informatics

MDPI

Article

Design, Development, and Initial Evaluation of a Web-Based Music Ear-Training Assessment System in Higher Education

Warakorn Seeyo, Sayam Chuangprakhon *, Jirawat Surasa and Watchara Homhuan

College of Music, Mahasarakham University, Khamreang, Kantharawichai 44150, Maha Sarakham, Thailand; warakorn.s@msu.ac.th (W.S.); jirawat.s@msu.ac.th (J.S.); watchara.h@msu.ac.th (W.H.)
* Correspondence: sayam.c@msu.ac.th

Abstract

Online assessment of music ear training requires the coordinated delivery of auditory stimuli, musical response formats, and examination records. In this study, a web-based assessment system for higher education was designed, developed, and initially evaluated. A React interface and Firebase services integrated reusable items, audio playback, a virtual pitch-reference piano, local and cloud response saving, and instructor-controlled grading. The initial evaluation combined a three-expert content review of 50 rhythmic-dictation and chord-quality items with a 12-item questionnaire completed by 101 consenting respondents. Mean item–objective congruence was 0.953. The broad user-evaluation composite averaged 4.30 out of 5 (SD = 0.60; 95% CI = 4.18–4.41; Cronbach's $\alpha = 0.943$). Exploratory factor analysis supported one dominant questionnaire factor (KMO = 0.907; Bartlett's $\chi^2(66) = 1063.49$, $p < 0.001$; 62.65% variance explained) rather than four distinct latent dimensions. Post-review controlled technical testing across 145 trials showed 25/25 successful audio-playback requests, 25/25 same-device recoveries, 20/20 acknowledged Firestore writes, and 20/20 successful final submissions with no duplicate records or scoring discrepancies. These findings provide preliminary content-alignment, user-perception, questionnaire-structure, and selected technical-performance evidence. Examination-score reliability, item discrimination, performance on physical mobile hardware, and learning effects remain to be evaluated.

Keywords: music ear training; online assessment; higher education; web-based system; Firebase; question bank; user evaluation

1. Introduction

Music ear training helps individuals develop the ability to recognize, discriminate, and represent musical structures through listening. Assessing these skills requires a clear definition of the intended listening task and evidence that responses support the proposed score interpretation. Wolf and Kopiez developed the Musical Ear Training Assessment (META) to measure practice-dependent analytical hearing [1]. Swaminathan et al. reported normative and psychometric evidence for the Musical Ear Test (MET), which assesses melodic and rhythmic abilities [2]. Correia et al. subsequently reported evidence supporting online administration of the MET [3]. These studies provide methodological precedents for auditory assessment, although a standardized musical-ability test and achievement in a particular university course serve different purposes.

Digital tools also support the organization of aural-skills teaching and practice. Chen and Li examined the use of digital notation software in sight-singing and ear-training instruction [4], and Buonviri and Paney report the use of websites, software, and mobile

Check for updates

Academic Editor: Yang Shi

Received: 15 September 2026
Revised: 25 October 2026
Accepted: 28 October 2026
Published: 2 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

applications in high-school aural-skills instruction [5]. Pesek et al. evaluate the Troubadour platform using questionnaires and platform activity over three months, illustrating how music-learning interfaces can be studied through both reported experience and recorded interaction [6]. Chen and Lin review digital and AI-supported approaches in higher music education while identifying continuing questions about access, pedagogical integration, and cross-cultural applicability [7]. Hwang and Chu's adaptive error-feedback design offers a further example of connecting auditory tasks with interface support [8].

Within Thai higher music education, Seeyo et al. emphasize listening, imitation, and performance-based approaches in Isan folk music and identify a need for locally adaptable assessment resources [9]. The present study addresses a complementary system-development problem: managing auditory content, response records, and instructor decisions within an online environment. Broader instruments such as Gold-MSI and PROMS provide further examples of explicit construct definition and validation [10,11], but cannot substitute for evaluating the content and conditions of a specific course examination.

Online assessment also creates an information-management problem. Instructors need to associate an item with its media, answer key, examination configuration, student response, and subsequent score. Reviews of online formative assessment and online examinations identify feedback, assessment design, authentication, and integrity as interconnected concerns [12,13]. Formative-assessment research additionally shows why the availability of a score should be distinguished from feedback that students can use to improve their work [14–17]. Consequently, a music-assessment system should preserve the context of a response, as well as its numerical outcome.

User perceptions provide another, complementary source of evidence. Perceived usefulness and ease of use are established concepts in information-systems research [18]. The information-systems success model further distinguishes system quality, information quality, user satisfaction, and benefits [19]. These distinctions are useful here because a positive satisfaction rating cannot demonstrate accurate scoring, uninterrupted service, or improved musical performance. Similarly, the assessment-integrity issues discussed in *Informatics* by Kaisara and Bwalya require attention to information and procedural safeguards, not simply a claim that online monitoring prevents cheating [20].

The primary problem addressed in this study is the fragmentation of music ear-training assessment across auditory stimulus delivery, reusable item management, response preservation, scoring, and instructor review. The system integrates these functions within one browser-based workflow rather than treating them as separate teaching or testing utilities. Its primary contribution is therefore an integrated information architecture linking media, question definitions, responses, scoring, review, browser-based pitch reference, local persistence, and cloud persistence. Expert content review, questionnaire findings, post hoc exploratory factor analysis, and controlled technical benchmarks are used as complementary forms of initial evaluation evidence; they are not presented as validation of a new standardized measure of musical ability. Three research questions guide the report: (1) how can a web-based architecture support the preparation, delivery, preservation, and review of ear-training assessments; (2) what content-alignment evidence supports the initial examination form; and (3) how do respondents evaluate the content, interface, perceived benefits, and use of the system?

2. Materials and Methods

2.1. Study Design and Evaluation Framework

The study combined system development, expert content review, and a descriptive questionnaire evaluation at the College of Music, Mahasarakham University, Thailand. Design-science research provides a framework for relating the educational problem, the

software artifact, and the evaluation evidence [21,22]. This report uses that perspective to organize the implemented system and its initial evaluation; it does not claim to present a prospectively completed formal design-science protocol.

The evaluation distinguishes between content representation, user experience, and examination-score interpretation. Validity frameworks emphasize the interpretation and use of assessment scores [23,24]. Expert judgments address alignment between items and objectives, whereas questionnaire responses address perceived experience of the system. Examination responses would be needed to investigate discrimination and score reliability. Table 1 summarizes the evidence sources and their scope.

Table 1. Evidence sources and the claims they can be used to support.

Evidence Source	Unit and Scope	Supported Interpretation	Important Boundary
System implementation and documentation	Application design and workflows	Implementation description	No independent source-code audit; selected functions were evaluated in a controlled post-review benchmark under declared conditions
Expert-review summary	50 items; three reviewers	Reported item–objective alignment	Original item-by-reviewer ratings unavailable
User questionnaire	101 consenting respondents; 12 ratings	User perceptions and questionnaire consistency	Not an achievement test or task-based usability test
Implemented-system screenshots	Screenshots of the implemented instructor and student interfaces	Evidence of implemented interface functions and organization	Not operational logs, examination outcomes, or evidence of the exact software configuration experienced by every respondent

2.2. System Architecture and Data Model

The application was implemented as a single-page application using React 18.3.1, Vite 5.2.12, Firebase SDK 10.12.2, and Tailwind CSS 3.4.4. Firebase Authentication (Firebase SDK 10.12.2) manages sign-in and identity, Cloud Firestore stores structured examination records, and Firebase Storage stores uploaded audio and image assets. These managed services support the core examination workflow without a dedicated application server. Figure 1 shows the relationships among browser interaction, persistence, data storage, and instructor activities. The frontend hosting provider is not specified.

The data model comprises users, examinations, and submissions. User records contain an identifier, contact information, role, and student identifier. Examination records contain question definitions, answer keys, time settings, access restrictions, ordering settings, and feedback configurations. Submission records link an examination and respondent to answers, status, scores, manual-grading information, feedback, aggregate visibility-switch counts, and available timestamps. A reusable question bank enables instructors to store questions, retrieve existing items, and assemble new examination sets. Item preparation is therefore separated from examination assembly. A separate version-controlled item-bank schema is a proposed extension, rather than a feature of the present implementation.

Instructors can create and edit examinations, import questions, assign media, restrict access, and review submissions. Figure 2 presents the organization of these functions. Figures 2 and 3 are publication-oriented composite illustrations based on screenshots captured from the implemented online assessment system developed by the authors. They show the instructor and student interfaces and the available system functions. The records

visible in the screenshots were not included as questionnaire or examination outcome data in the present analysis.

