

학교 음악 연주 수행평가에 나타난 인간 평가의 맥락적 특성과 AI기반 평가 모델 설계의 고려 요소*

Contextual Characteristics of Human Assessment in School Music Performance and Design Considerations for AI-based Assessment Models

정주연** · 강주현*** · 신지혜****

Joo Yeon Jung^{ID} · Joo Hyun Kang^{ID} · Jihae Shin^{ID}

Abstract This study examines the characteristics of human evaluation in secondary school music performance assessments and explores how these school-based tendencies can inform the development of AI algorithmic models. Although prior research has noted the influence of intuitive and holistic impressions in music performance evaluation, such insights have rarely been connected to the realities of assessing non-major middle school students in authentic school settings. In this study, music-major evaluators assessed actual middle-school vocal and instrumental performance recordings collected during the development of an AI assessment platform, and subsequently participated in in-depth interviews with AI developers. The analysis showed that evaluators placed particular importance on students' sincerity, effort, and engagement, and that expressive elements and pitch accuracy required flexible, context-dependent interpretation rather than strict precision. These tendencies highlighted the difficulty of converting school-based evaluative practices into forms that AI can learn reliably. By situating human evaluation patterns within the contextual and pedagogical aims of school environments, this study offers insight into how AI systems can be designed to align with the performance characteristics of non-major learners and emphasizes the need for human-informed, context-sensitive AI models in music education.

Key words: middle school music performance assessment, school music education assessment, human evaluation, AI algorithmic model design, AI-based music assessment

초록 본 연구는 중등 음악 수행평가에서 나타나는 인간 평가의 특성을 검토하고, 학교 현장에서의 평가 경향이 AI 알고리즘 모델 개발에 어떤 함의를 제공하는지 탐색한다. 선행연구에서는 연주에 대한 직관적, 총괄적 인상이 음악 연주 평가에서 나타난다고 보고해 왔으나, 이러한 발견이 비전공 중학생을 평가하는 실제 학교 맥락과 연결되어 논의된 사례는 드물다. 본 연구에서는 AI 연주 평가 플랫폼 개발 과정에서 실제 중학교 가창 및 기악 수행평가 녹음 자료를 음악전공 평가자들이 평가하도록 하였으며, 이후 평가자와 AI 개발자를 대상으로 심층 인터뷰를 실시하였다. 분석 결과, 평가자들은 학생의 성실성·노력·참여도를 특히 중요하게 고려하였으며, 표현적 요소와 음고 정확성은 엄격하고 정밀한 기준보다는 맥락에 따라 유연하게 해석될 필요가 있음을 확인하였다. 이러한 인간 평가의 경향은 학교 현장의 평가 특성을 AI가 학습할 수 있는 형태로 구현하는 데 어려움이 있음을 시사하였다. 본 연구는 인간 평가 경향을 학교 환경의 맥락과 교육적 목표 속에서 재해석함으로써, 비전공 중학생의 실제 수행 특성과 학교 음악 교육에 적합한 AI 시스템 설계에 새로운 통찰을 제공하며, 음악교육에서 인간 기반·맥락 친화적 AI 모델의 필요성을 강조한다.

주제어: 중학교 음악 연주 평가, 학교 음악교육 평가, 인간 평가, AI 알고리즘 모델 설계, AI 기반 음악평가

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2023S1A5A2A03084082)

** First author, E-mail: jung.jooyeon@ginue.ac.kr

Assistant Professor, Gyeongin National University of Education, Department of Music Education, 155 Sammak-ro, Manan-gu, Anyang-si, Gyeonggi-do, 13910, South Korea

*** Corresponding author, E-mail: jhk1011@gmail.com

Lecturer, Korean University of Education, Department of Music Education, 250 Taeseongtabyeon-ro, Gangnae-myeon, Heungdeok-gu, Cheongju-si, Chungcheongbuk-do, South Korea

**** Co-author, E-mail: jhshin99@ewha.ac.kr

Associate Professor, Ewha Womans University, Department of Music Education, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea

I . Introduction

Music performance assessment plays a critical role in school music education, offering students meaningful feedback on their musical growth and informing instructional decisions (Moura et al., 2024; Stanley et al., 2002). Effective assessment in music education provides specific feedback about students' current performance levels while offering clear guidance for improvement, which enables educators to deliver more targeted and constructive feedback that promotes individual musical growth and development (Moura et al., 2024). Teachers also could continuously analyze student assessment results to understand what students have learned and adjust their teaching strategies accordingly as an essential part of effective music education (Pellegrino et al., 2015).

Despite its educational value, human evaluation of student performances often faces issues related to subjectivity, holistic versus specific criteria, and the lack of support. While the aesthetic nature of performance assessment inevitably involves an element of tacit, intuitive subjectivity that cannot be fully captured by verbal or numerical rating scales, this subjectivity—though valuable—can contribute to inconsistent feedback and highlights the need to balance it with more objective assessment methods (Wrigley & Emmerson, 2011). Previous studies also found that while examiners view assessment criteria as helpful for guiding evaluations and providing student feedback, they also express concern that such criteria can restrict holistic judgment and promote a narrow perspective on music performance (Moura et al., 2024; Stanley et al., 2002). Additionally, administrative supports—such as assessment guidelines and access to resources—are closely related to music teachers' confidence in assessment. When teachers receive minimal support or guidance regarding their assessment practices, they tend to avoid assessing affective domains such as student attitude and engagement because of concerns about the subjectivity of such assessments (Russell & Austin, 2010; Wesolowski et al., 2017).

