City University eJournal of Academic Research (CUeJAR)
e-ISSN : 2682-910X
CUeJAR Homepage: https://www.city.edu.my/CUJAR

OPEN ACCESS

Curriculum Standards to Immersive Practice: Development and Preliminary Evaluation of a Virtual Reality Rhythm Module for Secondary School Music Education

Keran Li[1]*, Raja Zulkarnain Raja Yusof[2] & Hao Yi Ho[3]

Faculty of Creative Industries, City University of Malaysia, Selangor, Malaysia

Abstract: *Immersive technology is often introduced into music education through a technology-first process in which an application is selected before its curricular purpose and assessment logic are defined. This study instead developed and preliminarily evaluated a curriculum-first virtual reality (VR) rhythm module for secondary school music education. A four-stage design and development research process translated curriculum standards into a ten-week system linking VR task families, teacher guidance, weekly lesson scripts, fidelity checkpoints, and measures of beat stability, strong-weak beat perception, rhythm-pattern accuracy, and textbook-song transfer. The retained archive contained favourable aggregate content-appraisal indices (I-CVI = .95; S-CVI = .92), although panel metadata and item-level decisions required for procedural replication were unavailable. The human-rated transfer measure showed good agreement (weighted kappa = .82; ICC = .89). Fifty-four students were allocated within one school to VR instruction or matched conventional teaching (27 per group). In the VR group, mean attendance was 9.70 of 10 sessions, fidelity was 95.67%, module completion was 94.63%, and mean immersive time was 27.72 minutes per 45-minute lesson. Gain-score effects favoured VR across the four outcomes (d = 2.450 to 3.636), with noncentral-t 95% confidence intervals of 2.410 to 4.046 for beat stability, 2.488 to 4.151 for strong-weak beat perception, 2.751 to 4.506 for rhythm-pattern accuracy, and 1.732 to 3.154 for textbook-song transfer. Separate ANCOVA models controlling for the corresponding pre-test were significant for all four outcomes (all p < .001). Partial eta squared was .742 for beat stability, .742 for strong-weak beat perception, .774 for rhythm-pattern accuracy, and .615 for textbook-song transfer; the archived F statistic for beat stability was unavailable. These setting-specific findings provide preliminary support for the curriculum-first model while requiring cautious interpretation and independent replication.*

Keywords: *curriculum alignment; virtual reality; rhythm pedagogy; secondary music education; design and development research; effect size; implementation fidelity*

Copyright © City University Press.

CUeJAR
Received: 6th July 2026
Revised: 18th August 2026
Accepted: 7th September 2026

1. Introduction

Secondary-school rhythm instruction involves more than counting beats and repeating short patterns. Learners must maintain a steady pulse, distinguish strong and weak beats, control duration and grouping, and apply these skills in curriculum repertoire. These competencies are related but distinct. A student may maintain tempo while misplacing metric accents, or reproduce a short pattern accurately but lose its organisation within a song. The curriculum design therefore preserved both the distinction among the four competencies and their progression toward integrated performance.

Virtual reality can support this progression by coordinating sound, visual cues, space, and movement while providing repeated practice and immediate timing feedback. These affordances are educational only when task design directs attention to the intended musical relationship and limits unnecessary cognitive load (Dede, 2009; Jensen & Konradsen, 2018; Mayer, 2021; Merchant et al., 2014). This distinction is especially important in music education because a technically engaging activity may not match the metre, repertoire, or performance forms that teachers assess. Systematic reviews report promising results alongside weaknesses in experimental design, pedagogical specification, and reporting (Hamilton & Alperović, 2021; Radianti et al., 2020). Experimental studies also show that immersion or engagement does not guarantee learning and may impede it when learners face excessive processing demands or insufficient guidance (Makransky et al., 2019; Parong & Mayer, 2018).

The central question is therefore not simply whether VR can be used for rhythm training, but how curricular expectations should guide cues, tasks, progression, feedback, teacher action, and assessment from the start of development. A technology-first process may create an attractive practice environment without explaining why success in the application represents progress in school music. A curriculum-first process begins with what students must know and do, defines acceptable evidence of that learning, and then determines how immersion can support the required actions. Learning outcomes, activities, and assessments were therefore designed as one aligned system following constructive alignment and backward design (Biggs & Tang, 2011; Wiggins & McTighe, 2005).

This article documents the design and development component of a doctoral project on curriculum-aligned VR rhythm learning and reports bounded downstream evidence. Its primary contribution is the intervention architecture: how four DDR stages translated curriculum and theory into a working module, how expert appraisal and measurement planning informed development, and whether classroom records support feasibility. The theoretical lenses were used prospectively to set design requirements rather than retrospectively to explain observed effects. Outcome reporting is limited to four prespecified learning domains and two implementation indicators. Gain-score Cohen's d represents the standardised between-group difference in improvement, whereas ANCOVA partial eta squared represents the adjusted post-test group effect after controlling for the corresponding pre-test.

Four research questions guided the article:

1. What design requirements follow from the target rhythm curriculum, intended performance outcomes, and constraints of a secondary-school music classroom?
2. How can these requirements be translated into an aligned system of curriculum competencies, VR tasks, teacher mediation, weekly progression, and assessment?
3. What retained evidence describes the module's content appraisal, measurement readiness, and implementation fidelity?
4. What patterns of learning gain and pre-test-adjusted group effects occur across the four curriculum-linked outcomes?

The article provides a design account that can be examined and adapted and tests whether the resulting system produced coherent implementation and outcome signals. Its claims are limited to the curriculum, school, sample, and ten-week implementation. They do not establish the effectiveness of all VR rhythm applications or remove the need for independent replication.

