Jurnal Pendidikan Malaysia 51(2) (2026): 58-69
DOI: http://dx.doi.org/10.17576/JPEN-2026-51.02-05

Evaluating the Effectiveness of ChatGPT as an AI-Assisted Tool in Music Performance Pedagogy

(Menilai Keberkesanan ChatGPT sebagai Alat Berbantukan Kecerdasan Buatan dalam Pedagogi Persembahan Muzik)

YUCHENG MA, AMELIA ALIAS*, NURFARADILLA MOHAMAD NASRI

ABSTRACT

Artificial intelligence (AI) has produced measurable benefits in general education, yet quantitative evidence remains limited in the embodied and interpretive domain of music performance pedagogy. This one-group pretest-posttest intervention study with correlational analyses examined changes in students' blind-scored music performance during a ChatGPT-assisted period, teachers' perceptions of preparation time, lesson accuracy and assessment reliability, and associations between technology-use variables and teaching-learning outcomes. A four-phase 12-week study, including a four-week ChatGPT-assisted phase (Weeks 7-10), involved 247 students and 52 teachers from five Chinese higher-education institutions selected through stratified purposive sampling. Data were collected using adapted student and teacher outcome questionnaires, three study-specific technology-use indicators, and a five-dimension performance rubric, blind-scored by three judges at pretest and posttest, and were analysed using t-tests, ANOVA, correlation, regression and MANOVA. Students' performance increased significantly (dz = 0.74), with a significantly greater gain in musical expression than in the embodied technical dimensions. Teachers rated all instructional-efficiency outcomes significantly above the neutral midpoint. Among students, usage frequency predicted self-reported outcomes; among teachers, only perceived usefulness was significant. The linear association between usage frequency and blind-scored performance gain was nonsignificant. The findings support ChatGPT's effectiveness as a supervised aid for interpretive learning, practice planning, lesson preparation and rubric-aligned work, but not as a substitute for embodied technical feedback from human teachers.

Key Words: ChatGPT; artificial intelligence in education; music performance pedagogy; technology acceptance; higher education

ABSTRAK

Kecerdasan buatan telah menghasilkan manfaat yang boleh diukur dalam pendidikan umum, namun bukti kuantitatif masih terhad dalam pedagogi persembahan muzik yang bersifat jasmani dan interpretif. Kajian intervensi praujian-pascaujian satu kumpulan dengan analisis korelasi ini meneliti perubahan persembahan muzik pelajar yang dinilai secara buta, persepsi guru tentang masa penyediaan, ketepatan pengajaran dan kebolehpercayaan pentaksiran, serta hubungan antara penggunaan teknologi dengan hasil pengajaran-pembelajaran. Kajian empat fasa selama 12 minggu, termasuk fasa berbantukan ChatGPT selama empat minggu (Minggu 7-10), melibatkan 247 pelajar dan 52 guru daripada lima institusi pengajian tinggi di China melalui pensampelan bertujuan berstrata. Data dikumpulkan menggunakan soal selidik hasil pelajar dan guru yang diadaptasi, tiga indikator penggunaan teknologi khusus kajian, serta rubrik prestasi lima dimensi yang dinilai secara buta oleh tiga juri pada praujian dan pascaujian, dan dianalisis menggunakan ujian-t, ANOVA, korelasi, regresi dan MANOVA. Prestasi pelajar meningkat secara signifikan (dz = 0.74), dengan peningkatan ekspresi muzikal yang secara signifikan lebih besar daripada dimensi teknikal jasmani. Guru menilai semua hasil kecekapan pengajaran secara signifikan melebihi titik tengah neutral. Bagi pelajar, kekerapan penggunaan meramalkan hasil laporan kendiri; bagi guru, hanya persepsi kebergunaan signifikan. Hubungan linear antara kekerapan penggunaan dan peningkatan prestasi yang dinilai secara buta tidak signifikan. Dapatan menyokong keberkesanan ChatGPT sebagai alat bantu yang diselia untuk pembelajaran interpretif, perancangan latihan, penyediaan pelajaran dan kerja berasaskan rubrik, tetapi bukan sebagai pengganti maklum balas teknikal jasmani daripada guru manusia.

Kata Kunci: ChatGPT; kecerdasan buatan dalam pendidikan; pedagogi persembahan muzik; penerimaan teknologi; pendidikan tinggi

INTRODUCTION

Artificial intelligence (AI) has become a general-purpose technology reshaping industries worldwide (Rashid & Kausik 2024). In education, systematic reviews report that AI-supported personalised learning, intelligent tutoring and automated assessment yield measurable gains in achievement and engagement (Merino-Campos 2025; Wang et al. 2024), with substantial time savings for AI-supported instruction (Belloula 2025) and lesson preparation (Lee et al. 2024). ChatGPT, the most widely adopted conversational AI system, delivers individualised, on-demand textual guidance at negligible marginal cost (Nah et al. 2023).

Music performance education, however, is fundamentally different from the contexts in which evidence was produced. Performance learning is embodied and interpretive, contingent upon fine-grained motor skills, expressive decision-making and aural judgments that are acquired through guided practice. Its dominant mode of instruction is individual tuition in a studio, which is limited by the hours of direct contact and the lag between lessons and feedback, as well as variable levels of assessor consistency (Gaunt 2008; Wesolowski et al. 2016). While the use of ChatGPT for practice planning and lesson preparation could be seen as an affordable and accessible option for text-based guidance, textual guidance is insufficient for embodied technical feedback. Furthermore, most technologies for supporting musicians are focused on production-oriented processes and do not produce analysable learning data (Deruty et al. 2022), and the discussion around AI in this field is largely conceptual rather than empirical (Bell et al. 2025; O'Leary 2025). Higher-education establishments do not have field-specific evidence from their practice about whether using ChatGPT does coincide with demonstrable change in performance and whether teachers find it useful for their practice.

