

Interactive Podcast-Based Learning Model to Enhance Speaking Skills of University Students

Rio Kurniawan^{1*}, Azhar Aziz Lubis², Panut Setiono³, Didi Yulistio⁴

¹⁻⁴ Universitas Bengkulu, Indonesia

Abstract

University speaking courses require learning designs that provide repeated oral production, self-monitoring, and usable feedback. This study developed and preliminarily evaluated an interactive podcast-based learning model for third-semester students in the Indonesian Language Education Study Program at Universitas Bengkulu. A development-oriented mixed-methods design (QUAL → quan) was embedded in the ADDIE stages. Formal needs assessment involved three classroom observations, interviews with 15 students, two lecturers, and one educational technology specialist, and document analysis. One intact class of 30 students participated in five podcast cycles across ten meetings. Expert appraisal increased from 85.0% to 88.5% of the maximum possible score after revision, while students reported high practicality ($M = 4.18$, $SD = 0.46$); the two lecturers provided a combined descriptive mean of 4.38. The three-component speaking-performance score increased from 63.33 ($SD = 8.10$) to 81.00 ($SD = 6.70$), $t(29) = 12.12$, $p < .001$, $d_z = 2.21$. Observable speaking assurance, reported separately from linguistic performance, also increased. Because the study used one intact class without a comparison group and did not systematically code feedback quality or individual change trajectories, the findings indicate feasibility and preliminary improvement rather than causal effectiveness. The study contributes a transferable practice–replay–feedback–revision sequence integrating culturally responsive task criteria and privacy-conscious implementation supports.

ARTICLE HISTORY

Received : 30 March 2026

Revised : 4 July 2026

Accepted : 12 July 2026

KEYWORDS

Feedback literacy; Instructional design; Interactive podcast; Learner-generated audio; Speaking skills

PUBLISHER'S NOTE

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license.



CORRESPONDING AUTHOR

*Rio Kurniawan, Universitas Bengkulu, Indonesia. Email: kurniawanrio@unib.ac.id

Introduction

Speaking is a core professional capability for prospective language teachers because classroom work depends on clear explanation, questioning, storytelling, feedback delivery, and audience-sensitive interaction. Contemporary research on technology-enhanced language learning shows that speaking development is strengthened when learners receive frequent opportunities to produce language, revisit their performance, and interact around explicit criteria rather than merely consume digital content (Huang & Li, 2024; Perry, 2024). Learner-generated podcasting is therefore potentially valuable, but its pedagogical effect depends on the design of the production and feedback process, not on the novelty of audio technology itself (Wakefield et al., 2023).

The study was initiated through informal scoping in May 2025 and a formal needs assessment in July 2025. The May classroom visits were used only to identify a potential instructional problem and were not treated as research evidence in the analysis. The formal dataset comprised three structured observations of the subsequently selected 30-student class, interviews with 15 students, interviews with two Speaking Skills lecturers and one educational technology specialist, and analysis of the course syllabus and Semester Learning Plan. The archived observation format documented the form of speaking opportunities and feedback but did not quantify speaking-turn duration or feedback frequency; consequently, the baseline findings are reported as cross-source qualitative requirements rather than prevalence estimates.

Across the formal sources, the central problem was not the complete absence of oral tasks. Students presented in class, but individual speaking turns were concentrated in a small number of performances, feedback was often general, and students had limited opportunities to listen to themselves and revise before final assessment. Interview and observation records also indicated uneven device quality and concern about public posting. The design challenge was therefore to create a low-threshold, privacy-conscious sequence that distributed practice across all students and connected self-review, peer dialogue, lecturer feedback, and revision.

Four complementary design principles guided the instructional sequence. First, repeated individual production increases opportunities for every learner to practise rather than allowing a few confident speakers to dominate classroom time. Second, replay supports self-monitoring by making pauses, articulation, and intonation available for inspection. Third, feedback literacy positions students as active users of criteria who learn to give, interpret, and act on comments; peer-feedback interventions are more productive when learners receive scaffolding and when feedback uptake is built into the task (Carless & Boud, 2018; Little et al., 2024; Man et al., 2024). Mobile-assisted speaking research similarly shows that the quality and timeliness of peer feedback depend on motivation, guidance, and the design of the interaction environment (Xu & Peng, 2022). Fourth, culturally responsive digital pedagogy treats learners' cultural resources as assets while requiring choice, respectful representation, and alignment with disciplinary objectives (Gay, 2018; McDaniel, 2024; Wesley-Nero & Donley, 2024).

Recent studies demonstrate several relevant affordances but leave an integration gap. Learner-generated podcasts can function as authentic assessment and can create a less intimidating speaking environment (Perry, 2024; Wakefield et al., 2023). Digital storytelling can support speaking proficiency and willingness to communicate when learners create meaningful narratives (Huang, 2023), while technology-enhanced environments may reduce communication barriers when pedagogy, privacy, and learner diversity are addressed (Huang & Li, 2024). However, these strands are often reported separately: podcast projects may prioritize the final product, peer-feedback studies may not require replay and revision, and culturally situated tasks may not be linked to a criterion-referenced oral-performance cycle. The present sequence combines these components as a testable design proposition.