2. Research Background and Design Framework

2.1 Rhythm learning as an embodied and hierarchical process

Rhythm depends on coordinated temporal processes. Beat tracking supports temporal prediction, metre organises events hierarchically, and performance requires motor responses that conform to that structure. Sensorimotor synchronisation and embodied music learning are therefore part of rhythmic knowledge rather than optional additions to it (Juntunen & Hyvönen, 2004; London, 2012; Repp & Su, 2013). Familiar classroom activities such as clapping, tapping, stepping, and conducting help learners organise time through action. Experimental evidence that body movement can influence subsequent metre perception further supports an embodied component of rhythmic understanding (Phillips-Silver & Trainor, 2005).

Embodied activity can make rhythm instruction engaging, but it also constrains the design. Movement, visual displays, and sound must correspond to the musical structure. Prominent animation, rewards, or spatial effects can draw attention away from the target skill. Cognitive Load Theory distinguishes the intrinsic demands of simple and compound metre from avoidable processing demands. The module therefore used synchronized cues, brief instructions, stable response rules, and feedback directed at musically relevant errors to reduce extraneous visual and procedural load (Mayer & Moreno, 2003; Moreno & Mayer, 2007; Sweller et al., 2011). Immersive presentation may increase interest or arousal without improving learning when learners cannot process the experience or when guidance is inadequate (Makransky et al., 2019; Parong & Mayer, 2018).

Embodied learning and cognitive load meet in the coordination of action, perception, and feedback. When movement, the heard beat, and the visual response share the same temporal

structure, learners have fewer arbitrary mappings to maintain between the interface and the musical concept. The action-feedback loop can then direct attention to pulse, accent, and grouping rather than add a separate mechanical demand. By contrast, a gesture that changes the virtual environment without representing the intended beat has little musical value. The module therefore combined a single-task cognitive-economy rule with an embodied-correspondence principle that distinguishes content-congruent gestures from movements used only to operate an interface (Johnson-Glenberg, 2018). The Cognitive-Affective Model of Immersive Learning similarly treats presence and agency as affordances whose effects depend on subsequent cognitive and affective processing (Makransky & Petersen, 2021).

The module applied cognitive economy by making one primary relationship salient in each task: pulse regularity, accent hierarchy, pattern organisation, or transfer to a song excerpt. Multimodal information was included only when it supported that relationship.

2.2 Curriculum alignment and transfer

The target curriculum requires students to work with 2/4, 3/4, 4/4, 3/8, and 6/8 metres and apply rhythmic understanding in textbook-based musical activities (Ministry of Education of the People's Republic of China, 2022). Although the range contains only five metres, it includes different beat groupings and contrasts between simple and compound organisation. Recognising the time signature alone is therefore insufficient.

Curriculum alignment connected four elements: learning objectives, activities designed to achieve them, observable student responses, and assessments of those responses. This approach avoided weak alignment based only on repeating labels from the textbook. For example, displaying 6/8 notation does not demonstrate alignment unless the task requires learners to perceive and apply compound-duple organisation.

Transfer was planned from the design stage because estimates can be inflated when training and assessment use the same digital tasks. Transfer requires learners to apply a developed competence in a new but related context (Barnett & Ceci, 2002). The module therefore progressed from isolated pulse and metre tasks to integrated patterns and textbook-song excerpts, and the song outcome was assessed with a separate analytic performance rubric. However, the retained archive does not establish whether the BSTT, SBMT, and RPAT used items, task formats, or apparatus fully independent of the VR practice tasks. Train-test overlap for those three instruments therefore cannot be ruled out.

2.3 Teacher orchestration and TPACK

The intervention was an orchestrated set of teaching materials rather than a standalone software product. TPACK holds that effective educational technology integrates content, pedagogy, and technological conditions (Mishra & Koehler, 2006). In this study, music-content requirements

determined the target material; pedagogical knowledge informed sequencing, scaffolding, correction, and classroom transitions; and technological knowledge guided the use of embodied responses and feedback without disrupting the lesson. Research on music-specific VR likewise identifies responsive action-sound mapping, usability, and performer-centred evaluation as central design concerns (Serafin et al., 2016).

The teacher therefore remained central to delivery. Each week, the teacher introduced the objectives and safety procedures, organised participation and headset transitions, provided corrective feedback, and connected the immersive tasks to textbook study. The guide specified these essential functions without scripting every utterance, allowing the teacher to respond professionally to learners and classroom conditions while preserving intervention fidelity (Bellg et al., 2004; Toomey et al., 2020). Classroom orchestration also required deliberate management of equipment, time, participation structures, and transitions (Dillenbourg, 2013).

2.4 Theoretical lenses as design functions

A combined framework of Self-Determination Theory (SDT), Cognitive Load Theory (CLT), TPACK, and embodied or multimodal learning addressed distinct design problems. Graduated challenge and informative feedback were intended to make progress visible and support competence (Ryan & Deci, 2000). CLT guided the reduction of unnecessary interface demands and the focus on one musical relationship at a time. TPACK connected music content, teacher practice, and technological conditions. Embodied and multimodal learning guided the temporal coordination of movement, sound, visual cues, and correction.

These theories were translated into design features rather than tested as mediating mechanisms. The framework therefore provides a prospective rationale for the intervention package, not a causal model explaining differences in outcomes.

2.5 Six integrated design principles

The literature and curriculum analysis produced six design principles. Table 1 expresses each principle as an operational rule linked to a specific design risk.

Table 1: Design Principles and Operating Rules for the VR Rhythm Module.