Three research gaps in the literature provide a rationale for this investigation. First, the quantitative ChatGPT studies were conducted in broad educational settings, including STEM, and little is known about pretest-posttest changes in performance-based outcomes like blind-scored playing quality. Second, there was a lack of quantitative evidence on teachers' perceptions of preparation time, lesson accuracy, and assessment reliability in music performance instruction. Third, it remains unclear whether usage frequency is associated with self-reported outcomes beyond perceived usefulness and ease of use. These research gaps are important to address because conservatoire departments considering restricting or allowing generative AI in studio teaching settings rely heavily on findings from other educational contexts. Although ethical and cognitive risks (Wiese et al., 2025) remain important contextual concerns, they were beyond the scope of this study's measurements and will be considered in future investigations.

To address these gaps, the study pursued two primary objectives and one secondary correlational objective:

1. To examine pretest-posttest changes in students' blind-scored music performance during the ChatGPT-assisted period.
2. To examine teachers' perceptions of ChatGPT-supported instructional efficiency in terms of preparation time, lesson accuracy and assessment reliability.
3. To examine whether ChatGPT usage frequency explains self-reported teaching-learning outcomes beyond perceived usefulness and perceived ease of use.

Accordingly, two primary questions and one secondary question were formulated:

1. What changes occurred in students' blind-scored music performance from pretest to posttest during the ChatGPT-assisted period?
2. How do teachers perceive ChatGPT-supported instructional efficiency in terms of preparation time, lesson accuracy and assessment reliability?
3. How does ChatGPT usage frequency relate to self-reported teaching-learning outcomes after accounting for perceived usefulness and perceived ease of use?

H1a and H2 were primary, H3 was secondary, and one contrast was exploratory.

H1a: Students' posttest blind-scored performance scores will be significantly higher than their pretest scores. Exploratory secondary contrast: The pretest-posttest gain in musical expression was compared with the mean gain across pitch and rhythm accuracy, fluency, and tone quality and control.

H2: Teachers' ratings of ChatGPT's contribution to preparation time, lesson accuracy and assessment reliability will be significantly above the neutral midpoint.

H3: After accounting for perceived usefulness and perceived ease of use, ChatGPT usage frequency will be positively associated with self-reported teaching-learning outcomes.

LITERATURE REVIEW

The review rests on searches of Scopus, Web of Science, ERIC and Google Scholar covering 2015-2026, combining ChatGPT and generative artificial intelligence with music education, performance pedagogy, effectiveness and quantitative-outcome terms, restricted to English-language publications; Chinese-language databases were not systematically searched, a boundary registered in the limitations.

AI IN EDUCATION AND CHATGPT

Educational AI stems from recent advances in natural language processing that allow large language models to demonstrate proficiency in contextual understanding and generate human-like text (Supriyono et al. 2024). ChatGPT then brought these capabilities to a popular audience (Nah et al. 2023), while systematic reviews confirm improvements in both performance and engagement in higher education (Merino-Campos 2025; Vieriu & Petrea 2025). Additionally, educator surveys report decreased preparation time spent (Belloula 2025; Lee et al. 2024). Most compelling are the aggregate results from Deng et al. (2025), a meta-analysis that included 69 experimental studies published in 2022-2024. These articles reported substantial benefits when using ChatGPT, including 0.712 standard deviation (Hedges' g) increases in academic performance, 0.881 in affective-motivational states, and 0.703 in higher-order thinking. Its evidential basis is narrower than these figures suggest: the interventions were primarily language-focused and included few conservatoire-specific contexts, while the effect size was found in grades. Additionally, the authors note that the lack of power analyses in individual studies and questions about posttest validity "warrant cautious interpretation". Overall, a 0.71 g improvement is a substantial increase in test scores in standard classroom settings, but not blind-scored performance assessments such as those used in conservatoires. The extrapolation to music education contexts is precisely what this study attempts to evaluate. In turn, Kamalov et al.'s (2023) review is better seen as an agenda-setting review than a systematic one.

The same literature catalogues genuine risks. Habitual delegation of thinking to AI is associated with cognitive offloading and weaker critical-thinking performance (Gerlich 2025), over-reliance on AI dialogue systems can erode students' independent cognitive abilities (Zhai et al. 2024), dependence has behavioural costs such as procrastination (Ali et al. 2024), and the tendency of language models to produce confident but incorrect content makes human oversight indispensable (Nah et al. 2023). From this strand follows a competing expectation that intensive use carries cognitive costs offsetting its benefits; this study does not test that expectation and registers it as future work.

THE MASTER-APPRENTICE MODEL AND ITS LIMITS

Higher music performance education preserves its historical pedagogy of one-to-one tuition in a studio setting, in which techniques and interpretive approaches are passed from teacher to student. Research into conservatoire pedagogy highlights both the strengths of this model (provision of high levels of personal engagement and emphasis on the teacher-student relationship) and its weaknesses (variation in quality of individual tuition and potential undermining of students' independence) (Carey & Grant 2015; Gaunt 2008, 2010). Higher education more generally reveals trends towards larger instructional loads, with negative implications for learning, and less scope for individualised instruction (Shi 2019; Korthals Altes et al. 2024; neither specifically considering music, and both serving strictly as analogy). In the time between lessons, students have to rely on their own initiative to engage in largely unsupervised practice, in which even basic approaches like slow practice and careful attention to tempo are used inconsistently (Allingham & Wöllner 2022). Summative assessment presents additional challenges, with marking frequently shown to be inconsistent between assessors, even when using objective criteria such as mark schemes (Wesolowski et al. 2016). The context within which a low-cost, always available AI assistant might intervene is one of limited and inconsistent contact with teachers, variable practice habits, delayed and potentially inconsistent feedback, and difficulties in summative assessment.