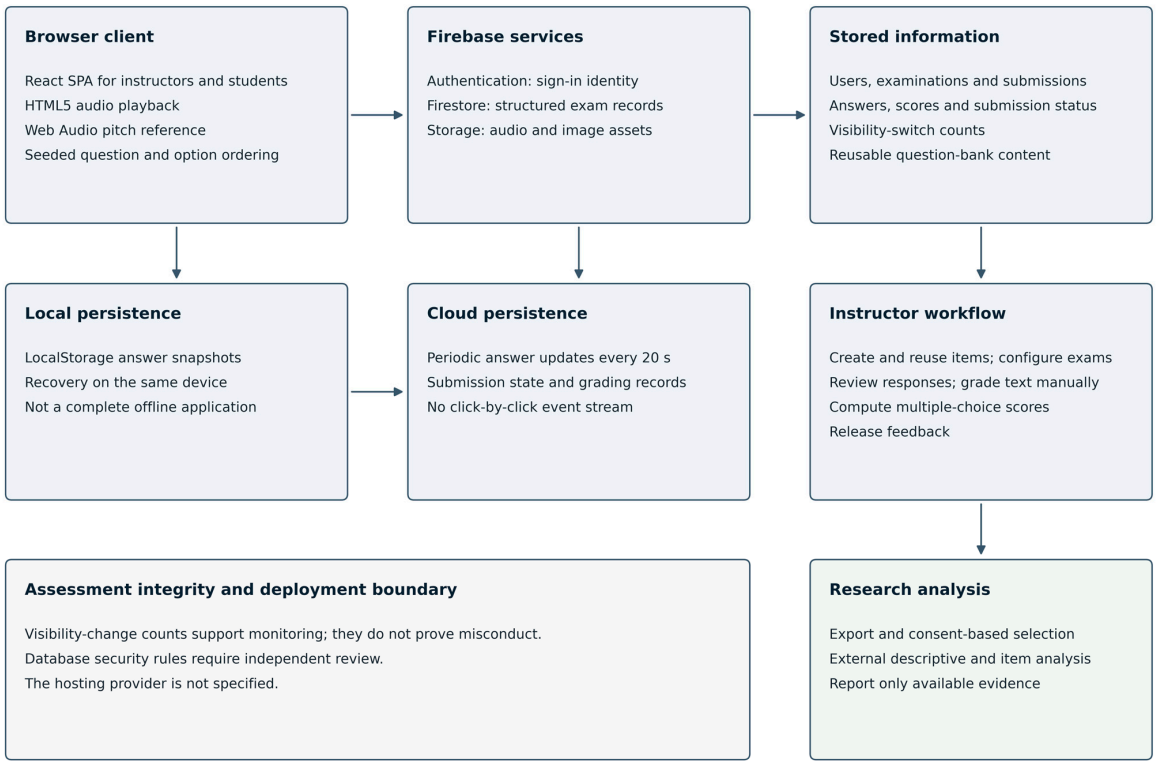


Figure 1. System architecture of the web-based music ear-training assessment platform, showing browser interaction, local persistence, Firebase services, structured data storage, instructor workflows, and external research analysis.

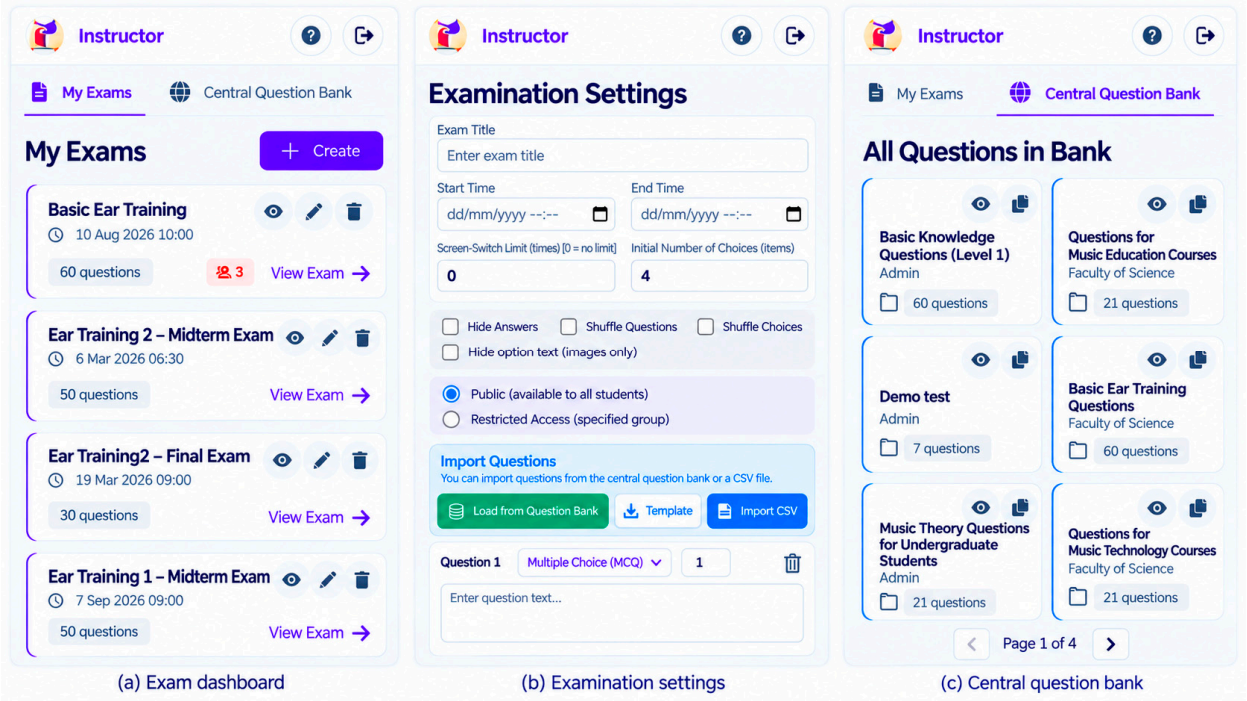


Figure 2. Publication-oriented composite illustration based on screenshots of the implemented instructor interface: (**a**) examination dashboard; (**b**) examination settings and question import; and (**c**) central question bank.