With recent advances in artificial intelligence (AI), AI-based assessment tools have emerged as promising alternatives or supplements to traditional human evaluations in music (Wei, Karuppiiah, & Prathik, 2022). Machine learning algorithms have demonstrated remarkable capabilities in analyzing musical performances based on technical aspects such as pitch accuracy, rhythm precision, and tonal quality (Evin, 2024; Giraldo et al., 2019; Pati, Gururani, & Lerch, 2018). In fact, educational technology companies, including MakeMusic and MusicFirst, have begun developing AI-powered platforms that can provide immediate, consistent feedback on student performances. These technological developments have sparked considerable interest among educators and researchers, who recognize the potential for AI to enhance both the efficiency

and objectivity of music assessment while preserving the pedagogical benefits of performance evaluation (Gurley, 2012; Tucker, 2016).

While significant efforts have been dedicated to refining music performance assessment approaches, particularly in terms of technical and expressive evaluation criteria, these developments have largely overlooked the integration of technology, such as AI, into educational settings. Numerous studies have focused on constructing and validating assessment rubrics and rating scales grounded in human judgment, emphasizing dimensions like technique, tone, rhythm, and interpretation across various instruments and proficiency levels (e.g., Abeles, 1973; Bergee, 2003; Thompson & Williamon, 2003; Russell, 2010; Wesolowski et al., 2017). Such instruments are also often tailored to advanced learners, such as high school or university-level students, with expert musicians or faculty serving as evaluators. However, very few studies have addressed how assessment practices might be adapted to younger learners or to the specific developmental needs of secondary school students, where ongoing feedback and encouragement for improvement is critical (Moura et al., 2024).

To address this gap, this study investigates two key questions related to the perspectives of humans and AI developers on secondary school music performance assessment. First, what are the characteristics of music teachers' assessment tendencies in the evaluation process of secondary school students' musical performances? Second, what are AI assessment platform developers' perceptions of human evaluation in the assessment process of secondary school students' musical performances? By examining these interconnected questions, this study aims to provide a comprehensive understanding of the current state of music performance assessment and the feasibility and implications of integrating AI technologies in secondary school music education contexts.

In this research, school contexts are defined as environments based on the Korean National Music Curriculum (Ministry of Education, 2022), which prioritizes holistic growth and the cultivation of future competencies, such as self-directedness, rather than the pursuit of professional musical excellence. The educational goals focus on guiding students to engage in comprehensive musical activities and to integrate music into their daily lives. Consequently, the purpose of assessment is to evaluate both the students' achievement and their learning process as they develop musically. Within this framework, the teacher's role is to design instruction based on the national curriculum and to facilitate this development by evaluating student progress to ensure the fundamental objectives of music education are met.

II. Literature Review

1. The characteristics of human music performance assessment

Music performance assessment involves a complex process in which both technical skills and affective aspects are evaluated. Understanding how music experts assess music performance is therefore crucial for refining music performance assessment approaches. Recognizing this complexity, Stanley et al. (2002) examined how examiners perceived criteria for music performance assessment by analyzing the stages of their evaluation process. This study showed that at the beginning of the assessment, experts focused on global assessment, known as a holistic approach to performance. During the second step, however, these examiners tried to identify performance characteristics which might support their decisions. At the final stage of assessment, they considered specific assessment criteria (e.g., technical accuracy, tone quality) to finally check their decisions. This study concluded that these stages may vary across examiners, highlighting that expert evaluation is inherently multidimensional and not strictly sequential, but rather grounded in an overall impression in which multiple performance elements interact in complex and interdependent ways.

Similarly, Russell and Austin (2010) found that school-based performance assessments also extend beyond strictly musical achievement. In practice, teachers incorporated both achievement-related factors, such as performance skills and musical knowledge, and nonachievement factors including attendance, attitude, and practice charts into students' grades, with many teachers placing greater emphasis on nonachievement rather than achievement criteria. Taken together, these studies underscore that music performance assessment—whether conducted by expert judge or classroom teachers—involves synthesizing diverse types of information that are not always easily quantifiable. This convergence suggests that evaluating musical performance is fundamentally a complex, multilayered process that requires careful consideration when developing fair and systematic assessment practices.

2. The characteristics of AI-based music assessment

Representative examples of AI-based music assessment software are MakeMusic and PracticeFirst. In this section, several functions of these software programs will be introduced.

These programs have thousands of musical works for different instruments including wind,

strings, and voice. While students play their music along with recorded accompaniments, the programs assess students' rhythm (duration) and pitch (intonation). The difference between these two programs is how they display their assessment. Both programs use red, green, and yellow colored notes to show whether notes are correct or incorrect, but MakeMusic uses real notes on a music sheet while PracticeFirst focuses on graphical notation with colors below traditional music notation. An important characteristic of these programs is that students can adjust their level of assessment rigor, from easy to hard, based on their performance level. All of this feedback from the programs are immediate and real-time.

After practice, students can get feedback based on a percentage score for not only overall practice but also detailed numbers for pitch and rhythm. These practice records are shared with the teacher, which provides detailed information about individual students' practice processes. Also, teachers can provide students with additional practice tasks based on the programs' feedback. All these programs are web-based, which means they can be used on computers, phones, and tablets.

Overall, current AI-based music assessment systems primarily focus on judging pitch and rhythmic accuracy, as these aspects can be readily quantified and processed through computational analysis. This design reflects the fundamentally rule-based nature of such systems, in which a student's performance is compared against predefined pitch and timing specifications extracted from the musical score—that is, fixed, codified rules that the computer can evaluate numerically.

III. Method

1. Research overview and purpose

The purpose of this study was to provide practical insights into the development of an AI-based music performance assessment tool that can be utilized by both secondary music teachers and students in schools. In this regard, interviews with music performance evaluators and AI developers were conducted regarding their perspectives, system goals, and challenges they faced during the development.

To develop a platform that reflects the assessment practices commonly used in schools, performance data from secondary school students was collected and evaluated by music performance evaluators using a rubric modeled after those employed in school settings. Based

on these human-generated evaluation results, the AI was then trained to learn and emulate human-like judgment using the numerical scores produced through this process. However, this process revealed limitations in the direct application of the existing rubric's scoring method and evaluation criteria to AI learning, leading to this research.