Design principle	Operational rule	Main design risk addressed
Curriculum first	Derive every task from an identified meter, rhythm, or transfer competency before selecting an immersive interaction.	Technology-driven activities with weak curricular relevance

Design principle	Operational rule	Main design risk addressed
One task, one primary relation	Make pulse, accent, grouping, pattern, or transfer the dominant object of attention in each activity.	Competing cues and ambiguous learning targets
Embodied correspondence	Synchronize movement, auditory events, and visual feedback so that they express the same temporal structure.	Decorative movement that does not support musical understanding
Cognitive economy	Limit interface complexity, stabilize response rules, and deliver feedback on the musically relevant error.	Extraneous load and novelty distraction
Teacher orchestration	Integrate task use with explanation, transitions, safety procedures, correction, and textbook application.	Treating software as a substitute for pedagogy
Transfer and fidelity by design	Build song applications, lesson scripts, attendance records, log summaries, and fidelity checks into the module from the outset.	Assuming transfer or implementation quality after the event

Together, the principles define a traceable design chain connecting curriculum, tasks, teacher action, student response, assessment, and implementation evidence.

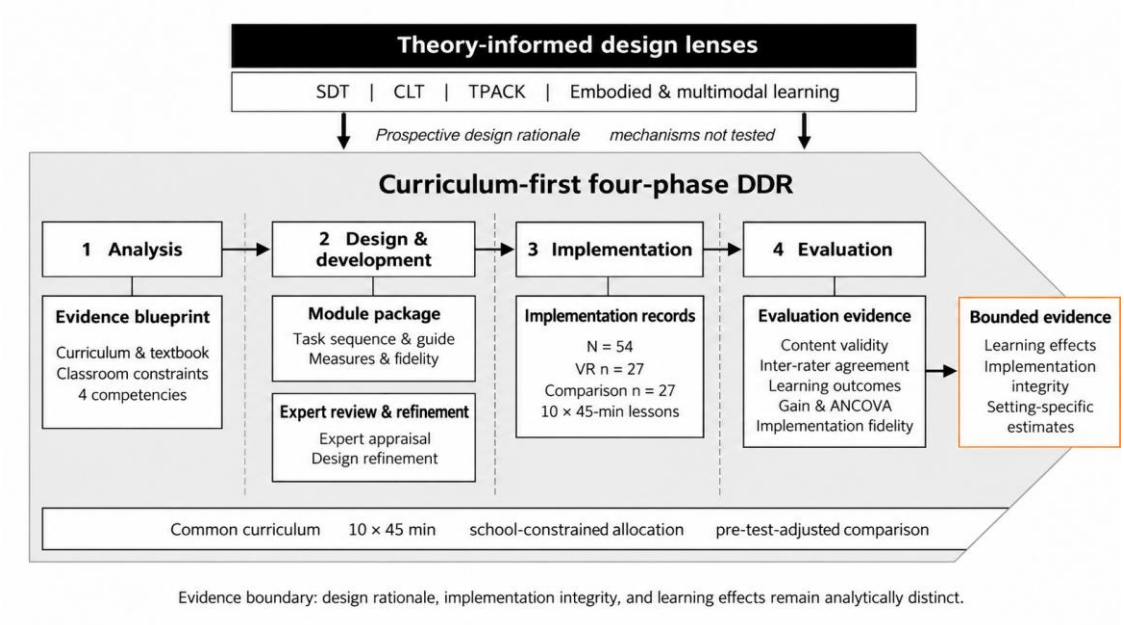


Fig. 1 | Curriculum-first DDR, allocation, and evidence architecture. a, Four DDR phases translate the curriculum and four prospective design lenses into a reviewed module, delivery records, and differentiated evaluation; the theories were not tested as mediators. b, The

implemented sample comprised 54 students under school-constrained allocation (27 VR; 27 conventional), with common curriculum targets across ten 45-minute lessons. c, BSTT, SBMT, RPAT, and the song rubric map to four competencies; process indicators describe implementation rather than learning. Outcome estimates are presented in Fig. 2.

3. Methodology

3.1 Research design, DDR process, and scope

Design and development research (DDR) was selected because the study aimed to produce an educational intervention and document the logic linking its requirements, construction, implementation, and evaluation (Richey & Klein, 2007). DDR guided the development and refinement of the module, followed by a school-based quasi-experiment. This article reports the lesson architecture, teacher guidance, implementation records, and bounded outcome evidence. Educational design research similarly combines analysis, construction, evaluation, and implementation in authentic settings through iterative refinement (McKenney & Reeves, 2018; Wang & Hannafin, 2005).

The four stages were analysis, design and development, implementation, and evaluation (Fig. 1a). Analysis produced a competency and evidence blueprint; design and development produced the reviewed module package; implementation generated delivery and exposure records; and evaluation examined design, delivery, and bounded outcome evidence. Figure 1b shows the implemented allocation and shared instructional conditions. The pre-implementation check was a classroom-timing rehearsal, not an individual-student pilot, and it generated no student outcome data.

3.2 Context and design requirements

The module was designed for lower-secondary music instruction in Liaocheng, Shandong Province, China. The analysis identified a fixed 45-minute weekly lesson, variation in learner readiness, supervised headset use, device rotation, transition and sanitation requirements, and the need to retain teacher-led music instruction. Open-ended immersion was excluded because it would be difficult to standardise and connect to weekly curriculum goals.

Curriculum and textbook review identified required actions involving pulse, metre, accent, grouping, pattern execution, and song performance. Classroom observation, teacher consultation, and timetable analysis informed constraints on task length, vocabulary, difficulty, physical response, device management, and feedback. These activities supplied design inputs rather than formal interview or qualitative data. An assessment blueprint was prepared before task construction so that each design objective had a planned source of evidence.