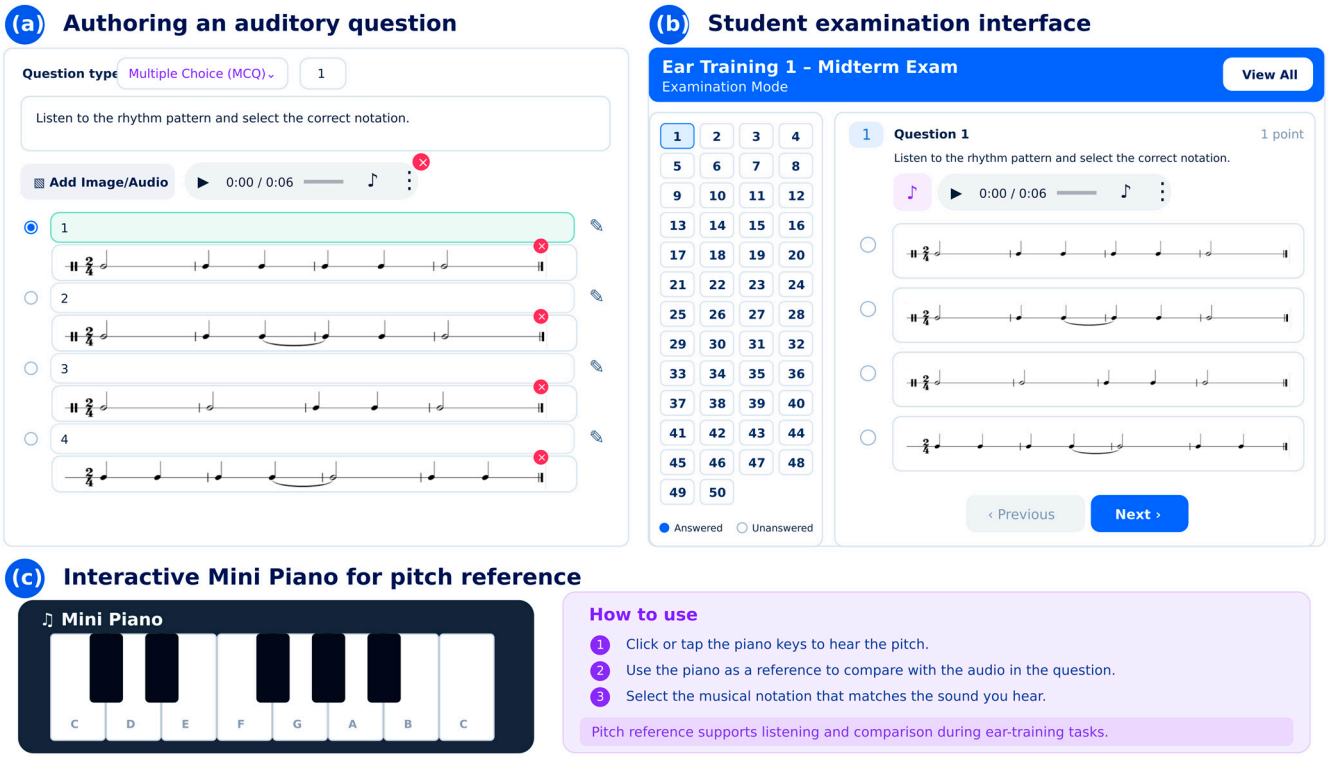


Figure 3. Publication-oriented composite illustration based on screenshots of the implemented auditory assessment interface: (**a**) question authoring; (**b**) student examination interface; and (**c**) interactive virtual piano for pitch reference. The letters C, D, E, F, G, A, and B denote musical pitch names; the virtual piano spans C4–C5.

2.3. Audio Delivery, Response Support, and Scoring

The client presents auditory stimuli through HTML5 audio controls and supports notation images and text responses. The audio preload setting is none, so media are requested on demand instead of all items being downloaded at page load. The virtual piano generates C4–C5 reference pitches in the browser through the Web Audio API; during the controlled technical evaluation, this functionality was tested in Safari 26.3 (AppleWebKit 605.1.15) It supports pitch comparison during an auditory task; it does not record or automatically evaluate singing. The original field survey did not measure media-loading latency or bandwidth. During revision, controlled tests measured browser playback-start/event timing and Web Audio dispatch timing; these measurements do not represent physical audible-onset latency or bandwidth performance.

Figure 3 illustrates item authoring, auditory item presentation, and the pitch-reference piano. A reference-assisted comparison differs from unaided pitch identification, so piano availability and replay rules are part of the assessment conditions. Subsequent examination-score evaluation should fix these conditions, the listening equipment, and the audio-production settings. The questionnaire export does not link individual respondents to a software build, piano setting, or item set, so effects of later piano and question-bank updates cannot be isolated.

Responses are saved in LocalStorage and periodically synchronized to Firestore at 20-s intervals. Recovery depends on the same device and retained browser storage; uncached media may remain unavailable during disconnection. Multiple-choice answers are compared with the answer key when the instructor initiates score calculation, and text responses require manual grading. Instructors control the release of scores and feedback. Aggregate visibility-switch counts support monitoring and configured submission policies,

but do not identify why a window lost focus or demonstrate misconduct. The examination-settings interface labels a screen-switch limit of zero as "unlimited"; this study did not independently test the enforcement of this setting. Table 2 summarizes the implemented system features and their corresponding evaluation boundaries.

Table 2. System features and corresponding evaluation boundaries.

Feature	Implemented Feature	Evaluation Boundary
Question reuse	Select saved items to assemble new examinations	No demonstrated form equivalence or exposure control
Auditory interaction	HTML5 playback; notation images; browser pitch reference	Controlled browser-event timing was measured; physical audible-onset latency and acoustic calibration were not measured
Response preservation	LocalStorage snapshots; 20-s cloud updates	Recovery remains device-dependent; controlled same-device testing achieved 25/25 successful restorations
Scoring	Instructor-triggered MCQ calculation; manual text grading	Controlled dummy submissions showed score agreement for verified keys; examination-score psychometrics and manual-grading agreement were not evaluated
Integrity monitoring	Aggregate browser-visibility switch count	Not a cheating classifier; no sensitivity/specificity established
Data access	Authentication, roles, public/restricted examinations	Authentication and protected-route behavior passed in desktop and simulated WebKit workflow checks; this does not constitute an independent penetration or security audit

2.4. Examination Content and Expert Review

The initial examination form comprised 50 items: 30 rhythmic-dictation items and 20 chord-quality items. These domains define the content reviewed in this study, although the platform can present additional types of auditory tasks. Three experts reviewed item–objective congruence (IOC), using +1 for congruent, 0 for uncertain, and −1 for incongruent judgments. Expert 1 was an Assistant Professor with a doctoral degree in music education, aged 47 years, with more than 10 years of teaching experience in music education. Expert 2 was an Associate Professor with a doctoral degree in education, approximately 50 years of age, with expertise in education, pedagogy, and instructional management and more than 15 years of teaching experience. Expert 3 was a computer lecturer, approximately 35 years of age, responsible for computer-system support and experienced in computer programming and software design, with approximately 5–6 years of teaching experience. For each item, IOC equals the sum of the three ratings divided by three. The reported acceptance threshold was two-thirds, displayed as 0.67.

The content-review analysis used the available aggregate review table. Three experts evaluated the item–objective congruence (IOC) of the 50 examination items. Counts and summary IOC values were used to describe the content alignment of the examination items with the intended learning objectives. This preliminary content evidence is distinct from evidence concerning the structure and validity of the user-evaluation questionnaire, which would require additional instrument-development and validation procedures [25].

2.5. Questionnaire, Respondents, and Ethics

The user-evaluation questionnaire yielded 102 submissions. One submission explicitly declining research participation was excluded, leaving 101 consenting respondents

with complete responses to all 12 items. The questionnaire assessed respondents' satisfaction with and perceptions of the online assessment system. The respondents comprised 100 undergraduate students and one postgraduate student; 14 respondents reported no prior ear-training study. These characteristics were retained as reported in the questionnaire, and all 101 complete and consenting responses were included in the descriptive analysis. The questionnaire was administered through Google Forms as a non-probability online survey. Students were invited to complete the questionnaire immediately after completing the online ear-training examination. Participation in the research survey was voluntary. The number of individuals invited to participate was not recorded; therefore, a response rate could not be calculated. All 101 included records had distinct normalized email addresses, although this check cannot rule out multiple accounts belonging to the same individual. Consent, analytical-sample selection, response completeness, and duplicate-entry checking are reported with reference to the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [26].

The 12 positively worded items use five response categories, coded 1–5 from the lowest to the highest degree of endorsement. They are organized in four three-item groups: content and appropriateness (Q01–Q03), interface and usability (Q04–Q06), perceived learning benefits (Q07–Q09), and satisfaction and adoption intention (Q10–Q12). Q10 is also the single explicit overall satisfaction item. The four group labels describe the questionnaire and are not empirically established factors. The instrument is not the System Usability Scale, nor is it a validated technology-acceptance-model scale. The exported final column is empty and supplies no qualitative comments; no thematic analysis is reported.

Research procedures followed the ethical principles of the Declaration of Helsinki. The study was approved under the exempt-review procedure by the Mahasarakham University Ethics Committee for Research Involving Human Subjects (English Approval No. 035-031/2026; Thai Approval No. 035-031/2569; 19 January 2026; valid through 18 January 2027). All survey records analyzed here contain affirmative electronic research consent. Names, email addresses, student identifiers, and individual timestamps are excluded from the public supplementary statistics. The public supplement provides aggregates and a covariance matrix rather than respondent-level records. The approval date precedes the recorded questionnaire period.

2.6. Statistical Analysis

Item-level analyses comprise response frequencies, means, sample standard deviations, medians, interquartile ranges, and the proportion choosing categories 4 or 5. Each conceptually defined three-item cluster score is the arithmetic mean of its items for a respondent; the broad evaluation composite is the mean of all 12 items. Standard deviations for these scores are calculated across respondent-level means, not by averaging item standard deviations. No missing-value imputation was necessary. Cronbach's alpha was calculated from the sample covariance matrix for the 12-item set and each conceptual cluster [27,28]. A high coefficient does not establish unidimensionality or construct validity [29,30]. In response to peer review, a post hoc exploratory factor analysis (EFA) was conducted on the 12 questionnaire items from the 101 consenting respondents. Sampling adequacy was assessed with the Kaiser–Meyer–Olkin (KMO) statistic and Bartlett's test of sphericity. Factor retention was evaluated using parallel analysis with 5000 randomly generated datasets matching the observed 101 × 12 data structure [31]. Factors were extracted using unweighted least squares (ULS). Because parallel analysis supported retention of one factor, no rotation was applied. Given the ordinal five-point response format, the one-factor extraction was repeated using Spearman correlations as a sensitivity analysis.