AI IN MUSIC PERFORMANCE PEDAGOGY

Although research into the implications of ChatGPT for music is nascent, there is a considerable body of work in two areas pertinent to this study. First, in the area of music information retrieval there is a long-standing history of computer-assisted evaluation of pitches, rhythms, timbres, and dynamics (Lerch et al. 2020). Artificial intelligence has been applied to the evaluation of student performance, sometimes using deep neural networks (Pati et al. 2018). Second, there is a related literature on computer-assisted practice, including early work on intelligent accompaniment (Sheldon et al. 1999) and more recent programs such as SmartMusic and YouSician (Waddell & Williamon 2019). These programs provide real-time auditory feedback on pitch and rhythmic accuracy, but not interpretive and contextual information such as historical background, composer intention, and expressive strategy, which a text-based language model could address. In other areas of education, data-driven artificial intelligence has been developed to suggest sequences of curricular topics (Chu & Ashraf 2025; Vieriu & Petrea 2025). By analogy, such an approach could be applied to recommend repertoires, although this remains speculation. Generative artificial intelligence has been used to mark complex student writing with a degree of accuracy approaching that of human markers (Awidi 2024), and preliminary studies suggest application to automated marking of rubric-based assessments (Legi et al. 2024). In other fields of education, educators have used ChatGPT to develop motivational

learning resources (Rutherford et al. 2025) and accepted its role in assessment (Khlaif et al. 2024). This complementarity suggested that change during the ChatGPT-assisted period might differ for expressive aspects of performance from the more technically oriented aspects, an idea that was explored in this study as an exploratory secondary contrast.

THEORETICAL FRAMEWORK

Three theories were integrated in the study. The Technology Acceptance Model (TAM) identifies perceived usefulness and perceived ease of use as evaluations underlying technology acceptance and use (Davis 1989). In the current study, these two constructs were treated as acceptance-related covariates rather than direct determinants of teaching-learning outcomes. According to deliberate practice theory, development results from purposeful activity involving feedback; thus, more frequent ChatGPT use could provide more opportunities for practice planning, repertoire analysis, and feedback support without replacing embodied correction (Ericsson et al. 1993). In turn, social cognitive theory highlights the role of goal-oriented mastery experiences in sustaining motivation, although self-efficacy was not measured in the study (Bandura 1986). Thus, H3 states that the association between usage frequency and self-reported outcomes will be positive and incremental after controlling for the two acceptance dimensions, and it does not posit causality or mediation.

METHODOLOGY

RESEARCH DESIGN

The study adopted a one-group pretest-posttest intervention design with correlational analyses. The study design consisted of four stages that lasted 12 weeks. The first two weeks were for pilotage and participant briefing, while weeks 3-6 were for the baseline, which ended with pre-tests. The intervention involving ChatGPT-assisted practice occurred during weeks 7-10, which concluded with post-tests. Finally, weeks 11 and 12 were for analysis. The same participants underwent performance assessments before and after the intervention, with scores determined by blind scorers. Additionally, the study design involved correlational analyses to establish relationships between acceptance and use variables and self-reported outcome measures. Randomization and a separate control group were not feasible within the participating conservatoire programmes. Therefore, the study analysed within-subject changes from pretest to posttest and correlational data rather than a controlled causal effect. Accordingly, maturation, concurrent regular training, and other time-related factors were considered alternative explanations for the observed changes. The study adhered to ethical principles, including informed consent, voluntary participation and withdrawal, confidentiality using anonymised identifiers, and institutional approval.

Throughout the training weeks, educators were instructed to use ChatGPT for lesson planning and assessment, while students learned to use it for practice planning and repertoire suggestions. In the four-week assisted trial, participants employed it for individualised practice plans, suggestions, and repertoire analysis, whereas educators used it for lesson planning, scheduling, and assessment-rubric development. Usage frequency was measured through a survey item asking participants to self-report their use of ChatGPT. Specifically, participants used the standard web version of ChatGPT (GPT-4o; free unless individually subscribed). Additionally, exposure was not standardised: there were no prescribed prompts, minimum usage frequency, platform-level logs, or fidelity checks. Moreover, the use of ChatGPT was not restricted during baseline.

RESEARCH SAMPLE/PARTICIPANTS

The target population was approximately 1,200 music performance students and teachers at five higher education institutions in China. The required sample size was determined using Cochran's (1977) approach and Krejcie and Morgan's (1970) table, yielding 293 and 291, respectively. These numbers served as guidelines because the sampling procedure was not fully representative but rather purposeful, given the combined nature of the population and the different measures for students and teachers. Thus, while this was the number for the combined population, the inference was made on n = 52 teachers, noting the caveat of its limited representativeness. The recruitment target was set at 300, with 351 questionnaires distributed (289 to students and 62 to teachers) and 318 returned (90.6% response rate). 19 questionnaires were removed due to substantial missingness or uniform response patterns, yielding 299 (247 students and 52 teachers). The stratified purposive sampling was conducted at the organisational level, with inclusion criteria including the inclusion of music performance programmes and basic technology proficiency (students) or at least two years of teaching experience (teachers). The questionnaires were distributed through Qualtrics and Google Forms to the departmental mailing lists, taking no longer than 20 minutes to complete. The distribution is described by Table 1 below. The students were male (63.6%), which departs from the typical ratio for music performance programmes in China, characteristic of the participating universities for

the study. Approximately 65% of students and 63% of teachers reported having prior experience with ChatGPT, with a mean weekly frequency of 2.31 and 2.60 (1-4 scale).