This demonstrated that human judgments of music performance are inherently complex and multifaceted aligning with what previous studies have shown. Because these classroom-based judgments are multifaceted and often implicit, the evaluation criteria had to be elaborated and operationalized so that they could be quantified for AI learning. To achieve this, the evaluation criteria and process were revised several times and more sophisticated and specific guidelines were developed. This process confirmed the essential need for ongoing collaboration between music experts and AI developers.

2. In-depth interview participants: Interview participants and profiles

To gain deeper insight into how human evaluators form judgments, in-depth interviews were conducted with two distinct groups: music performance evaluators and AI algorithm developers. The three music evaluators (A, B, C) were graduate students with undergraduate degrees in music—two in violin and one in composition. They were pre-service teachers in a secondary music teacher preparation program and worked as research assistants, evaluating student performance recordings. Because they were not the students' actual teachers, they were able to give more objective assessments. Additionally, since secondary music teachers in Korea are systematically required to establish their musical expertise through specific majors during their teacher education, the evaluative perspectives of the pre-service teachers in this study are fundamentally consistent with the professional standards of in-service teachers.

The AI algorithm was developed by a team of expert researchers. Two of them(D, E), who were the interviewees of this study, worked as research assistants responsible for data management and analysis—one was a master's student and the other a doctoral student specializing in artificial intelligence. Their undergraduate backgrounds were in Western music composition and Korean traditional music (daegeum). The brief information is as follows.

<Table 1> Interviewees’ profiles

Participants			
Group	Participant’s indicator	Academic background	Role in the project
Music performance evaluators	A	B.A. in Music (violin)	Evaluated recordings; research assistants
	B	B.A. in Music (violin)	
	C	B.A. in Music (western music composition)	
AI developers	D	B.A. in Music (Daegeum)	System design, algorithm development, data modeling
	E	B.A. in Music (western music composition)	

3. Pre-interview activities: Evaluation of student performance recordings

The evaluation of recorded student performances constituted the first phase of the study, and the interviews that followed were designed to examine the judgment processes exhibited during that phase. The dataset consisted of unedited audio recordings from actual middle school music performance assessments, which served as foundational data for algorithm development. The collected recordings were divided into vocal and instrumental categories (see Table 2).

<Table 2> Types of performances

Category	Types
Vocal	Western classical vocal music, Korean traditional vocal music
Instrumental	Recorder, Danso (Korean bamboo flute)

Three music-major research assistants evaluated the collected recordings in two rounds. These evaluations provided insight into how human evaluators form judgments and apply the evaluation criteria.

1) First-round

Music performance evaluators assigned an overall rating of high, medium, or low based on their general impression of each student performance. They also evaluated specific musical elements—including phrasing (breathing), rhythm accuracy, and pitch accuracy—using the same three-level scale. The music performance evaluators rated each performance and specific musical

element according to the high-medium-low criteria but did not calculate total scores. The detailed ratings were later converted into numerical values and summed during the analysis phase. This process was led by AI developers to generate structured data for algorithm training.

This analysis revealed a notable discrepancy between the sum of detailed scores and the overall ratings. In many cases, the combined detailed scores did not align with evaluators' holistic judgments based on intuitive impression. This inconsistency highlighted limitations in the first evaluation structure and indicated that a second evaluation with more refined criteria was necessary to develop an appropriate assessment framework for the AI platform.

2) Second-round

Following the first evaluation, music performance evaluators and researchers conducted in-depth discussions about the factors influencing overall assessments and concluded that a more detailed evaluation rubric was necessary. The music performance evaluators and researchers reviewed the elements that had been implicitly reflected in the overall scores during the first evaluation and developed a second evaluation rubric based on the 2022 revised national music curriculum. Key evaluation criteria included the ability to understand and express musical characteristics, proper posture and breathing, and expression appropriate to genre and instrument characteristics. Elements used in the first evaluation were reorganized to align with curriculum terminology and were divided according to the characteristics of vocal and instrumental music, incorporating components such as phrasing, rhythm, pitch, and *shigimsae* (ornamental techniques in Korean traditional music). This aimed to establish evaluation criteria that aligned with actual school music class objectives while creating a structure suitable for AI learning.

The revised rubric was refined to more accurately assess various performance types. Previously combined items, such as 'phrasing (breathing),' were separated into distinct criteria to allow for more precise and domain-appropriate judgments. For vocal performances, additional criteria for breath control, tone quality, and vocalization techniques were introduced. For instrumental performances, instrument-specific characteristics and appropriate playing techniques were added, including technical execution elements such as those for *danso*. An attitude category was incorporated to assess performers' sincerity, focus, and consistency, enabling evaluation not only of musical outcomes but also of affective aspects such as learning attitude. Each evaluation item—pitch accuracy, rhythm, tempo, phrasing, *shigimsae*, pronunciation and lyrics delivery, breathing, and other performance-type-specific elements—was rated on a three-level scale (high,

medium, low) and converted into numerical scores, with the scoring structure calibrated so that category scores corresponded appropriately to overall ratings. Moreover, ‘attitude’ category was added to evaluate performers' approach alongside musical outcomes. Attitude assessment focused on whether students performed diligently from beginning to end and maintained focus. Students were evaluated based on four levels: completing with focused attitude, completing despite occasional distractions, failing to complete due to insufficient effort, and showing effort but unable to finish. Table 3 and 4 presents example rubrics for Korean traditional singing and recorder performance.