Means and sample standard deviations are reported alongside ordinal response distributions and medians [32,33]. A two-sided 95% t-based interval summarizes precision of the composite mean under an independent-response model and does not correct selection bias. The main descriptive analysis includes all 101 consenting, complete records. A post hoc sensitivity analysis restricts the sample to the 100 self-reported undergraduates, reflecting the undergraduate focus of the original proposal. Prior ear-training experience was retained as a descriptive characteristic rather than used for retrospective exclusion. This analysis rule does not establish that every respondent belonged to the originally intended course cohort; the distinction is addressed in the limitations.

No linked examinee-by-item examination responses and verified answer keys were available for this analysis. Consequently, empirical item difficulty, discrimination, corrected item–rest correlation, and examination-score reliability were not estimated. Figure 4 specifies a proposed workflow for subsequent psychometric evaluation. For item i, $P_i = C_i / N_i$, where C_i is the number of valid responses scored as correct, and N_i is the number of valid responses. Difficulty is classified using unrounded values: $p < 0.20$ indicates very difficult items; $0.20 \leq p < 0.80$ is the operationally acceptable difficulty range; and $p \geq 0.80$ indicates very easy items. The broad middle range follows published item-review practice [34], with 0.80 assigned to the very-easy category in this study's operational scheme. These categories guide the review process and are not universal validity criteria. $D = P_{upper} - P_{lower}$ denotes an upper–lower discrimination index; corrected r denotes the correlation of an item score with the remaining item total. Difficulty and discrimination should be considered together [35]. Group definitions and missing-response rules must be specified before applying these statistics, and D and corrected r require separate interpretation. KR-20 may be used to examine internal consistency for a future dichotomously scored examination [36].

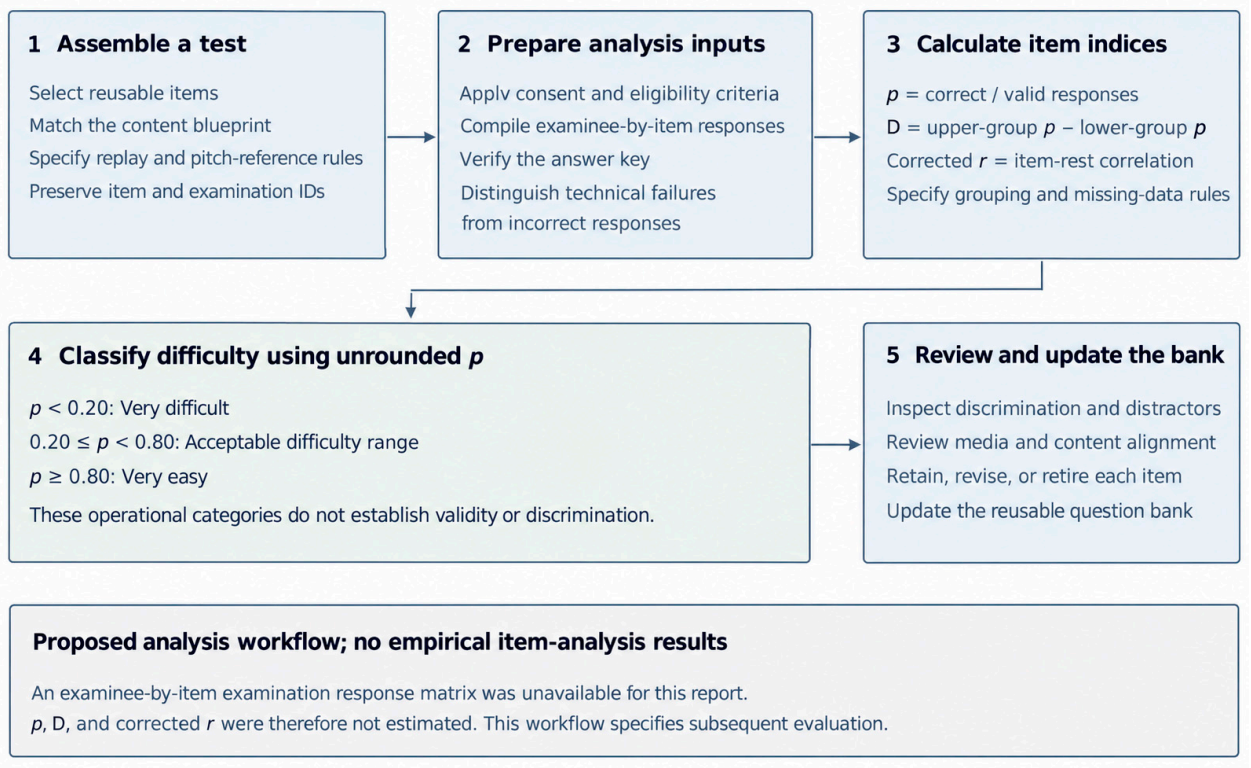


Figure 4. Proposed item-analysis and question-bank review workflow for subsequent psychometric evaluation. p denotes the difficulty index, D an upper–lower discrimination index, and corrected r an item–rest correlation. The workflow is a methodological specification rather than calculated examination results. Difficulty classification uses unrounded p values.

2.7. Controlled Technical Evaluation

In response to peer-review recommendations, a controlled technical evaluation was conducted during manuscript revision using dedicated dummy accounts and a restricted dummy examination; no real student records were used. Quantitative testing was conducted on 22–23 October 2026 with the production bundle (exam-system v0.0.0; Vite 5.2.12; React 18.3.1; Firebase SDK 10.12.2; Tailwind CSS 3.4.4) on a MacBook Pro (MacBookPro15,4; quad-core Intel Core i5, 8 GB RAM) running macOS 15.7.9. The 145 measurements comprised 87 desktop Safari and 58 simulated mobile WebKit trials (29 iPhone, 29 iPad) and covered HTML5 audio playback-start/event timing, Web Audio pitch-dispatch timing, LocalStorage persistence, same-device recovery after simulated interruptions, Firestore acknowledgment latency, and final-submission reliability. Live cloud tests used Wi-Fi broadband and Google Cloud Firestore. Complementary workflow checks used simulated iPhone 15/iOS 17.5 and iPad Pro 11-inch WebKit profiles on macOS, together with previously verified Windows 10 Chrome/Edge checks; the recorded iPad workflow user-agent reported CPU OS 12_2. The simulated mobile profiles reproduced touch-enabled viewport and user-agent conditions but did not constitute testing on physical iPhone or iPad hardware. Browser timing values represent application/event execution timing rather than physical acoustic output latency. Because the benchmark was performed after the questionnaire period, it characterizes selected functions under the declared revision-time conditions and does not establish the exact technical conditions experienced by every survey respondent.

2.8. Reproducibility and Manuscript Preparation

The supplementary JSON file preserves the original Thai questionnaire wording, and Supplement S1 provides English reporting translations. Together, the supplementary materials provide aggregate item distributions, a sample covariance matrix, executable code for reproducing the reported descriptive statistics and alpha coefficients, trial-level quantitative technical-benchmark logs, and simulated WebKit mobile-profile workflow results. The covariance structure and stated sample size support reproduction of the primary Pearson-correlation EFA and parallel-analysis procedure; the Spearman sensitivity analysis used the protected respondent-level ratings, which are not publicly distributed. Raw consent records and direct identifiers are not included. The supplementary code implements the statistical reconstruction and is not the application source code. Application credentials, production database records, and restricted assessment items are also excluded. The distinction between available statistical materials and unavailable operational artifacts is maintained throughout the data-availability statement.

ChatGPT-6, through Codex (OpenAI; accessed in 2026), assisted with drafting, literature searching, preparation of analysis code, and document formatting. Descriptive statistics and EFA summaries were calculated using deterministic analysis procedures and checked against the available questionnaire data and aggregate results. Figures 1 and 4 were drawn from system specifications and the proposed analysis procedure, and Figure 5 was plotted from observed category counts. Figures 2 and 3 were prepared as publication-oriented composite illustrations based on author-provided screenshots of the implemented online assessment system. OpenAI image-generation tools assisted with removal of personal identifiers, translation and standardization of interface labels, panel composition, and visual refinement. The authors reviewed the resulting figures against the source screenshots. Where displayed, dates were converted from the Buddhist Era to the corresponding Gregorian calendar dates without changing the underlying dates. No participant responses, examination outcomes, questionnaire values, EFA results, or technical-test results were generated or altered.