Table 1. Comparative synthesis of related instructional approaches

Approach	Typical emphasis	Common limitation for the present problem	Distinctive element in the proposed sequence
Teacher-produced podcast	Listening input and flexible access	Learners may remain consumers; no required speaking revision cycle	Every student produces and revises individual audio
Learner-generated podcast	Planning, recording, and sharing a product	Feedback and uptake may be optional or product-focused	Criterion-referenced peer and lecturer feedback is followed by revision
Mobile-assisted peer feedback	Digital exchange of comments on oral work	Quality depends on training, motivation, and task structure	Three-prompt feedback scaffold, moderation, and a lecturer-response rule

Approach	Typical emphasis	Common limitation for the present problem	Distinctive element in the proposed sequence
Digital storytelling	Meaningful narrative production and identity work	May be multimodal and not designed around specific oral-performance criteria	Local narratives are used within speaking objectives and ethical representation criteria
Proposed instructional sequence	Practice → replay → feedback → revision	Requires moderation and technical support	Platform-neutral workflow integrating self-monitoring, feedback literacy, cultural responsiveness, and restricted sharing

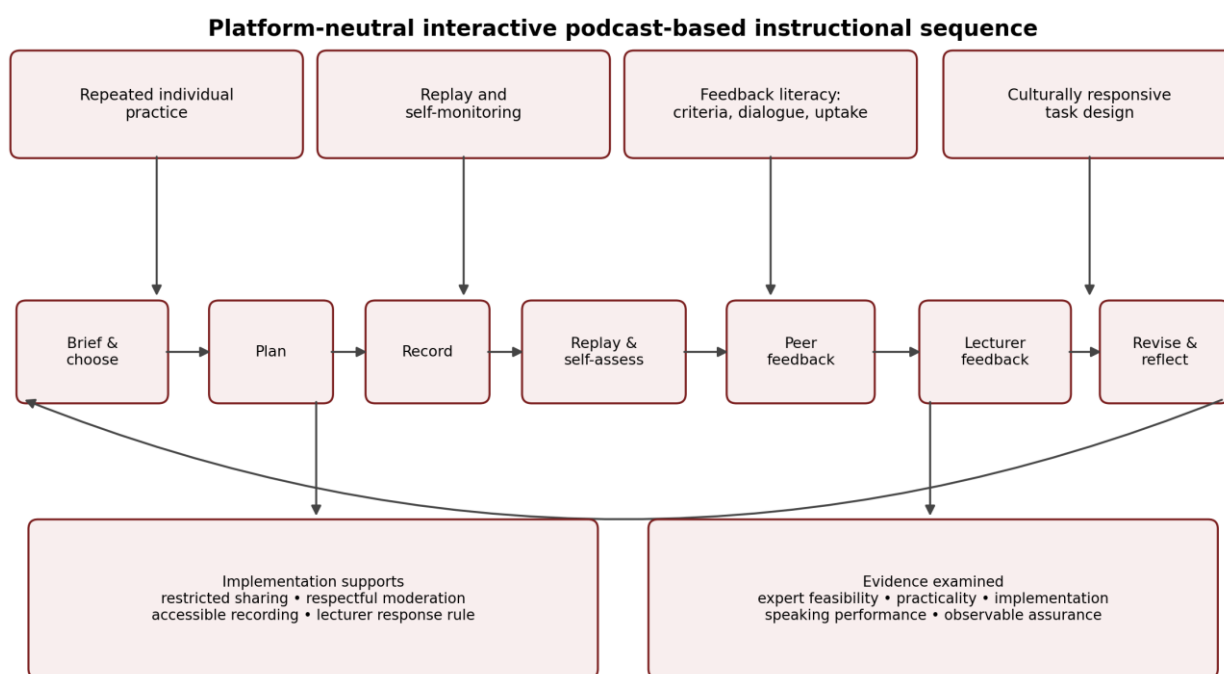


Figure 1. Conceptual architecture and boundaries of the interactive podcast-based instructional sequence

The product is therefore defined as an instructional sequence, not a podcast application or a stand-alone medium. Its boundaries include task syntax, student and lecturer guides, speaking criteria, feedback prompts, moderation and privacy rules, and revision procedures. Specific applications are replaceable. In this implementation, students used available audio recorders, restricted cloud links, and a moderated messaging group; public podcast distribution and Spotify were not components of the tested sequence.

Cultural responsiveness was operationalized through five design criteria: (a) relevance to the Speaking Skills learning objective; (b) use of local narratives as resources rather than stereotypes; (c) student agency to select or propose an appropriate topic; (d) lecturer review of source accuracy and respectful representation; and (e) reflection on audience, language choice, and ethical retelling. The trial documentation confirms lecturer consultation and alignment with course materials, but it does not provide a systematic audit of student topic substitution or cultural-accuracy decisions.

These aspects are therefore specified as requirements of the final sequence and priorities for subsequent implementation research.

The study addressed five questions: (1) What needs and design requirements emerged from the formal needs assessment? (2) How was the instructional sequence developed and revised through ADDIE? (3) What did expert appraisal, user practicality responses, and implementation records indicate about feasibility? (4) What preliminary change occurred in the three-component speaking-performance score? and (5) What preliminary change occurred in observable speaking assurance, treated separately from linguistic performance?

Method

Design and Mixed-Methods Integration

This research-and-development study used ADDIE (Analysis, Design, Development, Implementation, and Evaluation) as the product-development framework (Branch, 2009). The methodological sequence is represented as QUAL → quan: qualitative needs evidence had priority in Analysis and Design; descriptive quantitative evidence was embedded in Development and Implementation; and pretest–posttest statistics were used in Evaluation. Integration occurred through connecting, building, and joint interpretation. Qualitative themes were connected to design requirements, appraisal and pilot evidence built revisions, and implementation evidence was used to delimit the interpretation of quantitative change (Creswell & Plano Clark, 2018). The inference generated by integration was not that the sequence caused improvement, but that the design was feasible in one cohort and sufficiently promising for controlled evaluation.

