners, interpretive scope, and transparency, that is, the dimensions most consequential for conservatoire assessment practice. The pattern reflects a documented tendency in the current generation of published AI music assessment systems to operationalize performance quality through measurable technical proxies, while interpretive, stylistic, expressive, and embodied dimensions remain underrepresented in both the construct model and

the documented validation base.

The theoretical contribution is threefold. The article extends assessment literacy scholarship from the analysis of human assessors to the evaluation of assessment instruments, proposing that the same evaluative criteria apply regardless of whether the assessor is a faculty member or an algorithm. It develops a three-layer philosophical foundation, integrating hermeneutic-aesthetic theory of musical interpretation, post-phenomenological and critical-technology theory of mediation, and educational measurement philosophy. This foundation prevents the assessment literacy criteria from collapsing into a checklist and grounds the evaluation in defensible accounts of what musical interpretation is, how AI mediates its perception, and what evaluative soundness requires. And it demonstrates that applying this foundation to AI performance assessment tools reveals a specific and actionable pattern: assessment-illiterate instruments, that is, technically sophisticated but evaluatively ungrounded, are being deployed in institutional contexts whose pedagogical values and learning goals they are structurally ill-suited to serve.

The practical contribution is the Assessment Literacy Audit Protocol (ALAP): a six-criterion governance instrument for evaluating AI performance assessment tools before institutional adoption. Its philosophical scaffolding and operational specifications are presented in Table 1 and Figure 1. The ALAP does not argue against AI tools in conservatoire education; it argues for the institutional and scholarly infrastructure necessary to deploy them responsibly.

The final sentence of this analysis is a critical proposition rather than a conclusion. If conservatory music education is to integrate artificial intelligence into performance assessment, it must first become fluent in three languages at once, namely the language of assessment literacy, the language of musical interpretation, and the language of technological mediation. Not because these languages are final, but because they are the only languages in which the question of whether a machine can hear what a teacher hears can be asked rigorously enough to merit an answer.

Three specific research directions emerge from this analysis. First, empirical implementation of the ALAP in conservatoire contexts: institutional pilots in which AI tools are evaluated against the six-criterion matrix by mixed panels of music educators, assessment specialists, and students, producing publicly available audit reports. Second, training data transparency research: studies that document and analyze the demographic, stylistic, and repertoire composition of AI music assessment training corpora, establishing whether and how training data inequities produce assessment inequities in deployment, with reference to the precedent established in audio-domain bias research (Koenecke et al., 2020). Third, participatory AI design research: co-design projects involving conservatoire students and faculty in the specification and validation of AI assessment

constructs, ensuring that the dimensions that conservatoire pedagogy values most highly are addressed in future tool development rather than treated as residual categories.

A fourth direction concerns the design of hybrid evaluation frameworks that combine algorithmic measurement of technical parameters with human expert judgment of expressive, interpretive, and stylistic dimensions. The pattern across the six-dimension analysis is not that AI music assessment tools should be rejected wholesale but that their bounded affordances should be paired with assessment practices in which the dimensions outside the model are retained as expert-judged. Real-time multimodal processing (audio synchronized with video and symbolic notation) remains a computational challenge, and the absence of a transparent industry standard for combining algorithmic and human scoring in conservatoire-grade examination is itself a gap the research agenda should address.

References

Adorno, T., & Leppert, R. (2002). *Essays on music*. Univ of California Press.

Álvarez-Díaz, M., Muñiz-Bascón, L. M., Soria-Alemany, A., Veintimilla-Bonet, A., & Fernández-Alonso, R. (2021). On the design and validation of a rubric for the evaluation of performance in a musical contest. *International Journal of Music Education*, *39*(1), 66-79. https://doi.org/10.1177/0255761420936443

Apple, M., & Apple, M. W. (2004). *Ideology and curriculum*. Routledge.

Araújo, D., & Davids, K. (2011). What exactly is acquired during skill acquisition? *Journal of Consciousness Studies*, *18*(3-4), 7-23.

Arora, A. (2025). SoundRubric: Musical performance analysis for violin learning using audio deconstruction. *2025 International Conference on Responsible, Generative and Explainable AI (ResGenXAI)*, 1-5. IEEE. https://doi.org/10/1109/ResgenXAI64788.2025.11344053

Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability*, *21*(1), 33-46. https://doi.org/10.1007/s11092-008-9064-9

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Bowman, W. D. (1998). *Philosophical perspectives on music*. Oxford University Press.

Bozkurt, B., Gulati, S., Romani Picas, O., & Serra, X. (2018). MusicCritic: A technological

framework to support online music teaching for large audiences. In *Proceedings of the 33rd World Conference on Music Education (ISME)* (pp. 13–20). International Society for Music Education.