3.3 Curriculum-to-task-to-assessment mapping

The design distinguished four competence domains. Beat stability refers to maintaining a steady pulse in timed responses. Strong-weak beat perception refers to recognising and enacting the accent hierarchy of the five target metres. Rhythm-pattern accuracy refers to performing the durations, groupings, and accents of short structured sequences. Textbook-song transfer refers to applying these competencies in curriculum-linked excerpts.

Each domain was linked to a task family and a corresponding measure: the Beat Stability Task Test (BSTT), Strong-Weak Beat Meter Test (SBMT), Rhythm Pattern Accuracy Test (RPAT), and an analytic textbook-song transfer rubric (Fig. 1c). This mapping made it possible to distinguish weak learning from a mismatch between instruction and assessment.

Table 2. Curriculum-to-Task-to-Assessment Architecture

Competency	Curriculum demand	VR task family	Observable response	Assessment evidence
Beat stability	Sustain an even pulse through short musical activity	Pulse matching, synchronization, and steady-response tasks	Temporal consistency across a fixed task	BSTT score
Strong-weak beat perception	Distinguish and enact accent hierarchy in 2/4, 3/4, 4/4, 3/8, and 6/8	Accent selection, movement cueing, and meter-contrast tasks	Timed accent selection or movement cue that differentiates strong- and weak-beat positions	SBMT total and meter-component scores
Rhythm-pattern accuracy	Perform durations, groupings, and accents in structured figures	Pattern imitation, guided response, and error-correction sequences	Accurate execution of short rhythmic patterns	RPAT score
Textbook-song transfer	Apply rhythmic control in curriculum-linked repertoire	Integrated rehearsal and excerpt-based application	Pulse continuity, accent accuracy, grouping coherence, and rhythmic control	Analytic transfer rubric

3.4 Expert review and pre-implementation refinement

A small expert panel appraised curriculum alignment, task clarity, difficulty progression, lesson feasibility, teacher guidance, and assessment blueprints. Because the experts did not participate in the student outcome study, their evidence concerned the intervention package rather than learning effects. The retained record identifies the review domains but does not preserve a complete comment-by-comment revision history.

A classroom-timing rehearsal was also conducted before implementation. The teacher practised the introduction, task instructions, monitoring, transitions, and lesson closure. Retained notes document revisions to prompts, timing, and the order of supporting materials. No student completed the outcome instruments during this rehearsal, and no pilot outcome data were collected. Because the item-level decision log was not retained, a more detailed sequence of expert-driven revisions cannot be reconstructed.

Historical content appraisal was summarised with item-level and scale-level content validity indices (I-CVI and S-CVI) following established guidance (Lynn, 1986; Polit & Beck, 2006; Polit et al., 2007). The retained archive does not contain the number or profiles of reviewers, response-scale anchors, decision thresholds, or complete item-level ratings and decisions. It is unknown whether these details were never recorded or were lost during later document organisation. The four-point rubric used to score student transfer performance was separate from the expert-review scale. Consequently, the aggregate CVI values are reported as historical indices rather than as a reproducible validation procedure. For the human-rated transfer measure, raters were trained with common descriptors and anchor examples, and weighted kappa and the intraclass correlation coefficient (ICC) were used to assess agreement and consistency.

3.5 Ten-week sequence and teacher preparation

The ten-week sequence progressed from pulse and basic metre to a wider range of metres and patterns, followed by integrated pattern work and textbook-song performance. Both groups received the same weekly content goals and lesson duration, allowing delivery mode to vary while curriculum exposure remained matched.

Table 3. Ten-week curriculum and intervention sequence

Week	Meter/content focus	Primary learning emphasis	Application link
1	Pulse orientation	Baseline pulse control and device onboarding	Clapping and counting patterns
2	2/4	Strong-weak beat contrast	Short textbook-linked figures

Week	Meter/content focus	Primary learning emphasis	Application link
3	3/4	Triple-meter hierarchy and grouped accent	Pattern reproduction in triple meter
4	4/4	Stable pulse across longer sequences	Phrase-based counting and movement
5	3/8	Simple-compound discrimination	Short compound-meter excerpts
6	6/8	Compound-duple beat feel and accent	Pattern accuracy in compound meter
7	Mixed-meter review	Contrast, retrieval, and targeted correction	Matched error-correction tasks
8	Integrated patterns	Accurate response under greater complexity	Preparation for excerpt transfer
9	Song application	Rhythm integration in textbook repertoire	Rubric-guided rehearsal
10	Consolidation	Integrated review and performance preparation	Textbook-song transfer task

Teacher preparation covered the study purpose, weekly goals, equipment organisation, lesson scripts, transitions, feedback priorities, fidelity recording, and responses to minor technical failures. A rehearsal tested whether the introduction, focused practice, guided application, and closure could fit within the scheduled lesson. Backup devices and cleaning procedures were incorporated into the implementation plan.

3.6 Implementation evidence and ethics

As shown in Fig. 1b, 54 students were included in a structured within-school allocation: 27 received VR instruction and 27 received matched conventional instruction. The allocation sought to balance prior music participation and pre-test performance, but the archived materials do not permit reconstruction of the assignment algorithm and do not document full institutional, class-level, or individual randomisation. The design was therefore treated as quasi-experimental and used pre-tests and covariate adjustment. Attendance and classroom participation were recorded for both groups; VR duration, module completion, and task errors were recorded only for the VR group.