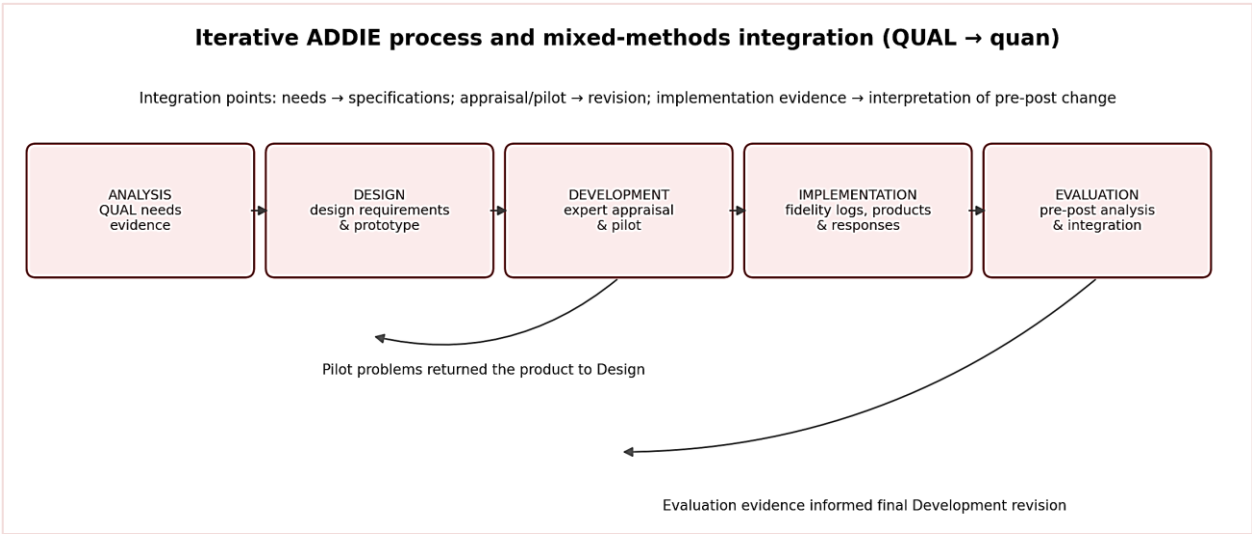


Figure 2. Iterative ADDIE process, decision points, and mixed-methods integration

Figure 2 illustrates the iterative application of the ADDIE framework within a qualitatively driven mixed-methods design. The Analysis, Design, Development, Implementation, and Evaluation stages are systematically connected through multiple integration and feedback points. Evidence obtained from expert appraisal and pilot testing informs revisions to the initial design, while implementation and pre–post evaluation data guide interpretation and final product refinement. This cyclical process ensures that development decisions remain evidence-based, responsive, and methodologically integrated.

Table 2. Alignment of research questions, evidence, analysis, and results

Research question	Primary evidence	Analysis	Result location
RQ1: Needs and requirements	Three observations; 15 student interviews; 2 lecturer interviews; 1 technology-specialist interview; syllabus/RPS	Cross-source coding and design-requirement matrix	Needs assessment and design requirements
RQ2: Development and revision	Design records, expert comments, pilot logs	Audit trail with decision triggers and iteration records	Development iterations and expert appraisal
RQ3: Feasibility	Expert appraisal, practicality questionnaire, implementation logs	Percentage of maximum descriptions; means/SD for students; descriptive lecturer summary; fidelity evidence table	Feasibility, practicality, and fidelity
RQ4: Speaking-performance change	Coded pretest/posttest audio ratings	Paired t test, 95% CI, dz; exploratory dimension tests with Holm threshold	Preliminary learning change
RQ5: Observable assurance change	Separate rubric indicator	Paired t test and cautious descriptive interpretation	Preliminary learning change

Development Procedure and Decision Rules

The procedure was iterative rather than strictly linear. Needs evidence generated the first specification. During Development, validator and pilot evidence returned the product to Design: privacy concerns led to restricted sharing, ambiguous rubric descriptors were rewritten, and the peer-feedback template was reduced to three prompts. During Evaluation, inconsistencies in score conversion and interpretation returned the product to Development for final clarification. A revision was triggered when at least one of the following occurred: two or more validators identified the same conceptual problem; a pilot participant could not complete a required step without assistance; a privacy or respectful-interaction risk was identified; or implementation logs showed that a procedure was not feasible with available devices or connectivity.

Table 3. Development audit trail and documented iteration

Stage and evidence	Decision trigger	Revision/iteration	Output
Analysis (July 2025): observations, interviews, documents	Repeated evidence of limited individual practice, general feedback, device inequality, and privacy concern	Specified individual recording, replay, criterion-based feedback, restricted sharing, and technical alternatives	Needs-to-design matrix

Stage and evidence	Decision trigger	Revision/iteration	Output
Design (August 2025): initial syntax, guides, tasks, rubric	Mismatch between public podcast conventions and participant privacy concerns	Returned to Design; public publication removed and platform-neutral functions specified	Prototype sequence and guides
Development (September 2025): three-expert appraisal and 10-student pilot	Repeated comments on instruction clarity, privacy, rubric ambiguity, file naming, and low bandwidth	Returned to Design; instructions shortened, privacy rules added, descriptors sharpened, feedback prompts reduced	Revised implementation package
Implementation (October–November 2025): ten meetings	Internet disruption and inadequate microphones for some students	Operational adaptation without changing core pedagogy: compressed audio, shared microphones, campus facilities	Implementation logs and completed recordings
Evaluation (December 2025): appraisal, practicality, pretest/posttest evidence	Need to separate linguistic performance from observable assurance and avoid causal claims	Returned to Development; outcome reporting, limitations, and final specification revised	Final instructional sequence and manuscript

Participants, Sampling, and Research Ethics

The population and accessible population each comprised 60 third-semester students registered in two Speaking Skills classes in the Indonesian Language Education Study Program at FKIP Universitas Bengkulu during 2025/2026. One intact class of 30 students became the intervention cohort because its schedule matched the development period and its lecturer agreed to integrate the sequence into regular instruction. All eligible students in that class participated. This convenience-based intact-class selection creates selection bias and limits transferability. The second class could theoretically have supported a quasi-experimental comparison, but different scheduling and instructional timing prevented synchronized baseline and intervention procedures; the project was consequently treated as a developmental feasibility trial rather than a comparative evaluation.

The 15 interviewees were purposively selected from the intervention class to represent lower, middle, and higher preliminary speaking levels. Because the same students contributed to needs assessment and implementation, expectations and social-desirability effects are possible. Interviews continued to 15 participants, but a formal saturation grid and negative-case log were not retained; therefore, the revision makes no strong saturation claim. Dissenting or technically difficult experiences are reported where available, including privacy concerns, unstable internet, and inadequate microphone quality. Students received an explanation of the activities and consented to voice recording and restricted class use. Recordings were shared only through a closed class space, and students were instructed not to redistribute peers' files.