TABLE 1. Demographic profile of respondents

Group	Variable	Category	n	%
Students (n = 247)	Gender	Male	157	63.6
		Female	90	36.4
	Year of study	Year 1	43	17.4
		Year 2	61	24.7
		Year 3	69	27.9
		Year 4	49	19.8
		Postgraduate	25	10.1
	Major instrument	Violin	58	23.5
		Piano	55	22.3
		Voice	52	21.1
		Other	44	17.8
		Wind	38	15.4
	Years of learning	Less than 3 years	68	27.5
		3-5 years	68	27.5
		6-10 years	44	17.8
		More than 10 years	67	27.1
	Prior ChatGPT use	Yes	161	65.2
		No	86	34.8
Teachers (n = 52)	Gender	Male	20	38.5
		Female	32	61.5
	Teaching experience	Less than 3 years	7	13.5
		3-5 years	10	19.2
		6-10 years	12	23.1
		11-20 years	14	26.9
		More than 20 years	9	17.3
	Instrument taught	Wind	14	26.9
		Piano	13	25.0
		Voice	7	13.5
		Strings	7	13.5
		Other	4	7.7
		Not reported	7	13.5
	Student level taught	Beginner	18	34.6
		Multiple levels	13	25.0
		Advanced	12	23.1
		Intermediate	9	17.3
	Prior ChatGPT use	Yes	33	63.5
		No	19	36.5

DATA COLLECTION METHOD/INSTRUMENTATION

Three instruments were employed. First, ChatGPT acceptance and use were operationalized by two study-specific single-item global measures labelled perceived usefulness and perceived ease of use, corresponding to the TAM constructs described by Davis (1989), along with one study-specific usage-frequency item. Focused single-item measures were

selected because the study treated each construct as a global judgement within its narrow domain of focus, which is in line with the general single-item measurement argument of Bergkvist and Rossiter (2007). Second, the student outcome questionnaire consisted of 20 items corresponding to practice efficiency, technical proficiency, repertoire mastery, and overall evaluation (five items per construct). The items were adapted and contextualized according to the objectives of the study from existing validated measures of self-regulated music practice, musical learning and performance self-efficacy, and ChatGPT-related efficiency and satisfaction (Ritchie & Williamon 2011, 2013; Miksza 2012; Madeira et al. 2018; Yu et al. 2024). Specifically, the five practice-efficiency items tapped into planning, monitoring, and time management, while the technical-proficiency items assessed pitch and rhythm accuracy, fluency, technical control, and musical expression. Similarly, the five repertoire-mastery items covered learning new musical material, achieving goals, and understanding styles, and the student overall-evaluation items addressed learning effectiveness and satisfaction. The 17-item teacher outcome questionnaire measured preparation time, lesson accuracy, assessment reliability, and overall evaluation and was adapted and contextualized from existing measures of educator-AI performance expectancy (Acikgul & Sad 2026) and GAI teaching use (Perez-Montesdeoca et al. 2026). Specifically, the four preparation-time items tapped into lesson-plan preparation, materials preparation, and time saving, while the lesson-accuracy items assessed systematic content organization and alignment with teaching objectives. The four assessment-reliability items addressed assessment design, objectivity and consistency, and the five overall-evaluation items corresponded to teaching effectiveness, personalization, and satisfaction. Across both questionnaires, generic wording related to music practice or AI use was substituted with ChatGPT-assisted music-performance scenarios, the referent was changed to perceived improvement following use, and the five-point Likert scales were standardised; no source item was reproduced verbatim. Third, a five-dimension performance rubric assessed pitch and rhythm accuracy, fluency, tone quality and control, musical expression, and overall performance from 1 to 5, with scores provided by three judges. To reduce expectancy bias, the recordings were anonymized with coded identifiers, pooled across the two occasions, and presented in a computer-generated random order. The judges scored independently, did not confer, and were blind to student identity and measurement occasion.

The development process involved three stages of content-validity evaluation. First, item coverage was compared against the cited source measures and construct domains for alignment. Second, five expert judges—qualified music educators, performance lecturers, and AI-in-education specialists—reviewed the draft items for wording clarity, construct relevance, pedagogical applicability, and contextual appropriateness for ChatGPT-assisted music-performance practice. The iterative discussion focused on item-level suggestions to refine the phrasing and operationalization before large-scale distribution, hence the absence of individual item scores that would allow computation of the CVI or CVR. Third, the student and teacher questionnaires were tested with a smaller sample of 30 students and 10 teachers to evaluate item clarity, response-format usability, respondent burden, and contextual fit. Content-validity evidence is therefore reported qualitatively on the basis of expert and pilot-tester feedback, consistent with the literature (Boateng et al., 2018). Main-study reliability was satisfactory, with student subscale Cronbach's alphas ranging from 0.796 to 0.829 (0.867 for the total scale) and teacher subscale alphas ranging from 0.716 to 0.746 (0.884 for the total scale). Construct validity of the student and teacher outcome measures was evaluated separately using exploratory factor analysis with principal axis factoring and Promax rotation (power = 4). The number of factors was determined through parallel analysis with 1,000 random correlation matrices, scree-plot inspection, and interpretability against the prespecified four-factor structures. For the student questionnaire (n = 247), the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO = 0.857) and Bartlett's test of sphericity ($\chi^2(190) = 1720.22$, $p < .001$) confirmed the data's suitability for factor analysis. Parallel analysis retained four factors; the first four eigenvalues of the unreduced correlation matrix explained 59.52% of total item variance. With the specified 0.50 minimum primary loading, pattern coefficients on the intended factors ranged from 0.63 to 0.82, with no secondary pattern coefficient exceeding 0.30. All 20 items were retained. For the teacher questionnaire (n = 52), KMO = 0.726 and Bartlett's test of sphericity ($\chi^2(136) = 405.92$, $p < .001$) indicated the adequacy of the data. Parallel analysis also retained four factors; the first four eigenvalues of the unreduced correlation matrix explained 67.91% of total item variance. Pattern coefficients on the intended factors ranged from 0.58 to 0.89, the highest secondary coefficient was 0.23, and all 17 items were retained. With 52 participants' responses for 17 items, construct validity of the teacher questionnaire is tentatively supported and requires replication in a larger sample. Finally, inter-judge reliability across the 494 scored performances (247 participants × two occasions, three judges) used two-way random-effects, absolute-agreement intraclass correlations. Average-measure ICC(2,k) values ranged from 0.788 to 0.820 across the five dimensions and reached 0.876 for the total score, whereas single-measure ICC(2,1) values ranged from 0.553

to 0.603 (0.702 for the total score). Dimension-level alphas ranged from 0.788 to 0.822 across the five performance dimensions, providing supplementary internal-consistency evidence.