Buck, M. W. (2008). *The efficacy of SmartMusic® assessment as a teaching and learning tool*. Doctoral dissertation, The University of Southern Mississippi.

Caruso, G. (2025). Biofeedback technology and artificial intelligence (AI) for music performance analysis and creation: Insights from an artistic practice as research. *Analitica. Rivista Online Di Studi Musicali, 17*(2024). https://doi.org/10.5281/zenodo.17176566

Chen, L. (2025). Unlocking the beat: how AI tools drive music students' motivation, engagement, creativity and learning success. *European Journal of Education*, *60*(1), e12823.

Chmurzyńska, M. (2023). Research on the assessment of music performance from Maria Manturzewska's early experiments to more recent studies. *Musicae Scientiae*, *27*(4), 862-874. https://doi.org/10.1177/10298649231188272

Clarke, E. (2005). *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford University Press.

Cook, N. (2013). *Beyond the score: Music as performance*. Oxford University Press.

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & Society, 36*(4), 1105-1116. https://doi.org/10.1007/s00146-021-01162-8

Cui, K. (2023). Artificial intelligence and creativity: piano teaching with augmented reality applications. *Interactive Learning Environments*, *31*(10), 7017-7028.

Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., Hsu, R., Katbamna, S., Olsen, R., & Smith, L. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology*, *6*(1), 35.

Djahwasi, H. R., Sampurno, M. B. T., & Nazeri, N. (2025). Expressive timing and tempo variance in six landmark recordings of Villa-Lobos' Etude No. 1: A sonic visualizer-based computational analysis. *Resital: Jurnal Seni Pertunjukan*, *26*(2), 360-401.

Djahwasi, H., Saidon, Z. L., Pongpradit, P., & Sampurno, M. B. T. (2023). Interpretation of technical and expression aspects on Mauro Giuliani's "Grand Overture, Op. 61." *Rast Müzikoloji Dergisi*, *11*(4), 563-587.

Eremenko, V., Morsi, A., Narang, J., & Serra, X. (2020). Performance Assessment Technologies for the Support of Musical Instrument Learning. *Proceedings of the 12th International Conference on Computer Supported Education CSEDU,* (1), 629-640.

Evin, M. (2024). A review on AI-enabled techniques for evaluating musician's performance.

AIP Conference Proceedings, *3149*(1), 140018.

Feenberg, A. (2002). *Transforming technology: A critical theory revisited*. Oxford University Press.

Fenwick, T., & Edwards, R. (2010). *Actor-network theory in education*. Routledge.

Ford, B. (2010). *What are conservatoires for? Discourses of purpose in the contemporary conservatoire*. Doctoral dissertation, University of London.

Gadamer, H.-G. (2004). *Truth and method* (J. Weinsheimer & D. G. Marshall, Trans.; 2nd rev. ed.). Continuum. (Original work published 1960)

Holmes, P. A. (2012). An exploration of musical communication through expressive use of timbre: The performer's perspective. *Psychology of Music*, *40*(3), 301-323.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth* (Number 560). Indiana University Press.

Kão, K., & Niitsoo, M. (2014). Matchmysound: Introducing feedback to online music education. *International Conference on Web-Based Learning*, 217-225. Springer International Publishing.

Kaufman, B., Kladder, J., & Palmer, E. S. (2025). Adjudicating Equity in Music Performance Assessments: A Literature Review. *Update: Applications of Research in Music Education*, *43*(3), 49-59. https://doi.org/10.1177/87551233241256288

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684-7689. https://doi.org/10.1073/pnas.1915768117

Levinson, J. (2015). *Musical concerns: Essays in philosophy of music*. OUP Oxford.

Li, P., & Wang, B. (2024). Artificial intelligence in music education. *International Journal of Human–Computer Interaction*, *40*(16), 4183-4192. https://doi.org/10.1080/10447318.2023.2209984

Li, W., Cui, X., Manoharan, P., Dai, L., Liu, K., & Huang, L. (2025). AI-assisted feedback and reflection in vocal music training: effects on metacognition and singing performance. *Frontiers in Psychology*, *16*, 1598867. https://doi.org/10.3389/fpsyg.2025.1598867

Liu, C. (2026). *A systematic review of artificial intelligence in piano education: Efficacy, ethics, and the transformation of the teacher's role (2020–2025)* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-8998934/v1

Lovett, M. (2024). Artificial creativity and tools for understanding: music, creative labour and AI. In J.-O. Gullö, R. Hepworth-Sawyer, J. Paterson, R. Toulson, & M. Marrington

(Eds.), *Innovation in music: technology and creativity* (pp. 267-277). Focal Press.