The bounded outcome analysis reported mean pre-post gains and 95% confidence intervals. Cohen's d represented the standardised between-group difference in gain scores. Separate ANCOVA models compared post-test scores after controlling for the corresponding pre-test,

with partial eta squared as the adjusted group-effect estimate. Confidence intervals for d were reconstructed by inverting the noncentral t distribution of the reported gain-score t statistic (df = 52; n = 27 per group), rather than by bootstrapping. Partial eta squared is reported as a point estimate because the archived model output did not support verified interval reconstruction. No a priori power calculation was retained, and observed post hoc power was not reported because it adds no information beyond the test result and confidence interval (Hoenig & Heisey, 2001). The gain-score d and covariate-adjusted ANCOVA effect answer different questions and should not be conflated (Morris, 2008). Confidence intervals are emphasised rather than threshold labels alone (Lakens, 2013).

This study was conducted in accordance with local educational research ethics guidelines for research involving minors in school settings. School administrative approval was obtained prior to implementation. Written informed consent was collected from parents or legal guardians, and verbal assent was obtained from all participating students. Only aggregate data are reported in the manuscript to protect participant privacy. All participants could withdraw from the study or discontinue use of the headset at any time without penalty. Equipment was disinfected after each use following hygiene protocols.

4. Results

4.1 A traceable intervention architecture

The first design objective was an auditable chain from curriculum requirements to classroom evidence. Each competence was linked to a musical object, task family, expected student response, and assessment. This chain addressed two common sources of ambiguity: selecting technically impressive activities without curricular relevance and using broad outcomes, such as general participation or platform satisfaction, to infer rhythm learning.

The mapping was specified before implementation. Weeks 9 and 10 did not introduce song performance as an unrelated activity after isolated training. Earlier metre and pattern tasks prepared learners to apply the same rhythmic relationships in curriculum excerpts. Transfer was therefore designed as progressive recontextualisation beyond the isolated VR practice tasks.

4.2 Preliminary expert appraisal and measurement readiness

The retained expert-review summary reported I-CVI = .95 and S-CVI = .92. These indices indicate a favourable appraisal of the intervention materials, but the missing panel size, reviewer profiles, scale anchors, decision rules, and item-level ratings prevent independent verification. They are therefore treated as preliminary evidence of content representation, not as evidence of learning effects or a fully reproducible validation study.

The textbook-song transfer measure showed substantial rater agreement (weighted kappa = .82) and strong consistency (ICC = .89) after rater preparation and calibration. These coefficients support the reproducibility of ratings for pulse continuity, accent accuracy, grouping coherence, and rhythmic control. Unlike the incompletely archived CVI process, the retained scoring method and rater-agreement statistics were sufficiently documented for reporting.

Table 4. Design-appraisal, Measurement, and Implementation Indicators

Evidence domain	Indicator	Value	Design interpretation
Content representation	I-CVI	.950	Archived aggregate; panel method, and item ratings were not retained
Content representation	S-CVI	.920	Archived aggregate; scale-level calculation cannot be independently reproduced
Transfer scoring	Weighted kappa	.820	Substantial agreement between raters
Transfer scoring	ICC	.890	Strong consistency of transfer scores
VR implementation	Attendance	9.70 of 10 sessions	Planned instructional exposure was largely retained
VR implementation	Fidelity adherence	95.67%	Essential weekly objectives and procedures were delivered consistently
VR implementation	Module completion	94.63%	Most assigned task progression was completed
VR implementation	Mean time-on-task	27.72 minutes	Immersive practice occupied a substantial but bounded part of each 45-minute lesson
VR implementation	Mean recorded errors	3.15	Aggregate correction opportunities; error categories were not retained

4.3 Feasibility within ordinary lesson constraints

Implementation records indicated a mean VR attendance of 9.70 of 10 sessions, 95.67% fidelity, 94.63% module completion, and a mean immersive time of 27.72 minutes per 45-minute lesson. The remaining 17.28 minutes covered onboarding, transitions, teacher explanation, guided textbook transfer, closure, and hygiene. Because the archive does not disaggregate that time, no more precise allocation is reported. VR formed one component of the lesson rather than the entire lesson.

The mean recorded error count was 3.15. The planned correction sequence combined immediate feedback with teacher prompts that related each attempt to pulse, accent, metre, or pattern. However, the archived logs contain only total errors and do not distinguish musical errors from task-execution or technical errors. No retrospective anecdotes were added because no qualitative protocol had been specified.

4.4 Learning gain and adjusted effect profile

Both groups improved. Mean gains for VR and control were 19.52 and 10.22 for beat stability, 21.44 and 12.56 for strong-weak beat perception, 21.07 and 10.33 for rhythm-pattern accuracy, and 24.56 and 15.19 for textbook-song transfer. The corresponding gain-score effects were d = 3.235 (95% CI [2.410, 4.046]), 3.326 [2.488, 4.151], 3.636 [2.751, 4.506], and 2.450 [1.732, 3.154], respectively (all $p < .001$). At baseline, between-group comparisons of the four outcome pre-tests were not significant ($p = .389$ to $.791$; $|d| = .072$ to $.236$). The archived materials did not retain equivalent analyses for age, gender, or previous musical training. Two archived VR post-test summaries were 79.52 +/- 4.86 and 80.67 +/- 6.09; because their verified scale bounds and participant-level ranges were unavailable, ceiling and floor effects cannot be excluded.