<Table 3> Rubric for Korean traditional singing

Criteria	High (18pts)	Medium (15pts)	Low(9pts)	Notes(Remarks)
Phrasing	Consistently observes breath marks with accurate phrasing (90-100%).	Generally observes breath marks with mostly accurate phrasing (70-89%).	Inconsistent breath mark observation; phrasing needs development (below 70%).	Scored by measure (errors within a measure result in deduction); percentage thresholds adjustable by class level; criteria vary by piece-specific <i>shigimsae</i> count.
<i>shigimsae</i>	Accurately performs <i>shigimsae</i> (tremolo and pitch bends) in 90-100% of occurrences.	Performs <i>shigimsae</i> with partial accuracy (70-89%).	Performs few <i>shigimsae</i> correctly (below 70%).	
Vocalization (vocal production)	Demonstrates appropriate vocalization for the piece in most or all measures (90-100%).	Demonstrates appropriate vocalization for the piece in most measures (75-89%).	Inconsistent or inappropriate vocalization (below 75%).	
Pronunciation	Sings with accurate pronunciation in most or all measures (90-100%).	Sings with accurate pronunciation in most measures (75-89%).	Pronunciation needs improvement (below 75%).	
Pitch	Sings with accurate pitch in most or all measures (90-100%).	Sings with accurate pitch in most measures (75-89%).	Pitch accuracy needs improvement (below 75%).	
Rhythm Korean rhythmic pattern	Sings with accurate rhythm in most or all measures (90-100%).	Sings with accurate rhythm in most measures (75-89%).	Rhythm accuracy needs improvement (below 75%).	

<Table 4> Rubric for recorder performance

Criteria	High (18pts)	Medium (15pts)	Low (9pts)	Notes
Phrasing	Performs with expressive phrasing in most or all measures (90-100%).	Performs with expressive phrasing in most measures (75-89%).	Phrasing needs improvement (below 75%).	Scored by measure (errors within a measure result in deduction); percentage thresholds adjustable by class level;
Breathing	Takes breaths at appropriate places in most or all measures (90-100%).	Takes breaths at appropriate places in most measures (75-89%).	Breath placement needs improvement (below 75%).	
Articulation	Applies appropriate tonguing technique for articulation in most or all measures (90-100%).	Applies appropriate tonguing technique for articulation in most measures (75-89%).	Tonguing technique needs improvement (below 75%).	
Rhythm	Plays with accurate rhythm in most or all measures (90-100%).	Plays with accurate rhythm in most measures (75-89%).	Rhythm accuracy needs improvement (below 75%).	
Pitch	Plays with accurate pitch in most or all measures (90-100%).	Plays with accurate pitch in most measures (75-89%).	Pitch accuracy needs improvement (below 75%).	

4. Interview data analysis

All interviews were conducted online via Zoom as video-based, semi-structured sessions lasting approximately two hours each. Every session was recorded and transcribed immediately afterward, with attention to preserving contextual meaning, and the transcripts were subsequently analyzed using thematic and qualitative methods. Any questions or points requiring clarification were followed up after the interviews.

Data analysis was conducted through a systematic thematic analysis process. The researchers repeatedly reviewed the transcripts to identify meaningful content units, which were then categorized through an inductive coding process. These emergent themes were refined and grouped into categories aligned with the research questions. Specifically, the coding system for human evaluation characteristics evolved from *Holistic and Flexible Judgment* from the first round of the evaluation process and *Structured and Context-Sensitive Judgment* from the second round. Simultaneously, developers’ perceptions were analyzed through themes such as *Technical Challenges in AI Algorithm Development* and *Developing Human-Informed AI Models for Classroom*

Evaluation. To ensure the trustworthiness and consistency of the analysis, researchers conducted collaborative cross-checking. In cases of conceptual ambiguity, iterative discussions were held until a final consensus on the theme coding system was reached. This process allowed for a deep exploration of school-based music assessment characteristics and their implications for the development of AI-based evaluation systems.

IV. Results

1. Characteristics of human evaluation

1) First round: Holistic and flexible judgment

Based on the analysis of interviews with three evaluators who participated in the evaluation of secondary school students' musical performances, several key characteristics of human evaluation emerged. In particular, during the first round of evaluation, student performances were assessed using categorical judgments based on specific musical elements—such as rhythm, pitch, phrasing (breathing), and articulation or ornamentation depending on the instrument—and an overall three-level rating (high-medium-low). For each category, evaluators considered how many measures or *jeonggan* (rhythmic units in Korean music) were performed in accordance with the musical criteria. Based on this, they also assigned a high, medium, or low rating per element.

Meanwhile, analysis of the evaluation results revealed that the overall ratings often did not align with the combination of category ratings. For instance, a student who received mostly 'low' marks across the categories might still be rated 'high' overall if the evaluator perceived a particularly compelling quality in the performance.

C: At first, we listened to the audio and scored it according to the overall high-middle-low scoring rubric. And when dividing it into high, middle, and low, even if the student's score was low, if there was something good in what we heard, we could still give them a 'high' or 'middle' rating—basically, the teacher, meaning the evaluator, could assign the level at their own discretion. That's how we did it at the beginning.

This illustrates the gap between the accuracy of individual musical elements and the evaluator's

holistic sense of overall performance quality. For AI models to be trained effectively, they must rely on quantifiable data and on a coherent mapping between subcategory scores and overall ratings. However, the evaluation patterns observed here reveal that this alignment is often absent. This mismatch raises the question of what evaluative dimensions—beyond the numerically defined musical elements—are influencing teachers' holistic judgments, making it essential to identify and model these factors.

Through the interview, we found that the evaluators frequently prioritized the overall musical impression—including tone quality, energy and fluency—even when these were not explicitly defined in the rubric. The three-level overall rating was used flexibly, allowing room for professional judgment shaped by how the performance sounded as a whole. In addition, evaluators attended to auditory cues that signaled performance-related behavior—such as confidence, hesitation, focus, or even whether the student was able to complete the piece.