Madnani, N., & Cahill, A. (2018). Automated scoring: Beyond natural language processing. *Proceedings of the 27th International Conference on Computational Linguistics*, 1099–1109. Association for Computational Linguistics.

Mangal, M., & Pardos, Z. A. (2024). Implementing equitable and intersectionality-aware ML in education: A practical guide. *British Journal of Educational Technology*, *55*(5), 2003-2038.

Manganello, F., Nico, A., Ragusa, M., & Boccuzzi, G. (2025). Testing the applicability of a governance checklist for high-risk AI-based learning outcome assessment in Italian universities under the EU AI act annex III. *Frontiers in Artificial Intelligence*, *8*, 1718613.

Merchán Sánchez-Jara, J. F., González Gutiérrez, S., Cruz Rodríguez, J., & Syroyid Syroyid, B. (2024). Artificial intelligence-assisted music education: A critical synthesis of challenges and opportunities. *Education Sciences*, *14*(11), 1171. https://doi.org/10.3390/educsci14111171

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). American Council on education and Macmillan.

Moura, N., Dias, P., Veríssimo, L., Oliveira-Silva, P., & Serra, S. (2024). Solo music performance assessment criteria: a systematic review. *Frontiers in Psychology*, *15*, 1467434. https://doi.org/10.3389/fpsyg.2024.1467434

Nichols, B. D. (2014). *The effect of SmartMusic on student practice*. Doctoral dissertation, Kennesaw State University. https://digitalcommons.kennesaw.edu/teachleaddoc_etd/1

Ou, J., Nogueira, J., & Qin, C. (2025). Exploring the impact of AI-assisted practice applications on music learners' performance, self-efficacy, and self-regulated learning. *Frontiers in Psychology*, *16*, 1675762.

Ouyang, D. (2025). Voice, vulnerability, and expressive growth: investigating AI anxiety and performance appraisal in voice arts education. *Frontiers in Psychology*, *16*, 1653502.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. In *The black box society*. Harvard University Press.

Pastore, S. (2023). Teacher assessment literacy: a systematic review. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1217167

Phanichraksaphong, V., & Tsai, W.-H. (2023). Automatic assessment of piano performances using timbre and pitch features. *Electronics*, *12*(8), 1791. https://doi.org/10.3390/electronics12081791

Popham, W. J. (2018). *Assessment literacy for educators in a hurry*. ASCD.

Preda-Uliță, A. (2025). AI and metacognition in instrumental performance: An integrative model for critical thinking/Inteligența artificială și metacogniția în interpretarea instrumentală: Un model de integrare a gândirii critice. *ICT in Musical Field*, *16*(1), 41.

Rizal, H., & Milyartini, R. (2024). Improving the quality of music education through applications based on Artificial Intelligence (AI). *SHS Web of Conferences*, *197*, 01004.

Said, E. W. (1991). *Musical elaborations*. Columbia University Press.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*(7), 534-539.

Tao, C. (2026). The impact of AI-based educational applications on university students' piano skills and self-efficacy. *International Journal of Psychology*, *61*(2), e70167.

Tian, J. (2025). Intelligent information teaching method and its influence on instrumental music teaching. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, *20*(1), 1-22.

Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Penn State Press.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago, IL: University of Chicago press.

Walton, J. C. (2020). *Making the grade: Theorising musical performance assessment*. Doctoral dissertation, Griffith University. https://doi.org/10.25904/1912/2485

Wei, L., Qiu, L., & Man, S. (2025). The Dual Impact of AI Adoption on College Students' Academic Performance in Music Education: Evidence from China. *Sage Open*, *15*(4), 21582440251407504.

Wesolowski, B. C., & Wind, S. A. (2019). Validity, Reliability, and Fairness in Music Testing. In T. S. Brophy (Ed.), *The Oxford handbook of assessment policy and practice in music education* (pp. 437-460). Oxford University Press.

Wind, S. A., Wesolowski, B. C., & Jones, E. (2026). Rater connectedness affects student achievement estimates and ordered rankings in formal music performance assessments. *Journal of Research in Music Education*, *74*(1), 110-131. https://doi.org/10.1177/00224294251363327

Winner, L. (2017). Do artifacts have politics? In J. Weckert (Ed.), *Computer ethics* (pp. 177-192). Routledge.

Yoon, M. (2020). Trends of assessment research in music education. *Korean Journal of Research in Music Education*, *49*(4), 153-173.

Zhang, L.-X., Yan, Z., & Wang, X. (2026). Mapping formative assessment in higher music education: A scoping review of its implementation and impact on musical intelligence. *Research Studies in Music Education*, *48*(1), 65-86. https://doi.org/10.1177/1321103X251400571.

Zhang, M. (2025). Advancing deep learning for expressive music composition and performance modeling. *Scientific Reports*, *15*(1), 28007.