DATA ANALYSIS METHOD

We found no extreme outliers in the main observed variables ($|Z| > 3.29$). We found no significant deviation from the normal distribution for gain scores (Shapiro-Wilk $W = 0.993$, $p = .297$). The Harman single-factor variance was 28.44% for students and 41.99% for teachers. We analyzed RQ1 by performing paired-samples t-tests of the pre- and post-rubric scores and the exploratory expression-versus-embodied contrast along with the demographic ANOVAs. In the case of RQ2, we applied one-sample t-tests with the neutral value of 3 and teacher-subgroup ANOVAs. For the secondary correlational question (RQ3), we entered perceived usefulness, perceived ease of use, and usage frequency into a hierarchical regression analysis separately for students and teachers. Specifically, we tested the hypothesis by introducing perceived usefulness and perceived ease of use into Block 1 and usage frequency into Block 2 of the regression model. For H3, we used the coefficient of the usage-frequency predictor and its additional contribution to explain variance. We calculated Pearson's correlations for the descriptive purposes; Spearman's rank-order was used to evaluate the same associations using the four-point ordinal measure of usage frequency. We also explored the relationship between the usage frequency and the blind-scored performance gains with Pearson's correlation. Finally, we report effect sizes as dz for paired analyses, Cohen's d for one-sample tests, and η^2 for multivariate and subgroup effects. All statistical analyses were conducted with SPSS 29 and partly in R and Python.

FINDINGS AND DISCUSSION

Findings are reported by research question, each discussed immediately in relation to prior research.

RESEARCH QUESTION 1: STUDENT PERFORMANCE OUTCOMES

Students' blind-scored total performance rose from a pre-test mean of 14.06 (SD = 2.49) to a post-test mean of 16.02 (SD = 3.58), a gain of 1.96 points (SD = 2.64, 95% CI [1.63, 2.29]), or approximately 14%. As Table 2 shows, the total-score effect was moderate to large, $t(246) = 11.65$, $p < .001$, $dz = 0.74$, 95% CI [0.60, 0.88], but the dimension-level gains were sharply asymmetric: musical expression improved by dz = 0.64, whereas the three embodied dimensions improved only modestly (dz = 0.24-0.26; Table 2). H1a was therefore supported. An exploratory paired contrast compared each student's expression gain with the mean embodied gain: 0.672 versus 0.281, $t(246) = 5.09$, $p < .001$, $dz = 0.324$. The contrast was statistically significant but was interpreted as exploratory rather than as confirmation of a prespecified hypothesis. One-way ANOVAs found no significant subgroup differences: year of study and instrument were null ($p = .296$ and .684); years of learning was marginal but non-significant, $F(3, 243) = 2.48$, $p = .062$.

TABLE 2. Paired-samples t-tests of performance scores

Dimension	Pre-test M (SD)	Post-test M (SD)	Mean difference	t	p	dz
Pitch & rhythm accuracy	2.74 (0.93)	3.00 (1.02)	0.267	3.70	< .001	0.24
Fluency	2.82 (0.87)	3.11 (1.07)	0.296	4.16	< .001	0.26
Tone quality & control	3.02 (0.92)	3.30 (1.07)	0.279	3.89	< .001	0.25
Musical expression	2.60 (0.80)	3.27 (1.01)	0.672	10.04	< .001	0.64
Overall performance	2.89 (0.86)	3.34 (1.24)	0.445	5.35	< .001	0.34
Total score	14.06 (2.49)	16.02 (3.58)	1.96	11.65	< .001	0.74

Note. n = 247, df = 246. Values are M (SD); dz = Cohen's d for paired samples. The total score is the sum of the five dimension scores (possible range 5-25). Dimension means are rounded independently and may not sum exactly to the total.

Set against the classroom anchor of Deng et al. (2025), dz = 0.74 was numerically similar to the pooled g = 0.71 reported for graded studies, but the estimates arise from different designs and are not directly equivalent. Embodied-skill constraints were evident at the dimension level, where interpretive gain (dz = 0.64) exceeded technical gains (dz = 0.24-0.26) by approximately two and a half times. The single-group design warrants interpreting the observed pretest-posttest effect alongside maturation, concurrent regular training and other time-related explanations, particularly given the short, four-week assisted period. Although exploratory rather than prespecified, the asymmetry is consistent with the division-of-labour framework in the review: audio systems can support error shaping but do not provide interpretive guidance (Sheldon et al. 1999; Waddell & Williamon 2019).

The asymmetry complicates rather than simply confirms the theory. Deliberate practice as originally formulated is teacher directed, weakness targeted, and involves immediate corrective feedback (Ericsson et al.

1993). The changes were larger in the interpretive than in the embodied technical score. This suggests that although text-based support can help organize practice, it may not completely recreate the embodied feedback loop involving the instrument, ear, and teacher. Frequent low-pressure feedback may also have strengthened self-efficacy, possibly reflected in bolder expressive playing (Bandura 1986), but self-efficacy was not measured and this remains an inference.