C: We tried to include aspects like the students' attitude. ... So even if two students had the same score, if, when listening to the audio, a student sounded insincere or was chatting in between, we tried to differentiate their rating within the high-middle-low scale.

A: If a student seemed to give up or conveyed a nuance like 'I can't do this anymore,' we reflected that in the rating. Some students also hesitated a lot in the beginning, dragged out the time, or took breaks in the middle, and we considered that as well.

Students who performed with steady tone and clear musical intention were often rated more favorably than those whose playing suggested uncertainty, disengagement, or abrupt discontinuation. Such evaluative tendencies may be particularly evident in the context of assessing middle-school non-music majors, where school performance prioritizes participation, effort and musical engagement over technical completeness or musical expressiveness. This suggests that evaluators considered not only musical accuracy but also the degree of engagement and conveyed through sound.

Furthermore, a tendency toward relative judgment was observed when performances were reviewed in sequence.

C: But since we were thinking only in terms of high-middle-low, I think the students who came before and after ended up influencing the rating quite a bit. For example, I would feel something like, 'This student didn't do as well as the one before,' and even if it wasn't the student right before or right after, I'd recall that I had given at least a 'middle' to someone of about this level. Moments like that made the standards feel a bit vague.

Evaluators reported that their perception of a performance was influenced by those that came before it, meaning that the same performance might have been rated differently depending on its position in the evaluation order. This raised challenges for consistency and introduced potential for comparison-based bias in the process. This pattern also appears to be linked to the fatigue that can accumulate when one teacher evaluates a large number of student performances in sequence—a common condition in school settings—which can in turn make evaluative standards less stable, even when the evaluator is not fully aware of its influence.

2) Second round: Structured and context-sensitive judgment

To address the discrepancies observed in the first evaluation round, the evaluators discussed and refined the subcategory criteria to establish a more coherent basis for subsequent assessments. Therefore, the second-round performance evaluation employed a more structured and detailed evaluation rubric than the first. A key development in the second-round evaluation was the formal introduction of an assessment category for performance attitude. Particularly, evaluators rated how seriously and attentively students completed their performances using a three-tiered scale. This score was evaluated independently from musical-technical categories and was intended to capture aspects such as concentration, commitment, and perseverance. This expanded structure not only improved the evaluators' ability to distinguish among students with similar technical skill levels, but also allowed for more meaningful feedback that recognized both musical and behavioral dimensions of performance.

A: In the second-round evaluation, we evaluated the affective domain in three tiers. For the 'high' level, we defined it as performing from beginning to end with a serious and committed attitude. For 'medium,' we divided it into two cases: the student was able to perform but showed somewhat less seriousness, or the student showed seriousness but did not complete the performance. For 'low,' we categorized performances where the student did not demonstrate a serious attitude at all.

B: In the first-round, it was very difficult to quantify this aspect, so it was challenging to describe it clearly. But in the second evaluation, this element could finally be incorporated, which was really beneficial. After all, we can't judge a performer solely based on accurate pitch and rhythm. Whether the student completed the performance—or completed it but with an obviously careless attitude—these things can all be heard. Being able to evaluate such elements made each part of the second evaluation much more meaningful.

The second evaluation revealed that while structured criteria can enhance objectivity, human assessment remains inherently complex and sensitive to performance context, interpretive ambiguity, and educational realities. Particularly, evaluators encountered difficulties with elements that required interpretive judgment, such as phrasing and musical expression because these aspects were often barely perceptible in non-major middle school students' performances. Although phrasing carried a substantial weight in the rubric, evaluators found it challenging to deduct points strictly, as doing so would have lowered scores disproportionately relative to the overall performance level.

A: It felt like pitch and rhythm were the main factors being evaluated. As for phrasing, I generally tried to give a decent rating whenever possible, because I didn't really feel that the students were consciously shaping their phrases. If we had evaluated phrasing strictly, their scores would likely have dropped significantly. So I treated phrasing as acceptable as long as the musical line connected reasonably well. It also seemed that phrasing was closely tied to breathing—how well their breath supported the line.

B: Just as A mentioned, I also felt that phrasing wasn't very evident. From what I could sense, students seemed to think about phrasing maybe only about 10 percent of the time. But even so, we couldn't simply deduct the full points allotted to that category; if we did, the score would fall straight into the medium range. And in reality, given the overall average level of performance, we couldn't justify scoring that harshly.

This suggests that while expressive musical elements are important in music performance evaluation, their weighting should be reduced for non-major students—an adjustment that should also be reflected in AI-based assessment.

Moreover, even for musical elements typically considered quantifiable—such as pitch accuracy—the evaluators emphasized that strict pitch-precision metrics were not always suitable for middle school non-major students.

A: Especially when students sing, their pitch keeps shifting. Particularly, in Korean traditional (folksong) singing, almost two-thirds of them sang in completely different pitch centers. Listening to accompaniment might help, but students start on different pitches, and I tried to evaluate them using relative pitch. But when they suddenly moved to an entirely different pitch in the middle, I really struggled with how to assess that. We can't treat everything after the shift as wrong, but computer would definitely mark all of it as incorrect.

These observations indicate that even seemingly objective musical criteria require relative application when assessing non-major middle school students. Accordingly, AI systems aimed at evaluating their singing or instrumental performances must be developed with sensitivity to the inherently fluid and context-dependent pitch behaviors of this population, rather than relying on rigid, pitch-precision metrics—underscoring an important design implication for AI-based assessment in school music contexts.

Beyond these element-level considerations, fatigue became more clearly influential during the second evaluation as the evaluators progressed through a large number of recordings. Evaluators described how sustained listening led to decreasing concentration and growing uncertainty in the judgments.

A: As my ears became fatigued, the large number of audio files made my evaluations feel increasingly ambiguous. But when we conducted the second evaluation and scored each element according to the criteria, it actually helped. Even when I was getting tired, using the rubric as a reference allowed me to evaluate more objectively compared to the first time.