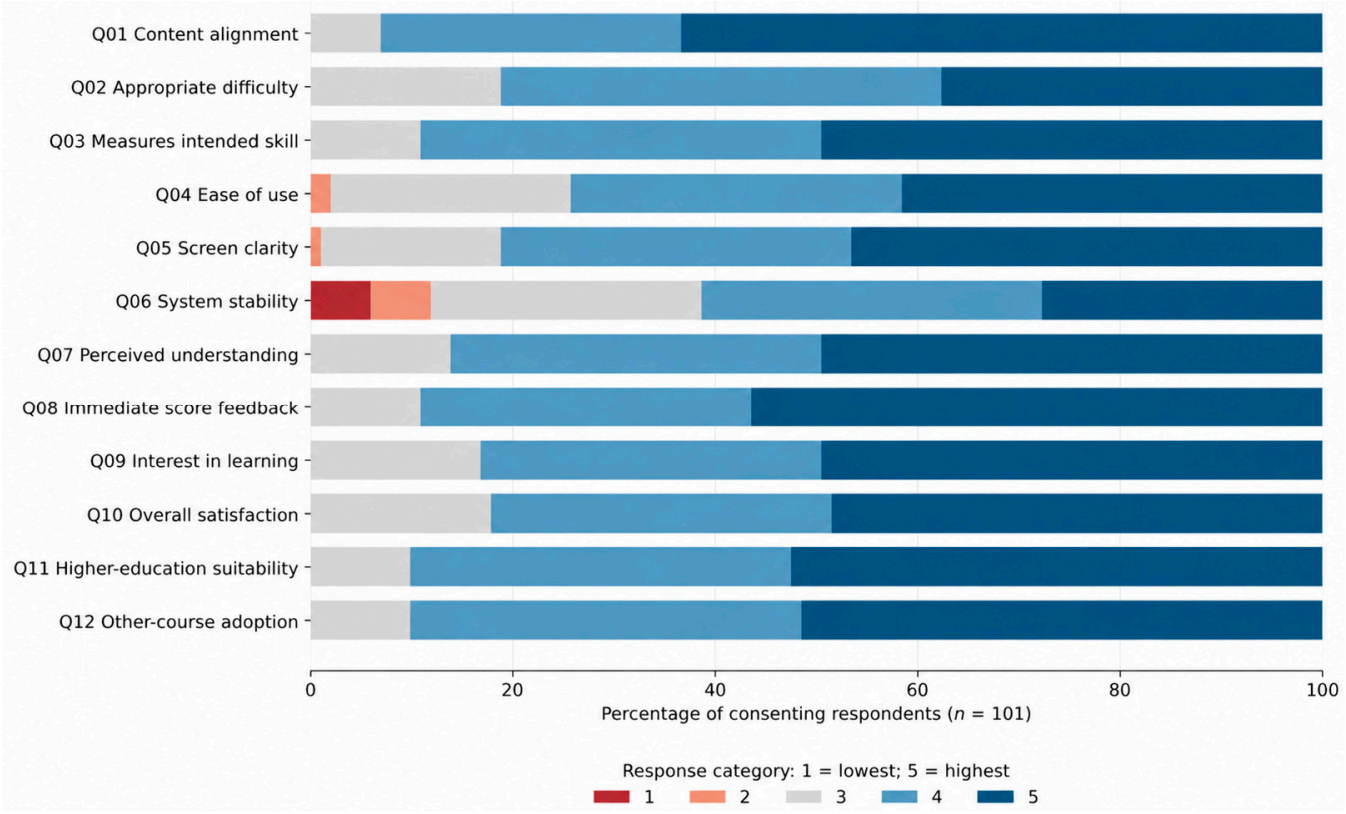


Figure 5. Distribution of questionnaire ratings for 101 consenting respondents. Five response categories are shown for each item; zero-frequency categories have no visible segment. Q01–Q12 denote the 12 questionnaire items. System stability (Q06) has the largest proportion of ratings in categories 1–3.

3. Results

3.1. Respondent Characteristics and Data Completeness

Of 102 submissions, 101 complete consenting records were included. Respondents comprised 100 undergraduates and one postgraduate. Seventy-eight reported currently studying ear training, nine reported previous study, and 14 reported never studying it. Table 3 describes this respondent sample; it is not a verified roster of the originally proposed 100-student course cohort.

Table 3. Characteristics of consenting questionnaire respondents ($n = 101$).

Characteristic	Category	*n*	%
Sex	Male	78	77.2
	Female	23	22.8
Age	18–20 years	97	96.0
	21–23 years	4	4.0
Study level	Undergraduate	100	99.0
	Postgraduate	1	1.0
Ear-training experience	Currently studying	78	77.2
	Previously studied	9	8.9
	Never studied	14	13.9

3.2. Content Alignment of the Initial Examination Form

The expert-review summary recorded 43 items with IOC = 1.00 and seven with IOC displayed as 0.67. All 50 met the reported two-thirds threshold (Table 4). Treating the latter category as exactly two-thirds yields a mean IOC of 0.953 overall, 0.956 for rhythmic dictation, and 0.950 for chord quality. These values summarize expert content judgments and do not estimate item difficulty or discrimination among examinees.

Table 4. Reported examination blueprint and aggregated item–objective congruence (IOC).

Content Domain	Item Numbers	Items	IOC = 1.00	IOC = 2/3	Mean IOC
Rhythmic dictation	1–30	30	26	4	0.956
Chord quality	31–50	20	17	3	0.950
Total	1–50	50	43	7	0.953

Note: Counts are transcribed from the existing expert-review summary. The category shown as 0.67 is treated as the rounded form of 2/3 when means are recalculated. Original item-by-reviewer ratings were not available.

3.3. Item-Level User Evaluation

The highest mean was observed for content alignment with the course (Q01: M = 4.56, SD = 0.62), followed by the perceived value of immediate score feedback (Q08: M = 4.46, SD = 0.69). The explicit overall satisfaction item averaged 4.31 (SD = 0.76), with 83 of 101 respondents selecting categories 4 or 5. System stability had the lowest mean (Q06: M = 3.71, SD = 1.12). For that item, 12 respondents selected categories 1 or 2, 27 selected category 3, and 62 selected categories 4 or 5. Thus, favorable overall perceptions coexisted with a substantial spread of stability ratings. Table 5 reports the complete item-level results, and Figure 5 shows the response distributions.

Table 5. User evaluation by questionnaire item (n = 101; response range 1–5).

Item	Statement (English Translation)	M	SD	Median (Q1–Q3)	4 or 5 (%)
Q01	The online test content is aligned with the music ear-training course.	4.56	0.62	5 (4–5)	93.1
Q02	The difficulty of the test is appropriate.	4.19	0.73	4 (4–5)	81.2
Q03	The test measures music listening skills in accordance with its objectives.	4.39	0.68	4 (4–5)	89.1
Q04	The online test system is easy to use and uncomplicated.	4.14	0.85	4 (3–5)	74.3
Q05	The screen layout is clear and easy to understand.	4.27	0.79	4 (4–5)	81.2
Q06	The system is stable, with no problems during use.	3.71	1.12	4 (3–5)	61.4
Q07	The online test helps me understand ear-training lessons better.	4.36	0.72	4 (4–5)	86.1
Q08	Receiving scores immediately helps me improve my learning.	4.46	0.69	5 (4–5)	89.1
Q09	The online test stimulates my interest in the ear-training course.	4.33	0.75	4 (4–5)	83.2

Table 5. *Cont.*

Item	Statement (English Translation)	M	SD	Median (Q1–Q3)	4 or 5 (%)
Q10	Overall, I am satisfied with using the online test.	4.31	0.76	4 (4–5)	82.2
Q11	The online test is appropriate for teaching and learning in higher education.	4.43	0.67	5 (4–5)	90.1
Q12	I would like online tests to be used in other music courses.	4.42	0.67	5 (4–5)	90.1

3.4. Composite Scores and Internal Consistency

The broad 12-item evaluation composite had an M = 4.30 and an SD = 0.60, with a 95% t-based interval of 4.18–4.41. This composite summarizes several aspects of use and should not be labeled solely as overall satisfaction. Content and appropriateness, perceived learning benefits, and satisfaction/adoption intention each rounded to 4.38, while interface/usability averaged 4.04. Cronbach's alpha for the complete 12-item set was 0.943; the four three-item coefficients ranged from 0.801 to 0.906 (Table 6). These coefficients pertain to questionnaire scores, not to the 50-item music examination.

Table 6. Respondent-level composite scores and Cronbach's alpha ($n = 101$).