RESEARCH QUESTION 2: TEACHER-PERCEIVED INSTRUCTIONAL EFFICIENCY

Teachers scored each of the three H2 constructs significantly higher than the neutral midpoint (Table 3): assessment reliability had the largest construct-level effect (d = 0.87), followed by preparation time and lesson accuracy. Overall evaluation as an additional outcome showed the largest standardized effect (M = 3.73, d = 1.16). H2 was supported at the level of perception. An exception was found, however, for perceived assessment reliability, which varied significantly depending on the level of students taught, $F(3, 48) = 3.20$, $p = .032$, $\eta^2 = .17$: teachers of intermediate students rated it M = 3.00, exactly the neutral point, while all other groups rated it above 3.6 on average.

TABLE 3. One-sample t-tests of teacher-perceived outcomes

Construct	M	SD	t	p	Cohen's d
Preparation time	3.51	0.75	4.89	< .001	0.68
Lesson accuracy	3.50	0.76	4.72	< .001	0.65
Assessment reliability	3.62	0.71	6.29	< .001	0.87
Overall evaluation	3.73	0.63	8.36	< .001	1.16

Note. n = 52, df = 51. One-sample t-tests against the scale midpoint of 3; d = Cohen's d.

Within the identified English sources, these are, to our knowledge, among the first quantitative studies exploring teachers' perceptions concerning music performance pedagogy; all investigate perceived rather than demonstrated effectiveness, and endorsements at the middle may be indicative of acquiescence bias. The pattern fits emerging qualitative literature suggesting teachers feel most supported in cutting down preparation time by using generative artificial intelligence (Belloula 2025; Lee et al. 2024), developing lesson plans with ChatGPT (Rutherford et al. 2025), and adopting it for assessment purposes (Khlaif et al. 2024). The relative emphasis on assessment reliability reflects the tool's demonstrated ability to approximate human consistency (Awidi 2024) and the challenges in music marking to achieve inter-rater reliability (Wesolowski et al. 2016). The only category of judgments that appeared lower than one would expect from the progression hypothesis concerned intermediate performances, possibly because of the unambiguous nature of beginner mistakes and the well-documented character of advanced interpretation, leaving intermediate assessments with a less objective basis than those at other levels. This represents only a hypothesis, however, as the sample size for this cell was small (n = 9). Overall, teachers' responses suggest intermediate-level student performance assessments would benefit from human judgment.

RESEARCH QUESTION 3: CORRELATES OF INTEGRATION

Across RQ3, the dependent variable was the self-reported outcome composite (not blind-scored gain). As Table 4 illustrates, for both students and teachers, all three variables were positively correlated with the outcome (usage frequency: r = .316, .412; perceived usefulness: r = .235, .479; perceived ease of use: r = .223, .391; all $p \leq .004$). Moreover, Spearman ranks also revealed a significant usage-outcome association for students ($\rho = .312$, $p < .001$) and teachers ($\rho = .401$, $p = .003$).

The hierarchical comparison in Table 5 provided the stricter test of H3. The inclusion of usage frequency increased students' explained variance from 8.6% to 15.4% ($\Delta R^2 = .068$, $p < .001$; $\beta = .267$), supporting H3. Student full-model VIFs were roughly 1.07. For teachers, the inclusion of usage frequency increased R^2 from .278 to .321, which was not significant ($\Delta R^2 = .043$, $p = .089$; $\beta = .229$). Thus, H3 was not supported, and perceived usefulness remained the only significant final-model predictor ($\beta = .309$, $p = .028$).

TABLE 4. Pearson correlations between the integration variables and the self-reported outcome composite

Variable	Students r	Students p	Teachers r	Teachers p
Usage frequency	.316	< .001	.412	.002
Perceived usefulness	.235	< .001	.479	< .001
Perceived ease of use	.223	< .001	.391	.004

Note. Students n = 247; teachers n = 52.

TABLE 5. Hierarchical regression of self-reported outcomes on acceptance-related covariates and usage frequency

PANEL A. HIERARCHICAL MODEL COMPARISON

Sample	Block	Variables entered	R^2	ΔR^2	ΔF(df; p)
Students	1	Perceived usefulness; perceived ease of use	.086	—	—
Students	2	Usage frequency	.154	.068	19.44 (1, 243; < .001)
Teachers	1	Perceived usefulness; perceived ease of use	.278	—	—
Teachers	2	Usage frequency	.321	.043	3.01 (1, 48; .089)

PANEL B. FINAL-MODEL COEFFICIENTS

Predictor	B	SE	β	t	p
Students (n = 247)					
Usage frequency	0.14	0.03	.267	4.41	< .001
Perceived usefulness	0.08	0.03	.174	2.87	.004
Perceived ease of use	0.06	0.03	.127	2.07	.040
Teachers (n = 52)					
Usage frequency	0.17	0.10	.229	1.74	.089
Perceived usefulness	0.17	0.07	.309	2.27	.028
Perceived ease of use	0.15	0.10	.199	1.51	.137

Note. Block 1 included perceived usefulness and perceived ease of use; Block 2 added usage frequency. Panel B reports final Block 2 coefficients. B = unstandardised coefficient; SE = standard error; β = standardised coefficient. Full models: students, $F(3, 243) = 14.72$, $p < .001$; teachers, $F(3, 48) = 7.55$, $p < .001$.

Supplementary MANOVAs based on prior ChatGPT use showed no effects for students, Wilks' Lambda = 0.998, $F(3, 243) = 0.19$, $p = .905$, or teachers, Wilks' Lambda = 0.925, $p = .287$. Because these analyses were supplementary, they did not change the conclusions for RQ1-RQ3.