At the same time, they noted that the introduction of structured and itemized criteria in the second round helped counteract this fatigue; using clearly defined reference points enabled them to maintain more consistent standards and rely less on moment-to-moment impressions than they had in the first round. While the rubric helped steady their judgments under fatigue, several evaluators noted that the cognitive demands of reviewing many recordings remained substantial—prompting the view that AI-assisted systems could provide additional support in maintaining consistency.

2. Developers' perceptions of human evaluation

1) Technical challenges in AI algorithm development

AI developers identified several technical and methodological challenges during the design of the performance evaluation algorithm. A major difficulty involved acquiring and processing performance data from non-music-major students, whose playing—unlike that of professional performers commonly used in existing music AI research—is often unpredictable and incomplete. Because of this, developers noted that existing algorithms can struggle to operate effectively or may generate inaccurate assessments in real school contexts. They further emphasized that current

models are highly sensitive to external variables, such as recording noise or time delays, highlighting the need for additional algorithmic development to accommodate real-world conditions found in school music class settings.

D: We found it difficult to apply existing AI-based evaluation models because most were developed for band ensembles or trained on recordings from highly skilled or professional performers—settings common in U.S. school band programs. In contrast, our task involved evaluating non-major middle school students. Many prior datasets were collected from music competitions or controlled recording environments, where performers are well-prepared and the audio quality is consistent. To our knowledge, there has been little research on how AI models should be adapted specifically for assessing non-major student performances, which highlights the need for frameworks tailored to this population.

E: However, the settings in which we evaluate music, the target learners, and the criteria applied in school contexts differ significantly from those conditions (professional, well-controlled performance settings). Consequently, we need to carefully reconsider how evaluation standards should be defined and how AI systems should be trained or configured to reflect these different environments.

In addition, as previously noted, musical elements such as phrasing and expression are inherently complex and ambiguous even for human evaluators, making them difficult to quantify. Developers acknowledged that current AI performance evaluation technologies face substantial limitations in learning or modeling these features through rule-based logic or pattern recognition.

E: I think the aspects that are subjective for humans are also difficult for AI. Pitch and rhythm are objective enough that even humans can give fairly quantifiable scores, but phrasing, expressiveness, and tone color are subjective even for humans, so it feels like there's no clear standard for training AI model on them.

Concerns were also raised regarding the rigidity of existing AI models, particularly rule-based systems that penalize minor deviations without considering natural performance variability. While human evaluators may overlook minor mistakes when the overall musical flow is maintained, AI often applies penalties mechanically, resulting in disproportionately low scores. This issue is especially problematic when assessing middle-school students or non-majors, whose performances naturally show greater variability. In school setting—where maintaining motivation and supporting musical growth are central goals—such rigid scoring practices may therefore be

misaligned with pedagogical needs.

These issues reveal the complexity of modeling student performance in ways that reflect real educational contexts. This cautious stance reflects a broader recognition that performance assessment involves not only measurable acoustic features but also interpretive and situational elements that are difficult to encode computationally. Therefore, AI-based performance evaluation systems must be applied with as awareness of their limitations, particularly in diverse educational contexts.

2) Developing human-informed AI model for classroom evaluation

While analyzing performance evaluation records, AI platform developers observed that human assessment involves a unique interplay of subjectivity, intuition, and contextual sensitivity. They noted that a single performance could receive a high overall impression score while simultaneously receiving lower ratings in specific rubric-based criteria. Although such mismatches might initially appear inconsistent, developers viewed them as characteristic features of how teachers make evaluative judgments, rather than as errors to be corrected.

E: I would not necessarily frame this as a problem; rather, it highlights certain characteristics of how human teachers evaluate performances. These characteristics should be thoughtfully integrated with more objective and categorized assessment features to build a more effective AI training model. ... After all, the overall impressions that teachers intuitively form are essential elements that must be reflected within the evaluation domain. ... Although it is difficult to articulate this precisely, I believe such intuitive judgments constitute highly meaningful data. Since we are ultimately developing an AI model intended to interact with human users, a system that aligns closely with human intuitive impressions may, in fact, be perceived as more relatable and effective from the user's perspective.

Developers emphasized that human evaluative tendencies—including affective judgments and contextual influences—should therefore be treated as intrinsic structural components of the assessment process. These elements were viewed as valuable resources for training models that aim to reflect the complexity of teacher judgment.

Building on this perspective, developers stressed that the central objective of AI-based evaluation is not simply to achieve quantitative accuracy but to approximate the decision-making patterns of human raters. To this end, they explored strategies for integrating both holistic, impression-based evaluations and detailed, analytic criteria. When systematic score gaps emerged

between human and AI assessments, they considered adjusting feature weights to better align the model with human evaluative tendencies.

E: If we can identify major categories that account for score differences between overall and categorical evaluations, we could adjust the AI model's learning process by giving certain categories more or less weight. By training models in multiple ways—each reflecting different characteristics of the data—we can then examine which model aligns most closely with human evaluations. That will likely be the direction we pursue moving forward.

Developers also highlighted the advantages of moving away from rule-based systems toward audio-driven, deep learning-based approaches that can learn evaluative patterns directly from data rather than relying on explicitly defined rules.

D: The goal is to develop models that follow human scoring tendencies as closely as possible. Unlike rule-based systems, more audio-driven models operate by analyzing the performance holistically and comparing it to reference materials, such as musical notation, to generate data-based assessments. Of course, we still need to experiment to determine which features, when explicitly extracted and fed into the model, produce outputs that most closely resemble human assessment. But ..., if we move toward more recent deep learning-based approaches rather than rule-based systems, the model essentially learns to mimic human scoring patterns without us needing to specify exactly how it makes each decision.