Instructional Sequence, Dosage, and Platform-Neutral Specification

The sequence consisted of five cycles implemented across ten regular class meetings. Each cycle included briefing, planning, recording, replay/self-review, peer feedback, lecturer feedback, and revision/reflection. Audio products lasted 3–4 minutes, except the 1–2-minute orientation recording. The duration was selected to make repeated individual production feasible within one course while

keeping files small enough for low-bandwidth submission. Exact out-of-class preparation time and lecturer workload were not logged, so workload feasibility is interpreted from completion records and practicality responses rather than time-on-task measurement.

Table 4. Core functions, implementation example, and cycle objectives

Core function / cycle	Platform-neutral requirement	Implementation example in the trial	Learning objective / evidence
Restricted capture and sharing	Any audio recorder and access-controlled sharing space	Student devices/Audacity; restricted cloud links	Protect privacy while enabling replay
Cycle 1: orientation and voice awareness	1–2-minute diagnostic recording and rubric orientation	Trial recording during Meetings 1–2	Recognize voice quality and recording procedure
Cycle 2: pronunciation-focused monologue	3–4-minute individual monologue, replay, criteria-based feedback	Meetings 3–4	Increase attention to articulation and fluency
Cycle 3: culturally grounded narrative retelling	Narrative selected under relevance, choice, accuracy, and respect criteria	Meetings 5–6	Develop expressive retelling and audience awareness
Cycle 4: teaching simulation	Explain a narrative text to an imagined secondary audience	Meetings 7–8	Practice teacher-oriented oral explanation
Cycle 5: reflective final performance	Final recording, post-task reflection, and posttest	Meetings 9–10	Integrate performance criteria and reflection
Feedback and revision	Structured peer prompt, lecturer rubric feedback, opportunity to revise	Moderated class messaging; lecturer response target within 48 hours	Develop feedback literacy and performance uptake

The study did not provide automated speech scoring, AI-generated feedback, or generative-AI tools. Students could plan scripts and use basic recording functions, but possible use of external editing, pronunciation, or AI assistance outside the formal workflow was not systematically audited. This boundary is acknowledged because uncontrolled external assistance could influence performance.

Instruments and Measurement

The needs-assessment instruments included a 12-question semi-structured interview guide, a 16-indicator observation sheet, and a document-analysis guide. Domains covered speaking opportunity, feedback experience, hesitation, technology access, attitudes toward recording, and cultural-task relevance. The original instrument package was not included in the manuscript files supplied for revision; accordingly, the article reports instrument domains and analytic use rather than reproducing unverified item wording.

The 18-item practicality questionnaire used a five-point response scale. Four item clusters—guide clarity, technological ease, feedback usefulness, and learning/cultural relevance—were used to organize descriptive reporting only; the study does not claim that these clusters constitute

validated latent subscales. The total student scale had Cronbach's $\alpha = .89$. Item-level distributions and cluster-specific reliability were not available, so the result is interpreted as a descriptive usability indicator subject to acquiescence and course-context social-desirability bias. Lecturer responses are presented descriptively because only two lecturers participated.

Speaking performance was assessed with parallel 3–4-minute teaching-simulation tasks using the same genre, audience, duration, and rating criteria but different narrative topics. The exact prompt wording and a formal linguistic-difficulty equivalence analysis were not available; expert review provided only a limited comparability check. The primary measured construct was a three-component speaking-performance score comprising fluency, pronunciation, and intonation. Content organization and audience awareness were instructional aims but were not separately scored, which narrows the construct and is acknowledged as a measurement limitation. The former 'self-confidence' indicator was renamed observable speaking assurance because it was inferred from performance behavior rather than a validated affective scale.

Each dimension was scored from 1 to 5 and converted to a 0–100 scale. Two trained speaking-course lecturers independently rated coded files without pretest/posttest labels. The raters were not fully independent of the instructional context, creating possible expectancy bias. The reported ICC(2,2) = .87, 95% CI [.78, .93], is interpreted as a two-way random-effects, absolute-agreement, average-measures coefficient (Koo & Li, 2016). Dimension-level ICC values were not retained, so reliability is established only for the reported composite and should be strengthened through external blinded raters in future work.

Expert Appraisal, Qualitative Analysis, and Statistics

Three experts appraised the prototype: two language-education lecturers with speaking-assessment expertise and one educational-technology lecturer with digital-learning expertise. They were selected for domain relevance; the available records do not document their institutional independence from the research team. The appraisal used 26 items on a 1–5 scale and is reported as the percentage of the maximum obtainable score. No external empirical cut-off was applied, and the percentages are not described as proof of content validity. Item-level agreement, validator-specific totals, and uncertainty estimates were not available; therefore, the results are framed as a limited expert feasibility appraisal. The available record indicates a second appraisal after revision but does not document blinding to earlier scores, so confirmation effects cannot be excluded (Polit et al., 2007; Yusoff, 2019).

Qualitative analysis followed coding, categorization, and cross-source comparison (Miles et al., 2014). Meaningful interview statements, observed events, and relevant document segments were coded, grouped into need categories, and translated into design requirements. Observation and document evidence were used to corroborate or qualify interview claims rather than assigned numerical weights. A peer debriefer reviewed the preliminary categories. Because verbatim transcripts, a formal codebook, saturation table, and negative-case register were not part of the supplied archive, the revision reports evidence-source matrices and limits claims to development-oriented requirements rather than rich interpretive themes.

The primary quantitative analysis compared the three-component speaking-performance score using a paired-samples t test, 95% confidence interval, and d_z . Fluency, pronunciation, intonation, and observable assurance were treated as exploratory analyses; a Holm-adjusted familywise threshold was applied, and all reported p values below .001 remained below the adjusted criteria. The original analysis reported Shapiro–Wilk $W = .96$, $p = .31$ for the change score. Given $n =$

30 and the very large standardized changes, effect sizes are treated as provisional. Aggregate outputs did not permit individual trajectory plots, robust sensitivity analysis, regression-to-the-mean checks, or a definitive ceiling-effect assessment (Lakens, 2013).