A supplementary analysis examined whether usage frequency was associated with blind-scored performance gain. The linear association was positive but nonsignificant ($r = .096$, $p = .132$, $n = 247$), while perceived usefulness was significantly associated with performance gain ($r = .186$, $p = .003$). This analysis did not test H3, which concerned the self-reported outcome composite.

Accordingly, the follow-up performance analysis was treated as supplementary and did not change the test or interpretation of H3.

On the self-reported side, behavioural engagement was related to AI-related outcomes beyond attitudes and exposure events (Lo et al., 2025), whereas ease of use appeared less important after initial adoption barriers (Tiwari et al., 2024). This pattern cautions against treating ever-users and never-users as groups with fundamentally different intentions. Blind-scored follow-up data qualified the self-reported association: the evidence was tentative, and self-reports remained vulnerable to common-method variance. Student-reported outcomes were most closely related to usage, whereas teacher-reported outcomes were most closely related to perceived usefulness. This nuanced pattern supports an associational rather than definitive causal interpretation of learning gains and adoption.

CONCLUSION

Students' blind-scored performance increased from pretest to posttest during the ChatGPT-assisted period ($dz = 0.74$), while teachers reported positive perceptions of instructional efficiency, led by assessment reliability. Usage frequency retained a significant adjusted association with student outcomes but not teacher outcomes; perceived usefulness was the only significant predictor in the full teacher model. The linear association between usage frequency and blind-scored gain was nonsignificant. Taken together, the findings support the effectiveness of ChatGPT as a supervised pedagogical assistant, particularly for interpretive and planning-related activities, but not as a substitute for embodied teacher feedback. Theoretically, the study extends technology-acceptance research into the performing arts and suggests an interpretive-embodied asymmetry as a hypothesis-generating pattern for future controlled research. Practically, students should use ChatGPT as a critically evaluated practice aid, teachers may employ it for rubric-aligned support and lesson planning under human oversight, and institutions should provide structured and equitable training.

The study has a number of limitations. First, using a single group, it was not possible to distinguish the effects of maturation or concurrent regular training from changes during the ChatGPT-assisted period. In addition, 65.2% of students had used ChatGPT before participating in the study; its use was not prohibited during baseline, and no prompt scripts, minimum usage, or logs were specified, limiting replication. Second, the ethical and cognitive risks mentioned in the literature were not assessed. Third, teacher inferences in the main study remain limited by n = 52, possible acquiescence bias in midpoint comparisons, and the use of focused single-item global indicators of perceived usefulness and perceived ease of use. Construct validity of the teacher outcome questionnaire received preliminary support from the four-factor EFA, but this conclusion requires further investigation in a larger sample because there were 52 respondents for 17 items. In addition, self-efficacy was not measured. Fourth, the majority of participants were male (63.6%), the literature search was limited to English publications, and the assisted period lasted only four weeks in the Chinese higher-education context, which may constrain generalizability. Therefore, future research should use a randomized longitudinal design, include platform logs, replicate instrument validity in a larger sample, directly assess the cognitive and ethical risks described in the literature, measure self-efficacy, and examine the generalizability of results across countries and disciplines.

REFERENCES

Acikgul, K. & Sad, S.N. 2026. Educators' acceptance and use of AI software: UTAUT2 model. *SAGE Open* 16(1): 1-19.

Ali, O., Murray, P.A., Momin, M., Dwivedi, Y.K. & Malik, T. 2024. The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change* 199: 123076.

Allingham, E. & Wöllner, C. 2022. Slow practice and tempo-management strategies in instrumental music learning: Investigating prevalence and cognitive functions. *Psychology of Music* 50(6): 1925-1941.

Awidi, I.T. 2024. Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence* 6: 100226.

Bandura, A. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs: Prentice-Hall.

Bell, A.P., Jiang, R. & Daley, M. 2025. Music education and artificial intelligence: A conversational editorial. *Action, Criticism, and Theory for Music Education* 24(3): 1-15.

Belloula, S. 2025. Empowering educators: Leveraging AI to revolutionize lesson planning. *International Journal of Research in Education and Science* 11(2): 264-280.

Bergkvist, L. & Rossiter, J.R. 2007. The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research* 44(2): 175-184.

Boateng, G.O., Neilands, T.B., Frongillo, E.A., Melgar-Quinonez, H.R. & Young, S.L. 2018. Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health* 6: 149.

Carey, G. & Grant, C. 2015. Teacher and student perspectives on one-to-one pedagogy: Practices and possibilities. *British Journal of Music Education* 32(1): 5-22.

Chu, T.S. & Ashraf, M. 2025. Artificial intelligence in curriculum design: A data-driven approach to higher education innovation. *Knowledge* 5(3): 14.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd Ed. New York: John Wiley & Sons.

Davis, F.D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13(3): 319-340.

Deng, R., Jiang, M., Yu, X., Lu, Y. & Liu, S. 2025. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education* 227: 105224.

Deruty, E., Grachten, M., Lattner, S., Nistal, J. & Aouameur, C. 2022. On the development and practice of AI technology for contemporary popular music production. *Transactions of the International Society for Music Information Retrieval* 5(1): 35-49.

Ericsson, K.A., Krampe, R.T. & Tesch-Römer, C. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review* 100(3): 363-406.

Gaunt, H. 2008. One-to-one tuition in a conservatoire: The perceptions of instrumental and vocal teachers. *Psychology of Music* 36(2): 215-245.

Gaunt, H. 2010. One-to-one tuition in a conservatoire: The perceptions of instrumental and vocal students. *Psychology of Music* 38(2): 178-208.

Gerlich, M. 2025. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies* 15(1): 6.

Kamalov, F., Santandreu Calonge, D. & Gurrib, I. 2023. New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability* 15(16): 12451.