These insights informed ongoing experimentation with multiple model configurations, combining impression-level data with detailed analytic features to determine which training approaches most effectively replicate teacher judgment.

In addition to technical considerations, developers stressed the importance of aligning the system with practical classroom needs. Teachers differ in how they envision using AI: some view it as a way to supplement human evaluation by offering objective insights into overlooked aspects, while others see it as a preliminary evaluator whose outputs they refine.

E: The direction we take can vary significantly depending on how we define the research goal. Some teachers may want AI to complement human evaluation by providing objective insights into aspects they might overlook. Others may prefer to use AI as a first-pass evaluator that automatically generates assessments which they then verify or refine. The approach largely depends on what teachers consider ideal.

D: Ultimately, it comes down to the kinds of feedback teachers want to provide to students in the school setting. AI is simply a tool—whether it functions by mimicking human scoring patterns or by using a different set of rules—and it can be applied in multiple ways depending on educational needs.

These perspectives underscore that decisions about whether AI should complement or partially substitute for human judgment must be guided through continued dialogue with educators and grounded understanding of classroom contexts. Developers emphasized that for AI to be adopted as a trusted support tool, the system must prioritize interpretability, usability, and pedagogical relevance from the early stages of design.

V. Discussion & Conclusion

This study examined the characteristics of human evaluation in secondary music education and their intersection with AI-based assessment systems. Unlike professional performance contexts (Álvarez-Díaz et al., 2020; Berge, 2003; Wesolowski, et al., 2017), school-based music education focuses on guiding students' musical development and encouraging them to lead an active musical life. As the subject ultimately aims to foster holistic growth, AI evaluation research in this specific context must reflect these educational values. Given the scarcity of theoretical frameworks and prior research on AI evaluation within this specific educational context, the findings of this study provide essential considerations for developing classroom-oriented AI models that reflect these unique pedagogical values.

Through interview with pre-service music teachers, who participated as performance evaluators, and AI developers, the study examined the school-context realities of performance assessment and the tendencies inherent in human evaluative judgment, deriving key implications for the design of AI-based assessment models. Longstanding challenges in music education—such as subjectivity, inconsistency, and limited technological support—also emerged, consistent with existing research (Wrigley & Emmerson, 2011; Moura et al., 2024). While previous studies have highlighted the need for more balanced and learning-supportive assessment in school music education (Russell & Austin, 2010; Wesolowski et al., 2017), this study further explored how AI systems could be designed to align with teachers' or human evaluative tendencies, and educational goals in secondary school settings.

The findings point to two interconnected implications for AI design in school music assessment. First, human evaluators placed particular importance on intuitive judgments shaped by students' performance attitude—such as sincerity, effort, and confidence—an emphasis that aligns with prior work highlighting the educational and motivational aims of school-based assessment (Russell & Austin, 2010; Wesolowski et al., 2017). Such tendencies suggest that these affective dimensions may need to be represented more explicitly in rubrics and incorporated into AI training data if AI systems are to approximate the evaluative priorities present in classroom settings. Second, the study shows that musical-technical criteria valued in professional performance contexts do not map neatly onto the realities of secondary school learners. Expressive features like phrasing were often too subtle to evaluate reliably, while even quantifiable elements such as pitch tended to require relative rather than rule-based interpretation. Teachers frequently assessed pitch leniently—focusing on the melodic contour or overall pitch trajectory rather than exact pitch matching—indicating that overly strict, precision-based algorithms may misrepresent student achievement. Together, these insights suggest that AI systems for school contexts must recalibrate the weighting and interpretation of both expressive and technical criteria, aligning evaluation with the holistic, educationally oriented judgment patterns observed among human teachers.

Building on these discussion points, this study highlights the need for AI-supported assessment in secondary school music education, particularly in contexts where a single teacher is responsible for evaluating large numbers of students. Prior research has shown that such instructional conditions can make it difficult for teachers to provide timely and detailed feedback (Russell & Austin, 2010; Wesolowski et al., 2017), suggesting that AI could serve as a complementary tool that reduces this burden. For AI to serve as a pedagogically meaningful, however, its design must be grounded in the realities of school-based assessment rather than driven solely by technical considerations. Research on music performance evaluation emphasizes that clearly defined assessment constructs are essential for reliability, yet also that evaluative criteria must remain responsive to diverse classroom contexts, instructional goals, and student performance levels (Wesolowski et al., 2017). In line with these insights, the present study suggests that AI systems should be designed to learn from localized teacher judgment patterns rather than enforcing uniform, one-size-fits-all criteria.

Moving forward, AI-based assessment could be extended beyond audio-only analysis to incorporate performance video, enabling timestamped feedback informed by gestures, posture, facial expression, and other visually available indicators of musical engagement (Moltisanti, Fidler, & Damen, 2019). Similarly, natural-language feedback generation may support more

instructionally meaningful guidance for students (Li, et al., 2025). However, such advanced functionalities must be adapted to the pedagogical realities of secondary school settings, ensuring that the quantity and type of feedback remain appropriate for non-major learners and aligned with teachers' instructional goals.

In conclusion, developing these adaptive, human-informed systems will require sustained collaboration among developers, teachers, and music education researchers, as well as diverse training data that reflect non-major learners and performances that fall outside formal norms. When guided by pedagogical insight and grounded in the complexity of actual classroom practice, AI can serve as a meaningful partner—enhancing equity, reducing teacher workload, and supporting individualized learning. However, a limitation of this study lies in the potential discrepancy between the assessment practices of pre-service teachers and those of experienced educators, despite their shared specialized musical backgrounds. To build upon these findings, future research should explore the perspectives of in-service teachers based on their actual classroom assessment experiences.