Results and Discussion

Needs Assessment and Design Requirements

The formal needs assessment generated five cross-source requirements. The strongest convergence concerned limited opportunities for every student to speak individually and the need for specific, usable feedback. Students also valued private rehearsal and replay before public performance. Technology access was available but device quality was uneven. Local narratives were considered useful when they were connected to course objectives and represented respectfully. Because the retained records did not include verified verbatim quotations or coded prevalence counts, Table 5 reports source convergence and observable examples rather than invented participant quotations.

Table 5. Cross-source evidence and resulting design requirements

Need / barrier	Evidence convergence	Observable or documentary example	Design requirement
Limited individual practice	Observations + student/lecturer interviews	Conventional presentations concentrated oral production in relatively few classroom turns	Five cycles requiring an individual recording from every student
General or delayed feedback	Student and lecturer interviews + observation sheet	Comments were often broad rather than tied to pronunciation, fluency, or intonation criteria	Three-prompt peer feedback plus lecturer rubric response
Fear of direct judgment	Student interviews + hesitant turns in observations	Students requested an opportunity to rehearse and listen before presenting	Restricted sharing, replay, and re-recording
Unequal device quality	Technology-access interview questions + implementation logs	Some devices had inadequate microphone quality; internet disruption occurred	Shared microphones, campus facilities, compressed/low-bandwidth submission
Need for meaningful content	RPS/syllabus + lecturer interviews	Narrative and teaching-simulation tasks aligned with Indonesian-language teaching objectives	Culturally responsive topic criteria and student topic agency

Development Iterations and Expert Appraisal

The first appraisal yielded 85.0% of the maximum possible score and identified four recurring revision areas: alignment between tasks and outcomes, usability and privacy of the workflow, clarity of instructions, and precision of assessment descriptors. After revision, the overall value was 88.5%. The improvement is descriptive and should not be interpreted as robust content validity because only three experts participated and validator-specific variation was unavailable.

Table 6. Expert appraisal by aspect and documented revision

Aspect (items)	Initial % of maximum	Revision	Post-revision % of maximum
Course-content alignment (8)	86.7	Clarified links between each podcast task and course outcomes	90.0
Media/workflow usability (7)	83.8	Simplified procedures; added restricted sharing and privacy safeguards	88.6

Aspect (items)	Initial % of maximum	Revision	Post-revision % of maximum
Language clarity (6)	86.1	Removed ambiguous wording and added procedural examples	91.7
Assessment alignment (5)	83.5	Separated pronunciation and intonation descriptors; clarified score conversion	87.5
Overall (26)	85.0	Minor but substantive workflow and assessment revision	88.5

The 10-student pilot was conducted with students from the same study program who were not members of the 30-student intervention class. The pilot tested recording, upload, restricted sharing, and comment procedures rather than learning outcomes. It led to standardized file naming, a three-prompt feedback template, and a low-bandwidth submission alternative. The records do not document formal monitoring of communication between pilot and intervention students, so diffusion of procedures cannot be entirely excluded.

Implementation Fidelity and Practicality

All 150 expected final recordings were submitted (30 students × five cycles), and all 30 students completed both pre- and posttests. Four students required additional technical support during at least one cycle. The original report summarized overall participation as 95%, but the retained manuscript did not provide separate numerators and denominators for attendance, peer-feedback completion, revision uptake, or lecturer-response timing. The revision therefore reports only directly supported completion evidence and explicitly identifies unmeasured fidelity dimensions.

Table 7. Fidelity evidence, interpretation, and limits

Fidelity dimension	Available evidence	Interpretation	Unresolved limitation
Recording completion	150/150 final cycle recordings submitted	Core production requirement was completed	Does not show number of drafts or replay frequency
Assessment completion	30/30 pretests and 30/30 posttests	Complete aggregate outcome dataset	Individual trajectories were not reported
Technical accessibility	Four students needed support; shared microphones/campus access used	Adaptation maintained access	Resource burden and time cost were not logged
Peer-feedback completion and quality	Template use was documented in class logs	Feedback procedure was implemented	Accuracy, specificity, and uptake were not systematically coded
Lecturer feedback timing	A 48-hour rule was specified and reported as implemented	Timeliness was an intended support	No numerator/denominator or timestamp audit
Moderation and privacy	Closed group, respectful-language rule, no public posting	Risk-reduction procedures were present	Rule violations and deletion schedule were not systematically audited

Students rated the total practicality scale at $M = 4.18$ ($SD = 0.46$). Descriptive item-cluster means ranged from 4.10 for technological ease to 4.25 for guide clarity. The two lecturers' combined descriptive mean was 4.38; inferential comparisons and lecturer standard deviations are not emphasized because $n = 2$. These ratings suggest usability in the supported class context but may reflect acquiescence, social desirability, and researcher/instructor involvement.

Table 8. Descriptive practicality results

Descriptive item cluster	Student M	Student SD	Lecturer combined mean (n = 2)
Guide clarity	4.25	0.52	4.50
Technological ease	4.10	0.61	4.25
Feedback usefulness	4.18	0.57	4.40
Learning and cultural relevance	4.20	0.54	4.35
Total descriptive scale	4.18	0.46	4.38

Preliminary Learning Change

The primary three-component speaking-performance score increased from $M = 63.33$ ($SD = 8.10$) to $M = 81.00$ ($SD = 6.70$), with a mean difference of 17.67 points, $t(29) = 12.12$, $p < .001$, 95% CI [14.69, 20.65], $d_z = 2.21$. Fluency, pronunciation, and intonation each showed positive exploratory change. Observable speaking assurance increased from 70.00 to 82.00 but is not interpreted as psychological self-confidence. All dimension-level p values remained below the Holm-adjusted thresholds; nevertheless, the magnitude of change requires cautious interpretation because the design lacked a comparison group, raters were embedded in the course context, and repeated exposure to the rubric may have influenced performance.