Composite	Items	M	SD	Alpha
Content and appropriateness	Q01–Q03	4.38	0.59	0.844
Interface and usability	Q04–Q06	4.04	0.78	0.801
Perceived learning benefits	Q07–Q09	4.38	0.65	0.890
Satisfaction and adoption intention	Q10–Q12	4.38	0.64	0.906
Broad evaluation composite	Q01–Q12	4.30	0.60	0.943

The undergraduate-only sensitivity analysis ($n = 100$) yielded an M = 4.30, an SD = 0.60, and a 95% interval of 4.18–4.42. Before rounding, the mean was 4.2975 compared with 4.2954 in the all-consenting analysis, a difference of approximately 0.002 points. The descriptive finding is therefore insensitive to the removal of the single postgraduate record. This comparison does not verify individual course enrollment or the eligibility of respondents reporting no prior ear-training study.

3.5. Exploratory Factor Analysis of the Questionnaire

The 12-item response matrix was suitable for exploratory factor analysis (KMO = 0.907; Bartlett's test of sphericity: $\chi^2(66) = 1063.49$, $p < 0.001$). Parallel analysis supported retention of one dominant factor: the first observed eigenvalue (7.859) exceeded the 95th-percentile random reference value (1.772), whereas the second observed eigenvalue (1.088) was below its corresponding reference value (1.549). The one-factor ULS solution explained 62.65% of the standardized item variance, with loadings ranging from 0.486 to 0.875. Q06 (system stability) had the lowest loading (0.486) and communality (0.236), whereas Q11 had the highest loading (0.875). A Spearman-correlation sensitivity analysis produced the same one-factor interpretation (KMO = 0.912; loadings = 0.526–0.872). These findings do not support four empirically distinct latent dimensions in this sample. Accordingly, the four prespecified three-item groupings are retained only as conceptual clusters for descriptive reporting.

3.6. Controlled Technical Benchmark Results

Across 145 revision-time WebKit trials (87 desktop Safari; 29 simulated iPhone; 29 simulated iPad), all 25 HTML5 audio playback requests completed successfully. Browser playback-start/event timing averaged 119.67 ms (SD = 43.31; range = 81.21–252.75 ms). Web Audio pitch-dispatch timing across 30 trials averaged 0.05 ms (SD = 0.02; range = 0.04–0.12 ms). LocalStorage serialization/write execution across 25 trials averaged 0.03 ms (SD = 0.01; range = 0.02–0.03 ms). All 25 same-device recovery trials restored the expected 50-item state, with recovery execution averaging 0.04 ms (SD = 0.01; range = 0.03–0.08 ms) and no missing or duplicated answers. Firestore acknowledgment latency across 20 controlled dummy writes averaged 754.64 ms (SD = 288.73; range = 373.08–1335.10 ms), with 20/20 acknowledged writes and no recorded write loss. Final-submission testing across 20 trials averaged 689.28 ms (SD = 243.91; range = 386.79–1090.82 ms) and produced 20/20 acknowledged submissions, no duplicate records, and no scoring discrepancies. Complementary 11-step workflows passed all 11 checks in each of two simulated mobile Safari profiles (22/22 checks overall), covering login, examination loading, HTML5 audio, the Web Audio mini piano, answer modification, LocalStorage autosave, reload recovery, layout overflow, final submission, score verification, and post-logout protected-route behavior. These mobile checks used simulated WebKit profiles on macOS rather than physical iPhone or iPad hardware. Previously verified Windows Chrome and Edge checks provide additional desktop-browser context. Table 7 summarizes the benchmark conditions, outcomes, and remaining boundaries.

Table 7. Controlled technical benchmark results and remaining boundaries.

Component	Test Conditions	Observed Outcome and Remaining Boundary
Audio delivery	25 HTML5 playback trials: 15 desktop Safari, 5 simulated iPhone, 5 simulated iPad	Browser playback-start/event timing: 119.67 ± 43.31 ms (81.21–252.75); 25/25 successful. Physical audible-onset latency was not measured.
Web Audio pitch reference	30 Web Audio dispatch trials: 18 desktop Safari, 6 simulated iPhone, 6 simulated iPad	Pitch-dispatch timing: 0.05 ± 0.02 ms (0.04–0.12); 30/30 successful. This is application dispatch timing, not acoustic output latency.
Local persistence	25 LocalStorage JSON serialization/setItem trials for a 50-item examination state	Execution timing: 0.03 ± 0.01 ms (0.02–0.03); 25/25 successful.
Recovery	25 same-device simulated interruption and recovery trials using LocalStorage snapshots	25/25 expected 50-item states restored; 0.04 ± 0.01 ms (0.03–0.08); 0 missing and 0 duplicated answers.
Cloud saving	20 controlled dummy Firestore writes to the live cloud endpoint	20/20 acknowledged; Firestore acknowledgement latency 754.64 ± 288.73 ms (373.08–1335.10); zero recorded write loss.
Final submission	20 controlled dummy end-to-end submissions with predefined responses, verified keys, cloud acknowledgement, and score read-back	20/20 acknowledged; 689.28 ± 243.91 ms (386.79–1090.82); 0 duplicate records; 0 scoring discrepancies.

Table 7. *Cont.*

Component	Test Conditions	Observed Outcome and Remaining Boundary
Compatibility and interaction	Desktop Safari/WebKit on macOS; previously verified Windows Chrome/Edge; 11-step workflow on each of two simulated mobile Safari profiles	22/22 simulated mobile-profile workflow checks passed. The iPhone profile used iOS 17.5; the iPad profile was a WebKit simulation on macOS (recorded user-agent: CPU OS 12_2). Neither profile used physical hardware.
Authentication and access	Login, restricted-examination access, final submission, and post-logout protected-route checks within both simulated mobile workflows	Passed on both profiles; these are functional checks and do not constitute penetration testing or an independent security audit.

3.7. Scope of the Examination Evidence

The study provides content-review and questionnaire findings, but no examination-score distribution, pass/fail rate, or empirical item-analysis results. Numerical records visible in the implemented-system screenshots were not included in the research results. The workflow in Figure 4 sets out how response-level examination data and verified keys can support subsequent item review and question-bank updating.

4. Discussion

4.1. Contribution to Digital Assessment in Music Education

The study shows how auditory delivery, reusable content, response preservation, and instructor review can be coordinated within a managed-backend educational information system. Its primary contribution is the integration of these functions for music assessment. The IOC findings support preliminary alignment of rhythm and chord items with objectives, the questionnaire describes favorable perceptions of content and use, the EFA provides preliminary evidence of a broad common questionnaire factor, and controlled benchmarks provide objective evidence for selected technical functions under declared conditions. Together, these results inform system development without establishing examination-score validity, broad physical-device performance, or learning effectiveness.

The reusable item bank is educationally useful because it separates question preparation from examination assembly. It may allow instructors to retain suitable audio stimuli, revise ambiguous alternatives, and create new forms without rebuilding each item. However, repeated use can also expose content and change item performance. A reusable bank is not automatically a calibrated bank: form comparisons require stable identifiers, versioned media and answer keys, and records of administration conditions. Unlike the adaptive feedback concept of Hwang and Chu [8], the documented seeded ordering changes presentation order rather than estimating ability and choosing items adaptively.

The local Thai context also limits what should be inferred from this form. The rhythm and chord tasks are not evidence that culturally specific Isan listening, timbre recognition, or performance practices have been validated. Seeyo et al. provide a rationale for developing such resources [9], but additional community and expert involvement would be required to define their constructs and choose suitable stimuli. The distinction between a curriculum-specific test and broader musical assessment remains important when comparing this work with Gold-MSI and PROMS [10,11].

The Troubadour evaluation combined user questionnaires with platform activity records [6]. That distinction is relevant here: the present questionnaire characterizes

reported experience, whereas interaction logs and controlled task observations would be needed to explain how students use the system. Likewise, the online MET study supplies a precedent for investigating auditory score properties directly [3], rather than inferring them from satisfaction.

4.2. Interpreting Favorable Ratings and the Stability Concern

Content alignment and perceived feedback benefits were the strongest aspects of the user evaluation. This pattern is consistent with the relevance of usefulness and information quality in information-systems research [18,19]. It is not a test of those models, because the questionnaire was neither their validated measurement instrument nor part of a structural model. The lower interface/usability composite points to a development priority that would be hidden by reporting only one favorable overall mean.