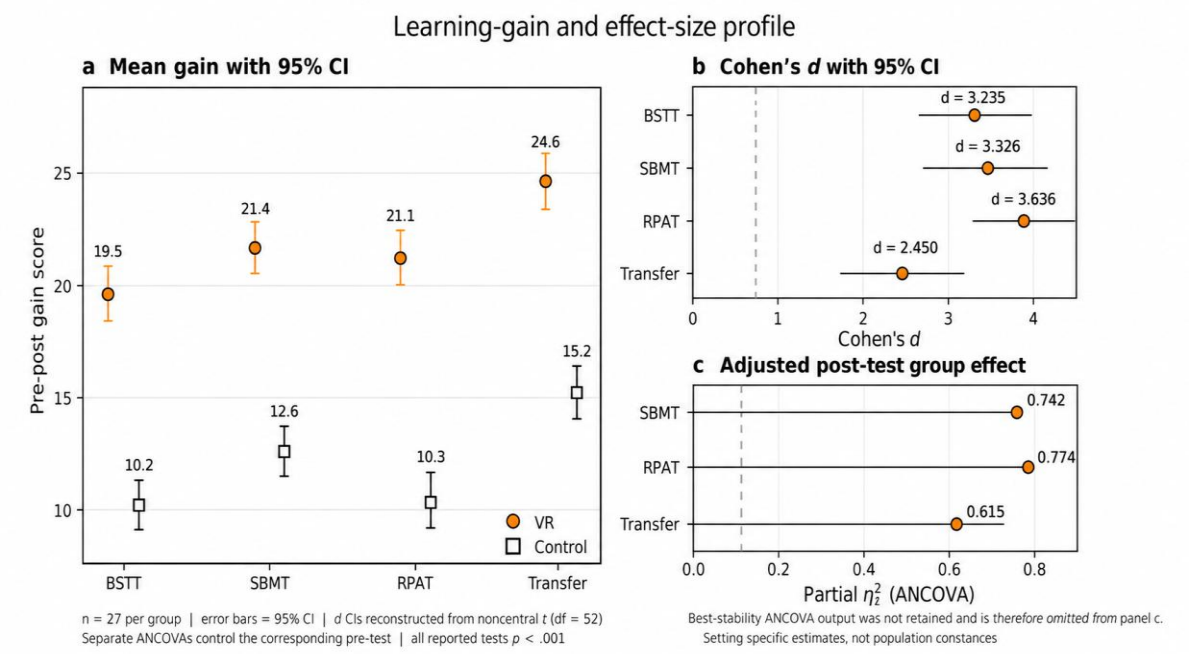


Fig. 2 | Learning gain and effect-size profile after the quasi-experimental intervention. a, Mean pre-post gains and 95% confidence intervals for the VR and control groups (n = 27 per group); participant-level distributions were not reconstructed. b, Cohen's d is the standardised between-group difference in gain scores; 95% confidence intervals were reconstructed by inverting the noncentral t distribution (df = 52). c, Partial eta squared values are the adjusted post-test group effects from ANCOVA models controlling for the corresponding pre-test where archived output was retained. The beat-stability ANCOVA statistic and effect estimate could not be

independently verified from the archived records and are therefore not interpreted as confirmed ANCOVA evidence. Dashed lines mark conventional large-effect reference values (d = .80; partial eta squared = .14; Cohen, 1988). Estimates are setting-specific.

Separate ANCOVA models controlling for the corresponding pre-test were significant for all four outcomes (all $p < .001$). The retained ANCOVA effect estimates were partial eta squared = .742 for strong-weak beat perception, .774 for rhythm-pattern accuracy, and .615 for textbook-song transfer. The corresponding beat-stability ANCOVA F statistic and effect estimate could not be independently verified from the archived output. The retained F statistics were 146.563 for strong-weak beat perception, 174.628 for rhythm-pattern accuracy, and 81.309 for textbook-song transfer. The beat-stability F statistic was not retained and was not reconstructed.

4.5 Documented refinements and classroom fit

Experts reviewed curriculum alignment, task clarity, difficulty progression, lesson feasibility, teacher guidance, and assessment blueprints. Classroom-timing rehearsal led to documented changes in prompts, transition timing, and the order of supporting materials. These records support feasibility within a 45-minute lesson but do not reconstruct the complete revision history.

Because no item-level revision table was retained, not every final feature can be linked to a specific comment or review round. The DDR contribution therefore rests on the documented architecture and retained evidence, without presenting unrecorded revisions as findings.

5. Discussion

5.1 Curriculum-first design as a constraint and a benefit

Curriculum alignment was most useful as a design constraint. It limited the target metres, defined the sequence of skills, excluded unrelated pitch and melody outcomes, and required performance-based evidence. These limits narrowed the technological design space while making the intervention easier to teach and audit. A teacher or reviewer can trace why each task was included and what evidence counted as successful learning.

This criterion also changes how immersive educational design is judged. Interface richness and interaction volume are insufficient unless sound, movement, visual cues, feedback, and teacher explanation all support pulse, metre, pattern, or transfer. Recent reviews similarly emphasise pedagogy, experimental design, and reporting quality rather than immersion alone (Hamilton et al., 2021; Radianti et al., 2020).

5.2 Embodiment requires cognitive economy

Embodied interaction can support rhythmic timing when movement, sound, and visual feedback represent the same musical relationship. When visual motion, spatial effects, or rewards are incongruent with the music, they may distract learners. The module therefore assigned one primary relation to each task: pulse tasks emphasised temporal regularity, metre tasks emphasised accent hierarchy, pattern tasks emphasised grouping and duration, and transfer tasks integrated these competencies. This design follows evidence that coordinated movement can support rhythmic perception while distinguishing musically congruent action from decorative interaction (Johnson-Glenberg, 2018; Phillips-Silver & Trainor, 2005).