Table 9. Pretest–posttest outcomes and construct interpretation

Outcome	Pre M (SD)	Post M (SD)	Mean difference	$t(29)$; p	95% CI	d_z
Speaking performance: 3 dimensions (primary)	63.33 (8.10)	81.00 (6.70)	17.67	12.12; $< .001$	[14.69, 20.65]	2.21
Fluency (exploratory)	60.00 (8.90)	78.00 (7.40)	18.00	10.20; $< .001$	[14.39, 21.61]	1.86
Pronunciation (exploratory)	62.00 (8.30)	80.00 (6.90)	18.00	11.50; $< .001$	[14.80, 21.20]	2.10
Intonation (exploratory)	68.00 (7.60)	85.00 (6.20)	17.00	9.80; $< .001$	[13.45, 20.55]	1.79
Observable speaking assurance (separate)	70.00 (8.00)	82.00 (6.80)	12.00	8.70; $< .001$	[9.18, 14.82]	1.59

Mixed-Methods Joint Display

Table 10. Integration of needs, design, implementation evidence, and quantitative interpretation

QUAL need / design inference	Instructional feature	Implementation evidence	Integrated inference
Every student needed more individual speaking time	Five required individual audio cycles	150/150 final recordings submitted	The sequence successfully distributed production opportunities; it does not establish that dosage alone caused gains
Students needed to hear and revise their own performance	Replay/self-review and opportunity to re-record	Self-review was required in the syntax; replay frequency was not logged	Self-monitoring is a plausible mechanism requiring direct process measurement

QUAL need / design inference	Instructional feature	Implementation evidence	Integrated inference
Feedback needed to be specific and usable	Peer prompts + lecturer rubric feedback	Template and response rule implemented; quality/uptake not coded	Interaction was structurally present, but feedback quality cannot explain effect magnitude
Technology quality was unequal	Low-bandwidth options and shared equipment	Four students received additional support	Practicality depended on institutional support, not only learner devices
Tasks needed meaningful content	Culturally responsive narrative and teaching tasks	High completion and positive relevance ratings	Cultural relevance was feasible, but no independent engagement effect was measured
Pretest–posttest change was large	Primary paired comparison	M increased 17.67 points; dz = 2.21	Promising within-cohort change warrants controlled replication, not causal effectiveness claims

Discussion

The findings demonstrate that the feasibility of learner-generated podcasting does not derive from the audio medium alone but from the structured instructional sequence through which the medium is implemented. The developed sequence comprised seven interconnected stages: understanding the brief and selecting a topic, planning, recording, replaying and self-assessing, receiving peer feedback, receiving lecturer feedback, and revising and reflecting. Expert appraisal, pilot implementation, practicality responses, and preliminary speaking-performance data collectively indicated that this sequence was sufficiently feasible for use in the target course. Nevertheless, feasibility remained conditional. Technological ease received the lowest descriptive rating, and some students required access to microphones, suitable recording spaces, or campus facilities. Although the pretest–posttest analysis showed a statistically large increase, this result should be interpreted as preliminary evidence of instructional promise rather than a definitive causal effect.

The findings are consistent with international studies that conceptualize student-generated podcasts as authentic assessment rather than simply an alternative submission format. Wakefield et al. (2023) found that learner-generated podcasts could create an authentic and enjoyable assessment experience while developing communication and collaborative capabilities. Similarly, Vainieri et al. (2024) reported that podcast-based assessment supported postgraduate students' confidence in communicating scientific knowledge, although students continued to require clearer task guidance and production support. Halabi (2021) likewise demonstrated that podcasting could function as an authentic alternative to conventional examinations, while Moore (2024) emphasized its potential to connect disciplinary learning, critical interpretation, and communication with audiences beyond the classroom. In foreign-language education, Acevedo de la Peña and Cassany (2024) further showed that podcasting can facilitate active language production, multimodal composition, and meaningful engagement with authentic audiences. The present findings extend these studies by indicating that educational value is generated less by podcast novelty than by the integration of explicit criteria, repeated production, self-review, feedback, and revision.

The identified sequence also aligns with research on formative assessment and technology-mediated feedback. Cadena-Aguilar and Álvarez Ayure (2021) found that self- and peer-assessment of student-generated podcasts supported undergraduate EFL learners' awareness of spoken

comprehensibility. Xu and Peng (2022) similarly demonstrated that mobile-assisted peer feedback in a Business English speaking course was shaped by learners' motivational profiles and the quality of their participation. More recent evidence suggests that mobile peer feedback can support speaking development when learners actively provide, interpret, and implement feedback rather than merely receive comments (Ding & Zhu, 2025; Weng et al., 2025). Nevertheless, Kerman et al. (2024) concluded that the effectiveness of online peer feedback depends on the interaction among technology, task design, learner characteristics, scaffolding, and teacher facilitation. The revision stage in the present study is therefore pedagogically important because feedback cannot improve performance unless learners can interpret and enact it. This interpretation is supported by evidence that self- and peer-assessment interventions can strengthen feedback literacy (Hoo et al., 2022), while peer and teacher feedback may produce different patterns of revision effectiveness (Cui et al., 2022).

Within the Indonesian context, the findings correspond with Khotimah et al. (2022), who documented behavioral, cognitive, emotional, and agentic engagement among Indonesian EFL students undertaking podcast-based speaking activities. Their findings, interpreted through self-determination theory, suggest that podcast tasks can support engagement when students experience meaningful choice, competence development, and interaction. The current study nevertheless adds an important qualification: learner choice must be accompanied by sufficient structure and technological support. Evidence from Indonesian online peer-feedback research also indicates that structured and managed feedback can improve performance more effectively than unsupported exchanges (Mulyati & Hadiano, 2023). In a comparable Southeast Asian context, Waluyo and Panmei (2024) found that students participated positively in online peer feedback but that engagement was not automatically associated with improved learning outcomes. This divergence may reflect differences in feedback quality, student proficiency, task complexity, and teacher moderation. It may also be influenced by contextual concerns such as unequal digital access, sensitivity to face-threatening criticism, and students' tendency to regard lecturers as the primary source of authoritative evaluation.