System stability warrants particular attention. Almost two-fifths of respondents selected a rating of 3 or below on this item, but the questionnaire cannot identify whether the cause was network access, a browser restriction, media loading, device behavior, or application logic. The later controlled benchmark found no failures in the selected audio-playback, same-device recovery, Firestore-write, and final-submission trials under the declared conditions, and the two simulated WebKit mobile profiles passed all 22 workflow checks. However, these revision-time tests were conducted after the questionnaire period and cannot retrospectively explain or invalidate the lower Q06 ratings. In addition, the mobile-profile checks were browser simulations on macOS rather than physical-device tests. The survey findings should, therefore, be interpreted as evidence of heterogeneous perceived stability in the field, while the benchmark provides separate evidence about selected functions under controlled conditions. Avoiding unnecessary interface demands is also consistent with cognitive-load considerations, although cognitive load was not measured here [37].

The perceived usefulness of immediate scores should likewise be interpreted cautiously. The system permits instructor-controlled grading and feedback release; the item wording does not demonstrate that every score was released immediately. Research on formative feedback emphasizes its content and use, not timing alone [15–17]. A later study could examine whether item-specific explanatory feedback helps students identify the musical feature they misheard, rather than merely learn which option was correct.

4.3. Interpretation of the Initial Evaluation Evidence

Expert content review and questionnaire coefficients address different aspects of evaluation. IOC summarizes judgments about correspondence between items and objectives, whereas Cronbach's alpha summarizes covariance among questionnaire scores. A high questionnaire alpha cannot be transferred to the 50-item music examination, and favorable user evaluations do not establish the validity or reliability of examination scores [23,24,29]. The findings should, therefore, be interpreted as initial content and user evaluation.

The four three-item conceptual clusters have high alpha values, but the post hoc EFA did not support four distinct latent dimensions. Instead, parallel analysis supported one dominant factor explaining 62.65% of the standardized item variance. This pattern is consistent with substantial shared covariance across the positively worded evaluation items. The four original group labels are therefore retained only for descriptive organization, not as validated subscales. Because the EFA is exploratory, based on a modest single-site sample, and not followed by confirmatory factor analysis in an independent sample, future measurement work should evaluate the structure again and consider model-based reliability estimates such as omega [38,39]. This distinction separates reliability reporting from instrument development and validation [25,30].

For examination scores, the future evidence requirements are different: verified keys, an examinee-by-item response matrix, item difficulty, discrimination, distractor functioning, and reliability under specified testing conditions. A high-*p* item may be appropriate for checking essential mastery, even if it is easy. A negative discrimination coefficient requires investigation of the key, wording, media, or construct before a retention decision can be made. The pitch-reference piano and replay rules must also be controlled, because assistance can change score meaning. Such work would extend this initial study beyond content review and user perception.

4.4. Integrity, Privacy, and Implementation Priorities

Browser-visibility monitoring should be positioned as an administrative signal. Losing focus may result from a notification, accessibility tool, operating-system event, or legitimate technical action. Automatic penalties based on this signal can therefore create fairness concerns. This interpretation aligns with treating assessment integrity as a broader information and procedural problem [13,20]. Instructors must establish documented exception handling and a review process before any consequence is attached to a count.

Persistence and security require interpretation within the limits of the controlled tests. Local snapshots reduce dependence on a single network transaction, and the revision-time benchmark achieved 25/25 same-device recoveries and 20/20 acknowledged Firestore writes. Final-submission testing also achieved 20/20 acknowledged submissions with no duplicate records or scoring discrepancies. In the simulated WebKit mobile-profile workflows, login, restricted-examination access, final submission, and post-logout protected-route behavior passed on both profiles. These results strengthen the implementation evidence, but they do not constitute an independent source-code review, penetration test, or comprehensive security audit. Local persistence on shared devices, broader network conditions, and physical-device variability therefore remain relevant deployment considerations.

4.5. Limitations and Subsequent Research

Several limitations restrict interpretation. The survey is a single institution, non-probability sample without a recorded invitation denominator. Its 100 undergraduate and one postgraduate records, including 14 respondents reporting no prior ear-training study, do not establish correspondence with the originally proposed course cohort. Eligibility and individual course membership cannot be reconstructed from these self-reports alone. The undergraduate sensitivity analysis addresses one study-level discrepancy but does not resolve that recruitment limitation. Furthermore, the survey records are not linked to application versions, so the effects of subsequent piano and question-bank changes cannot be separated. The EFA is post hoc and exploratory, and the one-factor solution requires confirmation in a larger independent sample.

The expert-review findings were available only as an aggregate summary. The original item-by-reviewer ratings and written comments were unavailable; therefore, reviewer agreement and individual item-level decisions could not be reanalyzed. The questionnaire measures perceptions rather than observed task performance or learning gains. Although EFA was added, it provides only preliminary exploratory structure evidence. There is no linked examination-response dataset, test–retest design, or comparison group. The revision-time technical benchmark provides objective evidence for selected functions, but it was conducted after the questionnaire period on declared testbeds and therefore cannot establish the exact stimuli, settings, network conditions, or software configuration experienced by every respondent. The mobile Safari-profile checks were simulated through WebKit on macOS rather than performed on physical iPhone/iPad hardware. Physical acoustic audible-onset latency and comprehensive security penetration testing were not evaluated.

Subsequent evaluation should preserve a documented software release, examination form, audio specifications, and consented course roster. Independent content review and verified answer keys should precede item analysis. Table 7 summarizes the controlled technical tests completed during revision and their remaining boundaries. Future work should extend these tests to physical iPhone/iPad hardware, native iOS/iPadOS Safari, additional network conditions, and independent security review before broader deployment claims are made, and should collect linked examination-response data before comparing assisted and unassisted assessment, equivalent forms, or learning outcomes.

5. Conclusions

This study designed, developed, and initially evaluated a web-based music ear-training assessment system integrating managed authentication, examination records, auditory media, reusable questions, browser-generated pitch references, local and cloud response saving, and instructor-controlled grading. Expert review of 50 items yielded a mean IOC of 0.953. The 101 consenting questionnaire respondents reported a broad evaluation mean of 4.30 out of 5 (SD = 0.60; Cronbach's $\alpha = 0.943$). Post hoc EFA supported one dominant questionnaire factor explaining 62.65% of standardized item variance, so the four prespecified item groups are interpreted as conceptual clusters rather than validated dimensions. Controlled revision-time testing provided selected technical evidence, including 25/25 successful audio-playback requests, 25/25 same-device recoveries, 20/20 acknowledged Firestore writes, and 20/20 successful final submissions with no duplicate records or scoring discrepancies; complementary simulated WebKit mobile-profile testing passed 22/22 workflow checks across iPhone and iPad profiles. These findings support preliminary content alignment, favorable user perceptions, exploratory questionnaire-structure evidence, and selected technical performance under the tested conditions. They do not establish examination-score validity, item discrimination, performance on physical mobile hardware, or learning effects.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/informatics13110182/s1. Supplement S1: Questionnaire wording, statistical methods, aggregate results, EFA details, and technical-benchmark notes; survey_aggregate_statistics.json: item distributions, means, and covariance matrix; reproduce_statistics.py: code to reconstruct reported descriptive and internal-consistency statistics; efa_aggregate_results.json: reported EFA outputs and reference values; reproduce_efa.py: code to reconstruct the primary Pearson-correlation EFA from the aggregate covariance matrix; technical_test_raw_log.xlsx and technical_test_raw_log.csv: trial-level controlled technical benchmark records.

Author Contributions: Conceptualization, W.S.; methodology, W.S. and S.C.; software, J.S. and W.H.; validation, S.C. and W.H.; formal analysis, W.S.; investigation, W.S.; resources, W.S., S.C. and J.S.; data curation, W.S. and J.S.; writing—original draft preparation, W.S.; writing—review and editing, W.S., S.C., J.S. and W.H.; visualization, W.S. and J.S.; supervision, S.C.; project administration, W.S.; funding acquisition, W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research project was financially supported by Mahasarakham University under Research Funding Contract No. 6901036/2569.

Institutional Review Board Statement: Research procedures followed the ethical principles of the Declaration of Helsinki. The study was approved under the exempt-review procedure by the Mahasarakham University Ethics Committee for Research Involving Human Subjects (English Approval No. 035-031/2026; Thai Approval No. 035-031/2569; 19 January 2026).

Informed Consent Statement: Electronic informed consent for research participation was obtained from all 101 respondents included in the analysis. Participation was voluntary. One submission indicating non-consent was excluded before analysis.

Data Availability Statement: Aggregate questionnaire statistics, analysis code, and trial-level technical benchmark logs are provided with the Supplementary Materials. Respondent-level records are not publicly distributed because the source export contains direct identifiers and the available consent wording does not establish permission for unrestricted public release. Requests for access can be directed to the corresponding author and are subject to institutional and ethics requirements. Original item-by-reviewer ratings, linked examinee-by-item examination-response data, and application source code are not included in the present supplementary package.