References

- Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, 21(3), 246-255.
- Álvarez-Díaz, M., Muñoz-Bascón, L. M., Soria-Aleman, A., Veintimilla-Bonet, A., & Fernández-Alonso, R. (2020). 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>
- Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education*, 51(2), 137-150. <https://doi.org/10.2307/3345847>
- Davies, P. (2000). Computerized peer assessment. *Innovations in Education and Training International*, 37(4), 346-354.
- Evin, M. (2024). A review on AI-enabled techniques for evaluating musician's performance. *AIP Conference Proceedings*, 3149(1), 140018. <https://doi.org/10.1063/5.0224734>
- Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: Reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education*, 6(2), 229-245.
- Giraldo, S., Waddell, G., Nou, I., Ortega, A., Mayor, O., Perez, A., Williamon, A., &

- Ramirez, R. (2019). Automatic assessment of tone quality in violin music performance. *Frontiers in Psychology*, 10, 1-12. <https://doi.org/10.3389/fpsyg.2019.00334>
- Gurley, R. (2012). *Student perception of the effectiveness of SmartMusic as a practice and assessment tool on middle school and high school band students*. Master's dissertation, Texas Tech University.
- Hewitt, M. P. (2002). Self-evaluation tendencies of junior high instrumentalists. *Journal of Research in Music Education*, 50(3), 215-226. <https://doi.org/10.2307/3345799>
- Hewitt, M. P. (2005). Self-evaluation accuracy among high school and middle school instrumentalists. *Journal of Research in Music Education*, 53(2), 148-161. <https://doi.org/10.2307/3345515>
- Jiang, Y. (2023). Expert and novice evaluations of piano performances: Criteria for computer-aided feedback. *Proceedings of the 24th ISMIR Conference* (pp. 367-374). International Society for Music Information Retrieval.
- Kim, H. J., Lee, J. Y., & Jang, S. Y. (2019). Pre-service teachers' perception on peer feedback in English writing. *International Journal of Contents*, 19(1), 513-523. <https://doi.org/10.15738/kjell.20..202008.335>
- Li, Y., & Sun, R. (2023). Innovations of music and aesthetic education courses using intelligent technologies. *Education and Information Technologies*, 28, 13665-13688. <https://doi.org/10.1007/s10639-023-11624-9>
- 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>
- Min, K. H., Kim, S. Y., Kim, Y. H., Bang, K. J., Seung, Y. H., Yang, J. M., Lee, Y. K., Lim, M. K., Cho, S., Joo, D. C., & Hyun, K. S. (2017). *Introduction to music education* (3rd ed.). Hakjisa.
- Ministry of Education (2022). *National music curriculum*. Ministry of Education Notice No. 2022-33. [Supplement No. 12]. Ministry of Education.
- Moltisanti, D., Fidler, S., & Damen, D. (2019). *Action recognition from single timestamp supervision in untrimmed videos*. 10.48550/arXiv.1904.04689. <https://doi.org/10.48550/arXiv.1904.04689>
- Moura, N., Dias, P., Verissimo, L., Oliveira-Silva, P., & Serra, S. (2024). Solo music performance assessment criteria: A systematic review. *Frontiers in Psychology*, 15, 1-32. <https://doi.org/10.3389/fpsyg.2024.1467434>
- Pati, K. A., Gururani, S., & Lerch, A. (2018). Assessment of student music performances using deep neural networks. *Applied Sciences*, 8, 1-18. <https://doi.org/10.3390/app8040>

- Pellegrino, K., Conway, C. M., & Russell, J. A. (2015). Assessment in performance-based secondary music classes. *Music Educators Journal*, 102(1), 48-55. <https://doi.org/10.1177/002743211555901>
- Richmond, J. W. (2002). Law research and music education. In R. Colwell & C. Richardson (Eds.), *The new handbook of research on music teaching and learning: A project of the Music Educators National Conference* (pp. 33-47). Oxford University Press.
- Russell, B. E. (2010). The development of a guitar performance rating scale using a facet-factorial approach. *Bulletin of the Council for Research in Music Education*, 184, 21-34. <https://doi.org/10.2307/27861480>
- Russell, J. A., & Austin, J. R. (2010). Assessment practices of secondary music teachers. *Journal of Research in Music Education*, 58(1), 37-54. <https://doi.org/10.1177/002242940936006>
- Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education*, 18, 46-56. <https://doi.org/10.1177/1321103X0201800106>
- Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception*, 21(1), 21-41. <https://doi.org/10.1525/mp.2003.21.1.21>
- Tucker, C. F. (2016). *A case study of the integration of SmartMusic into three middle school band classrooms found in update South Carolina*. Doctoral dissertation, Gardner-Webb University.
- Wei, J., Karuppiyah, M., & Prathik, A. (2022). College music education and teaching based on AI techniques. *Computers and Electrical Engineering*, 100, 1-12. <https://doi.org/10.1016/j.compeleceng.2022.10785>
- Wesolowski, B. C., Amend, R. M., Barnstead, T. S., Edwards, A. S., Everhart, M., Goins, Q. R., Grogan III, R. J., Herceg, A. M., Jenkins, S. I., Johns, P. M., McCarver, C. J., Schaps, R. E., Sorrell, G. W., & Williams, J. D. (2017). The development of a secondary-level solo wind instrument performance rubric using the multifaceted rasch partial credit measurement model. *Journal of Research in Music Education*, 65(1), 95-119. <https://doi.org/10.1177/0022429417694873>
- Wrigley, W. J., & Emmerson, S. B. (2011). Ecological development and validation of a music performance rating scale for five instrument families. *Psychology of Music*, 41(1), 97-118. <https://doi.org/10.1177/0305735611418552>

Zhang, S. (1995). Re-examining the affective advantages of peer feedback in the ESL writing class. *Journal of Second Language Writing*, 4(3), 209-222. [https://doi.org/10.1016/1060-3743\(95\)90010-1](https://doi.org/10.1016/1060-3743(95)90010-1)