The observed pretest–posttest improvement can be interpreted through three provisional mechanisms. First, requiring every student to produce an individual recording may have increased speaking opportunity and reduced dependence on more proficient group members. Second, replaying recordings against explicit criteria may have enabled students to notice pauses, pronunciation problems, grammatical inaccuracies, and disruptions in fluency that were less perceptible during real-time production. Third, peer and lecturer feedback followed by revision may have encouraged students to convert evaluative information into specific performance adjustments. However, these mechanisms were not measured directly. The increase may also have been influenced by maturation, concurrent coursework, repeated-task familiarity, increasing familiarity with the rubric, instructor enthusiasm, Hawthorne effects, regression to the mean, or growing comfort with recording technology. Consequently, the findings establish testable explanatory propositions but do not confirm mediation or causality.

The culturally responsive dimension must similarly be interpreted as a design condition rather than an independently verified engagement intervention. Allowing students to draw on local experiences, narratives, and cultural knowledge may enhance task relevance, but cultural responsiveness requires more than adding local names, locations, or folktales. Culturally digitized pedagogy involves critical attention to learners' identities, digital participation, cultural resources,

representation, and control over how knowledge is communicated (McDaniel, 2024). Culturally and linguistically responsive pedagogy also requires educators to recognize multilingual resources as legitimate components of learning rather than deficiencies that must be suppressed (Wesley-Nero & Donley, 2024). The sequence therefore needs to preserve learner agency, cultural accuracy, respectful representation, and opportunities to reflect on audience and authorship. Because no validated engagement measure or cultural-comparison condition was included, the study cannot conclude that culturally situated content independently improved student engagement.

Theoretically, the findings support an iterative rather than linear interpretation of the ADDIE framework. Expert appraisal and pilot problems returned the product to the Design stage, while implementation and evaluation evidence informed subsequent Development revisions. Evaluation thus operated throughout the development process instead of functioning solely as its final stage. The findings also extend self-determination theory by suggesting that autonomy is educationally productive when combined with competence-supporting criteria, accessible resources, and socially supportive feedback (Ryan & Deci, 2020). Pedagogically, educators should provide clear speaking criteria, examples of productive feedback, opportunities for replay-based self-assessment, and a compulsory revision stage. Students should also receive guidance in audio production, source attribution, audience awareness, responsible AI use, and the ethical circulation of digital content. To protect emotional well-being, instructors should offer restricted sharing, private submission options, flexible recording locations, and alternatives for students experiencing recording anxiety or limited technological access.

At the policy level, learner-generated recordings should be governed as student data rather than treated as disposable classroom products. Research has shown that students are concerned about institutional tracking, unclear data practices, and insufficient control over the collection and reuse of their educational information (Jones et al., 2020). Students from diverse backgrounds may also differ in their understanding of privacy and their willingness to adopt educational technologies (Kumi-Yeboah et al., 2023). Institutions should therefore establish transparent policies concerning consent, platform selection, ownership, storage duration, access permissions, redistribution, and deletion of student recordings. They should provide loanable recording devices, accessible campus facilities, and secure institutional platforms so that participation does not depend on personal economic resources. Policies should additionally regulate generative AI use in scripting, voice generation, transcription, translation, and pronunciation assistance, distinguishing legitimate learning support from the replacement of students' own oral performance.

The study's novelty lies in its platform-neutral specification of learner-generated audio as a testable instructional sequence rather than its promotion of a particular podcast application. It proposes that a seven-stage speaking cycle is feasible when four interdependent conditions are maintained: ethically restricted sharing, explicit performance criteria, scaffolded and moderated feedback, and accessible technological alternatives. Methodologically, the qualitatively driven mixed-methods design enabled needs evidence, expert appraisal, pilot findings, implementation records, practicality responses, and preliminary performance data to inform successive revisions. The contribution is therefore a theoretically grounded design proposition identifying the components that should be retained, manipulated, or measured in subsequent evaluations.

Several limitations constrain the interpretation and transferability of the findings. The study involved one convenience-selected intact class of 30 students from a single program and did not include a comparison group. The developer's participation in evaluation created a potential

developer–evaluator conflict, while the qualitative archive lacked a formal saturation record, verified quotation system, and systematic negative-case register. Expert appraisal was limited to three validators and did not provide validator-specific scores, item-level agreement, uncertainty estimates, or documented blinding during re-appraisal. The practicality instrument relied on course-based self-report, and dimension-level reliability and full item distributions were unavailable. Furthermore, the speaking assessment did not fully capture content organization, audience awareness, or validated affective outcomes. Delayed retention, student workload, replay frequency, feedback quality, revision uptake, external AI use, and data-retention practices were not systematically examined. Future research should employ multicampus samples, independent evaluators, randomized or quasi-experimental comparison groups, delayed posttests, validated engagement and emotional-well-being measures, and detailed process data. Comparisons among conventional presentations, simple record-and-submit tasks, and the complete seven-stage sequence would clarify which components produce meaningful and sustainable improvement.

Conclusion

This study developed and preliminarily evaluated a platform-neutral interactive podcast instructional sequence for a university Speaking Skills course. The product is defined by a repeatable practice–replay–feedback–revision cycle, culturally responsive task criteria, and privacy and access supports rather than by any specific application. Expert appraisal and user responses indicated that the sequence was feasible in the supported class, and the cohort showed large pretest–posttest improvement in a limited three-component speaking score. Because the evidence came from a one-group developmental trial with incomplete process and psychometric data, the appropriate conclusion is feasibility and preliminary promise, not established effectiveness. The next evaluation should retain the seven-step cycle, restricted sharing, criterion-referenced feedback, moderation, technical alternatives, and culturally responsive task criteria while using multiple classes, external blinded raters, validated affective measures, complete fidelity indicators, individual change displays, and delayed follow-up. This design would allow the proposed mechanisms and contribution to be tested rather than assumed.