Acknowledgments: During manuscript preparation, ChatGPT-6, through Codex (OpenAI; accessed in 2026), assisted with drafting, literature searching, analysis-code preparation, and formatting. OpenAI image-generation tools assisted with publication-oriented preparation of Figures 2 and 3 as described in Section 2.8. The authors reviewed and edited all outputs and are responsible for the accuracy and integrity of the analyses, references, figures, and final publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

API, application programming interface; CI, confidence interval; EFA, exploratory factor analysis; IOC, item–objective congruence; KMO, Kaiser–Meyer–Olkin; KR-20, Kuder–Richardson Formula 20; MCQ, multiple-choice question; MET, Musical Ear Test; META, Musical Ear Training Assessment; SD, standard deviation; SPA, single-page application; ULS, unweighted least squares.

References

1. Wolf, A.; Kopiez, R. Development and Validation of the Musical Ear Training Assessment (META). *J. Res. Music Educ.* **2018**, *66*, 53–70. [CrossRef]
2. Swaminathan, S.; Kragness, H.E.; Schellenberg, E.G. The Musical Ear Test: Norms and Correlates from a Large Sample of Canadian Undergraduates. *Behav. Res. Methods* **2021**, *53*, 2007–2024. [CrossRef] [PubMed]
3. Correia, A.I.; Vincenzi, M.; Vanzella, P.; Pinheiro, A.P.; Lima, C.F.; Schellenberg, E.G. Can Musical Ability Be Tested Online? *Behav. Res. Methods* **2022**, *54*, 955–969. [CrossRef] [PubMed]
4. Chen, B.; Li, C. A Quasi-Experimental Study on the Application of MuseScore Software to Improve Sight-Singing and Ear-Training Abilities in Music Education at Chuzhou University. *Int. J. Sociol. Anthropol. Sci. Rev.* **2025**, *5*, 609–626. [CrossRef]
5. Buonviri, N.O.; Paney, A.S. Technology Use in High School Aural Skills Instruction. *Int. J. Music Educ.* **2020**, *38*, 431–440. [CrossRef]
6. Pesek, M.; Klavž, F.; Šavli, P.; Marolt, M. Online and In-Class Evaluation of a Music Theory E-Learning Platform. *Appl. Sci.* **2022**, *12*, 7296. [CrossRef]
7. Chen, Z.-L.; Lin, H. AI and Digital Technologies in Sight-Singing and Aural Training: A Systematic Review of Innovations in Higher Music Education. *Eur. J. Educ.* **2026**, *61*, e70502. [CrossRef]
8. Hwang, Y.T.; Chu, C.N. The Design of Adaptive Error Feedback Music Ear-Training System with Image Cues. In *Human-Computer Interaction. Users and Applications*; Jacko, J.A., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 35–38. [CrossRef]
9. Seeyo, W.; Karin, K.; Chuangprakhon, S. Instructional Management Strategies for Isan Folk Music Ear Training in Thai Higher Education. *Afr. Educ. Res. J.* **2026**, *14*, 408–417. [CrossRef]
10. Müllensiefen, D.; Gingras, B.; Musil, J.; Stewart, L. The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE* **2014**, *9*, e89642. [CrossRef] [PubMed]
11. Kunert, R.; Willems, R.M.; Hagoort, P. An Independent Psychometric Evaluation of the PROMS Measure of Music Perception Skills. *PLoS ONE* **2016**, *11*, e0159103. [CrossRef] [PubMed]
12. Gikandi, J.W.; Morrow, D.; Davis, N.E. Online Formative Assessment in Higher Education: A Review of the Literature. *Comput. Educ.* **2011**, *57*, 2333–2351. [CrossRef]
13. Butler-Henderson, K.; Crawford, J. A Systematic Review of Online Examinations: A Pedagogical Innovation for Scalable Authentication and Integrity. *Comput. Educ.* **2020**, *159*, 104024. [CrossRef] [PubMed]
14. Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assess. Educ.* **1998**, *5*, 7–74. [CrossRef]
15. Nicol, D.J.; Macfarlane-Dick, D. Formative Assessment and Self-regulated Learning: A Model and Seven Principles of Good Feedback Practice. *Stud. High. Educ.* **2006**, *31*, 199–218. [CrossRef]
16. Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112. [CrossRef]
17. Shute, V.J. Focus on Formative Feedback. *Rev. Educ. Res.* **2008**, *78*, 153–189. [CrossRef]

18. Davis, F.D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q*. **1989**, *13*, 319–340. [CrossRef] [PubMed]
19. DeLone, W.H.; McLean, E.R. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. *J. Manag. Inf. Syst*. **2003**, *19*, 9–30. [CrossRef]
20. Kaisara, G.; Bwalya, K.J. Strategies for Enhancing Assessment Information Integrity in Mobile Learning. *Informatics* **2023**, *10*, 29. [CrossRef]
21. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design Science in Information Systems Research. *MIS Q*. **2004**, *28*, 75–105. [CrossRef]
22. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A Design Science Research Methodology for Information Systems Research. *J. Manag. Inf. Syst*. **2007**, *24*, 45–77. [CrossRef]
23. Kane, M.T. Validating the Interpretations and Uses of Test Scores. *J. Educ. Meas*. **2013**, *50*, 1–73. [CrossRef]
24. Messick, S. Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *Am. Psychol*. **1995**, *50*, 741–749. [CrossRef]
25. Boateng, G.O.; Neilands, T.B.; Frongillo, E.A.; Melgar-Quiñonez, H.R.; Young, S.L. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Front. Public Health* **2018**, *6*, 149. [CrossRef] [PubMed]
26. Eysenbach, G. Improving the Quality of Web Surveys: The Checklist for Reporting Results of Internet E-Surveys (CHERRIES). *J. Med. Internet Res*. **2004**, *6*, e34. [CrossRef] [PubMed]
27. Cho, E.; Kim, S. Cronbach's Coefficient Alpha: Well Known but Poorly Understood. *Organ. Res. Methods* **2015**, *18*, 207–230. [CrossRef]
28. Tavakol, M.; Dennick, R. Making Sense of Cronbach's Alpha. *Int. J. Med. Educ*. **2011**, *2*, 53–55. [CrossRef] [PubMed]
29. Sijtsma, K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika* **2009**, *74*, 107–120. [CrossRef] [PubMed]
30. Taber, K.S. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Res. Sci. Educ*. **2018**, *48*, 1273–1296. [CrossRef]
31. Horn, J.L. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika* **1965**, *30*, 179–185. [CrossRef] [PubMed]
32. Norman, G. Likert Scales, Levels of Measurement and the "Laws" of Statistics. *Adv. Health Sci. Educ*. **2010**, *15*, 625–632. [CrossRef] [PubMed]
33. Sullivan, G.M.; Artino, A.R. Analyzing and Interpreting Data from Likert-Type Scales. *J. Grad. Med. Educ*. **2013**, *5*, 541–542. [CrossRef] [PubMed]
34. Lim, B.C.; Kueh, Y.C.; Arifin, W.N.; Ng, K.H. Psychometric Properties of the Heart Disease Knowledge Scale: Evidence from Item and Confirmatory Factor Analyses. *Malays. J. Med. Sci*. **2016**, *23*, 33–45. [CrossRef] [PubMed]
35. Kumar, D.; Jaipurkar, R.; Shekhar, A.; Sikri, G.; Srinivas, V. Item Analysis of Multiple Choice Questions: A Quality Assurance Test for an Assessment Tool. *Med. J. Armed Forces India* **2021**, *77*, S85–S89. [CrossRef] [PubMed]
36. Kuder, G.F.; Richardson, M.W. The Theory of the Estimation of Test Reliability. *Psychometrika* **1937**, *2*, 151–160. [CrossRef]
37. Sweller, J.; van Merriënboer, J.J.G.; Paas, F. Cognitive Architecture and Instructional Design: 20 Years Later. *Educ. Psychol. Rev*. **2019**, *31*, 261–292. [CrossRef]
38. Dunn, T.J.; Baguley, T.; Brunsden, V. From Alpha to Omega: A Practical Solution to the Pervasive Problem of Internal Consistency Estimation. *Br. J. Psychol*. **2014**, *105*, 399–412. [CrossRef] [PubMed]
39. McNeish, D. Thanks Coefficient Alpha, We'll Take It from Here. *Psychol. Methods* **2018**, *23*, 412–433. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.