Khlaif, Z.N., Ayyoub, A., Hamamra, B., Bensalem, E., Mitwally, M.A.A., Ayyoub, A., Hattab, M.K. & Shadid, F. 2024. University teachers' views on the adoption and integration of generative AI tools for student assessment in higher education. *Education Sciences* 14(10): 1090.

Korthals Altes, T., Willemse, M., Goei, S.L. & Ehren, M. 2024. Higher education teachers' understandings of and challenges for inclusion and inclusive learning environments: A systematic literature review. *Educational Research Review* 43: 100605.

Krejcie, R.V. & Morgan, D.W. 1970. Determining sample size for research activities. *Educational and Psychological Measurement* 30(3): 607-610.

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D. & Palmer, E. 2024. The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence* 6: 100221.

Legi, M.B., Sengkay, D.F. & Sambul, A.M. 2024. An application of generative artificial intelligence for automated rubric-based grading. *Jurnal Teknik Informatika* 19(3): 183-192.

Lerch, A., Arthur, C., Pati, A. & Gururani, S. 2020. An interdisciplinary review of music performance analysis. *Transactions of the International Society for Music Information Retrieval* 3(1): 221-245.

Lo, N., Wong, A. & Chan, S. 2025. The impact of generative AI on essay revisions and student engagement. *Computers and Education Open* 9: 100249.

Madeira, L.R.B., Araujo, M.V., Hein, C.F. & Marinho, H. 2018. Adaptation of a self-regulated practice behaviour scale for Portuguese music students. *Psychology of Music* 46(6): 795-812.

Miksza, P. 2012. The development of a measure of self-regulated practice behavior for beginning and intermediate instrumental music students. *Journal of Research in Music Education* 59(4): 321-338.

Merino-Campos, C. 2025. The impact of artificial intelligence on personalized learning in higher education: A systematic review. *Trends in Higher Education* 4(2): 17.

Nah, F.F.-H., Zheng, R., Cai, J., Siau, K. & Chen, L. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research* 25(3): 277-304.

O'Leary, E. 2025. Considering the possibilities and problems of AI in music education: The need for critical literacies. *Action, Criticism, and Theory for Music Education* 24(3): 138-164.

Pati, K.A., Gururani, S. & Lerch, A. 2018. Assessment of student music performances using deep neural networks. *Applied Sciences* 8(4): 507.

Perez-Montesdeoca, H., Rodriguez-Rodriguez, D., Stendardi, D. & Fernandez-Sogorb, A. 2026. Design and validation of a questionnaire on teachers' uses of generative artificial intelligence. *Computers and Education Open* 10: 100332.

Rashid, A.B. & Kausik, M.A.K. 2024. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Advances* 7: 100277.

Ritchie, L. & Williamon, A. 2011. Measuring distinct types of musical self-efficacy. *Psychology of Music* 39(3): 328-344.

Ritchie, L. & Williamon, A. 2013. Measuring musical self-regulation: Linking processes, skills, and beliefs. *Journal of Education and Training Studies* 1(1): 106-117.

Rutherford, T., Rodrigues, A., Duque-Baird, S., Veng, S., Mykyta-Chomsky, R., Cao, Y., Chisholm, K. & Bergwall, E. 2025. "I just think it is the way of the future": Teachers' use of ChatGPT to develop motivationally-supportive math lessons. *Computers and Education: Artificial Intelligence* 8: 100367.

Sheldon, D.A., Reese, S. & Grashel, J. 1999. The effects of live accompaniment, intelligent digital accompaniment, and no accompaniment on musicians' performance quality. *Journal of Research in Music Education* 47(3): 251-265.

Shi, M. 2019. The effects of class size and instructional technology on student learning performance. *The International Journal of Management Education* 17(1): 130-138.

Supriyono, Wibawa, A.P., Suyono & Kurniawan, F. 2024. Advancements in natural language processing: Implications, challenges, and future directions. *Telematics and Informatics Reports* 16: 100173.

Tiwari, C.K., Bhat, M.A., Khan, S.T., Subramaniam, R. & Khan, M.A.I. 2024. What drives students toward ChatGPT? An investigation of the factors influencing adoption and usage of ChatGPT. *Interactive Technology and Smart Education* 21(3): 333-355.

Vieriu, A.M. & Petrea, G. 2025. The impact of artificial intelligence (AI) on students' academic development. *Education Sciences* 15(3): 343.

Waddell, G. & Williamon, A. 2019. Technology use and attitudes in music learning. *Frontiers in ICT* 6: 11.

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T. & Du, Z. 2024. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications* 252: 124167.

Wesolowski, B.C., Wind, S.A. & Engelhard, G., Jr. 2016. Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception* 33(5): 662-678.

Wiese, L.J., Patil, I., Schiff, D.S. & Magana, A.J. 2025. AI ethics education: A systematic literature review. *Computers and Education: Artificial Intelligence* 8: 100405.

Yu, C., Yan, J. & Cai, N. 2024. ChatGPT in higher education: Factors influencing ChatGPT user satisfaction and continued use intention. *Frontiers in Education* 9: 1354929.

Zhai, C., Wibowo, S. & Li, L.D. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments* 11: 28.

Yucheng Ma
Faculty of Education,
Universiti Kebangsaan Malaysia,
p154544@siswa.ukm.edu.my

Amelia Alias*
Faculty of Education,
Universiti Kebangsaan Malaysia,
amelia.alias@ukm.edu.my

Nurfaradilla Mohamad Nasri
Faculty of Education,
Universiti Kebangsaan Malaysia,
nurfaradilla@ukm.edu.my

*Corresponding Author: amelia.alias@ukm.edu.my

Dihantar: 20 Julai 2026
Dinilai: 5 Ogos 2026
Diterima: 20 Ogos 2026
Diterbitkan: 1 November 2026