This rule connects embodied learning with Cognitive Load Theory. It does not remove the inherent difficulty of rhythm; it reduces interface demands unrelated to the learning goal. The distinction is important in classrooms with varied prior preparation because an unfamiliar interface may widen differences if some students must learn both new technology and new musical concepts. It is also consistent with models that separate technological affordances from the cognitive and affective processes through which those affordances may influence learning (Makransky & Petersen, 2021).

The correction loop followed the same logic. Immediate auditory or visual feedback identified an error and allowed another attempt. If the mismatch persisted, the teacher gave a musical prompt or demonstration related to the relevant temporal structure rather than merely directing the learner to repeat an interface action. This sequence connected automated feedback with musical explanation.

5.3 The teacher remains part of the intervention

High fidelity and completion describe the orchestrated teaching package, not the software alone. Teacher preparation, lesson scripts, transitions, safety procedures, corrective priorities, and textbook links all contributed to implementation. Consistent with TPACK, technology was embedded in content-specific teaching and professional judgement (Koehler & Mishra, 2009; Mishra & Koehler, 2006). Managing multiple activities and devices was itself part of classroom orchestration (Dillenbourg, 2013).

Headset procurement alone is therefore insufficient for implementation. Schools also need teacher preparation, equipment logistics, supervision, and integration with regular classroom activity. In Week 6, for example, VR practice of compound-duple accent in 6/8 was followed by a planned teacher-led connection to a curriculum excerpt. This sequence was specified in advance and formed part of the 17.28 minutes outside immersive practice.

5.4 Unusually large effect estimates require a setting-bound interpretation

The study used different indicators for distinct evidential questions: content appraisal for representation, rater agreement for scoring, fidelity for enactment, gain-score d for standardised improvement, and partial eta squared for adjusted group effects. Delivery consistency does not establish the magnitude or cause of learning effects. The study did not test SDT, CLT, TPACK, or embodied learning as mediators, so the framework supports the design rationale but not post hoc causal claims about individual mechanisms.

The d values of 2.450 to 3.636 are unusually large and should not be treated as stable population effects. The sample came from one school, the outcomes closely matched the practised competencies, and the retained archive cannot verify that the BSTT, SBMT, and RPAT were independent of the VR task formats or apparatus. If practice and scoring overlapped in the VR condition, condition-specific familiarity could explain part of the effect. The textbook-song transfer outcome and ten-week duration reduce, but do not eliminate, concern about a brief novelty effect; novelty was not measured. Broad VR meta-analyses are not direct comparators because d depends on the measure and within-group dispersion (Merchant et al., 2014). The confidence intervals remain above zero but are wide, ranging from 1.732 to 3.154 for transfer and 2.751 to 4.506 for rhythm-pattern accuracy. These estimates describe one setting and require independent replication. Interpretation should therefore emphasise the metric, uncertainty, and possible train-test overlap rather than labels attached to the point estimates (Lakens, 2013; Morris, 2008).

5.5 Limitations and future development

The study has several limitations. First, the intervention was developed for one curriculum and one school; use elsewhere would require remapping the metre sequence, repertoire, and lesson logistics. Second, the archived expert-appraisal record contains only aggregate CVI values and lacks reviewer identities and profiles, scale anchors, decision thresholds, item-level ratings, and revision records. It is unknown whether these details were never recorded or were later lost, so the appraisal cannot be fully reproduced. Future studies should retain panel composition, rating rules, item-level decisions, and reasons for revision. Third, the sample was small, no a priori power plan was retained, and the allocation procedure cannot be fully reconstructed. Selection and site effects therefore remain possible. Baseline equivalence on outcome pre-tests does not establish equivalence in demographics or previous training. Given the magnitude of the effects, future work should independently recompute the standardised estimates and inspect raw-score distributions and scale bounds.

Fourth, the process indicators were limited. Time on task, completion, errors, attendance, and fidelity describe exposure and delivery but do not identify learning mechanisms. Future logs should distinguish pulse drift, accent-placement errors, duration or grouping errors, task-execution errors, and technical interruptions, while recording cue use and repeated attempts

under appropriate privacy safeguards. The fixed ten-week sequence also did not adapt to differences in student preparation.

Future adaptive versions could reduce tempo or return a learner to pulse matching after a defined instability threshold, then increase difficulty after several stable trials. Such rules were not implemented in this study and would require validation, fairness monitoring, auditability, and teacher override. Physiological or eye-tracking measures should be collected only for specific research questions, with separate consent and data-minimisation safeguards.

6. Conclusion

A four-stage DDR process produced a curriculum-aligned VR rhythm package linking tasks, teacher actions, assessments, and fidelity evidence. The retained expert-appraisal summary, rater-agreement statistics, and implementation records provide preliminary evidence on content representation, scoring readiness, and classroom feasibility. Gain scores and pre-test-adjusted ANCOVA results favoured VR across the four outcomes, although the estimates are setting-specific and may be influenced by train-test overlap.

The main contribution is a traceable curriculum-to-task-to-assessment design logic supported by bounded implementation and outcome evidence. It is not evidence that immersion is generally effective, nor is the archived validation record fully reproducible. Replication across schools and repertoires should retain complete ethics, allocation, instrument-independence, expert-review, and raw-score documentation so that design quality, implementation integrity, and learning outcomes can be evaluated separately.

Acknowledgements

The authors thank City University Malaysia, the participating school, music teachers, expert reviewers, raters, students, and families for their academic, advisory, and implementation support.

Funding Statement

City University Malaysia and the participating school provided institutional and in-kind support for academic guidance, technology, and implementation. No grant identifier was available in the retained study records.

Ethics and Participant Protection Statement

This study was conducted in accordance with local educational research ethics guidelines for research involving minors in school settings. School administrative approval was obtained prior to implementation. Written informed consent was collected from parents or legal

guardians, and verbal assent was obtained from all participating students. Only aggregate data are reported in the manuscript to protect participant privacy. All participants could withdraw from the study or discontinue use of the headset at any time without penalty. Equipment was disinfected after each use following hygiene protocols.

Data Availability Statement

This article reports design, measurement, implementation, and bounded outcome indicators from the school-based project. De-identified participant-level data are not publicly available because the study involves minors and the retained materials do not document institutional approval for public data sharing. Requests may be addressed to the corresponding author and remain subject to applicable institutional, consent, and privacy requirements.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest in the publication of this paper.

Relationship to Related Outputs

This manuscript presents the design and development research (DDR) component of the broader project, focusing on the curriculum-first intervention architecture, teacher orchestration framework, and implementation fidelity evidence. Aggregate outcome data are included only as a preliminary downstream coherence check of the designed system.

Complementary outcome-focused analyses based on the same implementation cohort will be reported separately in subsequent work. That work will present full outcome tables, subgroup comparisons, and correlational patterns, with a primary focus on learning effect profiles and their associated factors.

The two works address distinct research questions, adopt different analytical perspectives, and have no substantive overlap in core contributions or conclusions. A formal cross-citation will be added when the complementary outcome study is published.

AI-Assisted Editorial Support Statement

Generative AI was used for language and figure editing. The authors reviewed the resulting content and accept responsibility for the manuscript.

References

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612-637. https://doi.org/10.1037/0033-2909.128.4.612

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Ogedegbe, G., Orwig, D., Ernst, D., & Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*(5), 443-451. https://doi.org/10.1037/0278-6133.23.5.443

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Open University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(5910), 66-69. https://doi.org/10.1126/science.1167311

Dillenbourg, P. (2013). Design for classroom orchestration. *Computers & Education, 69*, 485-492. https://doi.org/10.1016/j.compedu.2013.04.013

Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education, 8*(1), 1-32. https://doi.org/10.1007/s40692-020-00169-2

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19-24. https://doi.org/10.1198/000313001300339897

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies, 23*, 1515-1529. https://doi.org/10.1007/s10639-017-9676-0

Johnson-Glenberg, M. C. (2018). Immersive VR and education: Embodied design principles that include gesture and hand controls. *Frontiers in Robotics and AI, 5*, 81. https://doi.org/10.3389/frobt.2018.00081

Juntunen, M.-L., & Hyvönen, L. (2004). Embodiment in musical knowing: How body movement facilitates learning within Dalcroze Eurhythmics. *British Journal of Music Education, 21*(2), 199-214. https://doi.org/10.1017/S0265051704005686

Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education, 9*(1), 60-70.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

London, J. (2012). *Hearing in time: Psychological aspects of musical meter* (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199744374.001.0001

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-385. https://doi.org/10.1097/00006199-198611000-00017

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to

a science lab simulation causes more presence but less learning. *Learning and Instruction, 60*, 225-236. https://doi.org/10.1016/j.learninstruc.2017.12.007
Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review, 33*, 937-958. https://doi.org/10.1007/s10648-020-09586-2
Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781316941355
Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43-52. https://doi.org/10.1207/S15326985EP3801_6
McKenney, S., & Reeves, T. C. (2018). *Conducting educational design research* (2nd ed.). Routledge. https://doi.org/10.4324/9781315105642
Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education, 70*, 29-40. https://doi.org/10.1016/j.compedu.2013.07.033
Ministry of Education of the People's Republic of China. (2022). Compulsory Education Arts Curriculum Standards (2022 Edition). Beijing Normal University Press.
Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017-1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x
Moreno, R., & Mayer, R. E. (2007). Interactive multimodal learning environments. *Educational Psychology Review, 19*, 309-326. https://doi.org/10.1007/s10648-007-9047-2
Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364-386. https://doi.org/10.1177/1094428106291059
Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology, 110*(6), 785-797. https://doi.org/10.1037/edu0000241
Phillips-Silver, J., & Trainor, L. J. (2005). Feeling the beat: Movement influences infant rhythm perception. *Science, 308*(5727), 1430. https://doi.org/10.1126/science.1110922
Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489-497. https://doi.org/10.1002/nur.20147
Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459-467. https://doi.org/10.1002/nur.20199
Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education, 147*, 103778. https://doi.org/10.1016/j.compedu.2019.103778
Repp, B. H., & Su, Y.-H. (2013). Sensorimotor synchronization: A review of recent research (2006-2012). *Psychonomic Bulletin & Review, 20*(3), 403-452.

https://doi.org/10.3758/s13423-012-0371-2
Richey, R. C., & Klein, J. D. (2007). *Design and development research: Methods, strategies, and issues*. Lawrence Erlbaum Associates.
Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68-78. https://doi.org/10.1037/0003-066X.55.1.68
Serafin, S., Erkut, C., Kojs, J., Nilsson, N. C., & Nordahl, R. (2016). Virtual reality musical instruments: State of the art, design principles, and future directions. *Computer Music Journal, 40*(3), 22-40. https://doi.org/10.1162/COMJ_a_00372
Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4
Toomey, E., Hardeman, W., Hankonen, N., Byrne, M., McSharry, J., Matvienko-Sikar, K., & Lorencatto, F. (2020). Focusing on fidelity: Narrative review and recommendations for improving intervention fidelity within trials of health behaviour change interventions. *Health Psychology and Behavioral Medicine, 8*(1), 132-151. https://doi.org/10.1080/21642850.2020.1738935
Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development, 53*(4), 5-23. https://doi.org/10.1007/BF02504682
Wiggins, G., & McTighe, J. (2005). *Understanding by design* (Expanded 2nd ed.). ASCD.