References

- Acevedo de la Peña, I., & Cassany, D. (2024). Student podcasting for foreign language teaching–learning at university. *Journal of Technology and Science Education*, 14(1), 123–141. <https://doi.org/10.3926/jotse.2509>
- Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer. <https://doi.org/10.1007/978-0-387-09506-6>
- Cadena-Aguilar, A., & Álvarez Ayure, C. P. (2021). Self- and peer-assessment of student-generated podcasts to improve comprehensibility in undergraduate EFL students. *Profile: Issues in Teachers' Professional Development*, 23(2), 67–85. <https://doi.org/10.15446/profile.v23n2.88928>
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
- Cui, Y., Schunn, C. D., & Gai, X. (2022). Peer feedback and teacher feedback: A comparative study of revision effectiveness in writing instruction for EFL learners. *Higher Education Research & Development*, 41(6), 1838–1854. <https://doi.org/10.1080/07294360.2021.1969541>

- Ding, Y., & Zhu, J. (2025). Mobile based peer feedback in EFL speaking: Learners' motivation, behavioral engagement in feedback provision, and achievement. *Asian-Pacific Journal of Second and Foreign Language Education*, 10, Article 10. <https://doi.org/10.1186/s40862-024-00314-9>
- Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice* (3rd ed.). Teachers College Press.
- Halabi, A. K. (2021). Pivoting authentic assessment to an accounting podcast during COVID-19. *Accounting Research Journal*, 34(2), 156–168. <https://doi.org/10.1108/ARJ-08-2020-0219>
- Hoo, H.-T., Deneen, C. C., & Boud, D. (2022). Developing student feedback literacy through self and peer assessment interventions. *Assessment & Evaluation in Higher Education*, 47(3), 444–457. <https://doi.org/10.1080/02602938.2021.1925871>
- Huang, H., & Li, M. (2024). The impact of technology-enhanced language learning environments on second language learners' willingness to communicate: A systematic review of empirical studies from 2012 to 2023. *Language Learning & Technology*, 28(1), 1–22. <https://doi.org/10.64152/10125/73593>
- Huang, H.-T. D. (2023). Examining the effect of digital storytelling on English speaking proficiency, willingness to communicate, and group cohesion. *TESOL Quarterly*, 57(1), 242–269. <https://doi.org/10.1002/tesq.3147>
- Jones, K. M. L., Asher, A., Goben, A., Perry, M. R., Salo, D., Briney, K. A., & Robertshaw, M. B. (2020). "We're being tracked at all times": Student perspectives of their privacy in relation to learning analytics in higher education. *Journal of the Association for Information Science and Technology*, 71(9), 1044–1059. <https://doi.org/10.1002/asi.24358>
- Kerman, N. T., Banihashem, S. K., Karami, M., Er, E., van Ginkel, S., & Noroozi, O. (2024). Online peer feedback in higher education: A synthesis of the literature. *Education and Information Technologies*, 29(1), 763–813. <https://doi.org/10.1007/s10639-023-12273-8>
- Khotimah, K., Cahyono, B. Y., & Batunan, D. A. (2022). Chronicling Indonesian EFL students' engagement in podcast-based speaking activities in online learning milieu: A self-determination theory perspective. *The JALT CALL Journal*, 18(3), 335–359. <https://doi.org/10.29140/jaltcall.v18n3.621>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kumi-Yeboah, A., Kim, Y., Yankson, B., Aikins, S., & Dadson, Y. A. (2023). Diverse students' perspectives on privacy and technology integration in higher education. *British Journal of Educational Technology*, 54(6), 1671–1692. <https://doi.org/10.1111/bjet.13386>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t* tests and ANOVAs. *Frontiers in Psychology*, 4, Article 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Little, T., Dawson, P., Boud, D., & Tai, J. (2024). Can students' feedback literacy be improved? A scoping review of interventions. *Assessment & Evaluation in Higher Education*, 49(1), 39–52. <https://doi.org/10.1080/02602938.2023.2177613>
- Man, D., Kong, B., & Chau, M. H. (2024). Developing student feedback literacy through peer review training. *RELC Journal*, 55(2), 408–421. <https://doi.org/10.1177/00336882221078380>
- McDaniel, D. S. (2024). Toward culturally digitized pedagogy: Informing theory, research, and practice. *Reading Research Quarterly*, 59(2), 193–210. <https://doi.org/10.1002/rrq.534>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.
- Moore, T. (2024). Pedagogy, podcasts, and politics: What role does podcasting have in planning education? *Journal of Planning Education and Research*, 44(3), 1134–1147. <https://doi.org/10.1177/0739456X221106327>
- Mulyati, Y., & Hadianito, D. (2023). Enhancing argumentative writing via online peer feedback-based essay: A quasi-experiment study. *International Journal of Instruction*, 16(2), 195–212. <https://doi.org/10.29333/iji.2023.16212a>
- Perry, S. (2024). Podcasting as a form of assessment: Increasing student motivation in academic English-speaking assessment. *Language Testing in Asia*, 14, Article 15. <https://doi.org/10.1186/s40468-024-00288-y>

- 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>
- Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61, Article 101860. <https://doi.org/10.1016/j.cedpsych.2020.101860>
- Vainieri, I., Thackeray, L., Hillman, S., Perez, A., Roberts, R., & Panagiotopoulou, E. (2024). Evaluating podcasts as a science communication assessment for postgraduate students. *Innovations in Education and Teaching International*, 61(6), 1446–1459. <https://doi.org/10.1080/14703297.2023.2267047>
- Wakefield, A., Pike, R. K., & Amici-Dargan, S. (2023). Learner-generated podcasts: An authentic and enjoyable assessment for students working in pairs. *Assessment & Evaluation in Higher Education*, 48(7), 1025–1037. <https://doi.org/10.1080/02602938.2022.2152426>
- Waluyo, B., & Panmei, B. (2024). Students' peer feedback engagements in online English courses facilitated by a social network in Thailand. *Journal of Technology and Science Education*, 14(2), 306–323. <https://doi.org/10.3926/jotse.2305>
- Weng, J., Noordin, N., Ahmad, N. K., & Chen, L. (2025). A systematic review of mobile-assisted peer feedback in enhancing speaking proficiency for EFL learners. *Cogent Education*, 12(1), Article 2504228. <https://doi.org/10.1080/2331186X.2025.2504228>
- Wesley-Nero, S., & Donley, K. (2024). Culturally and linguistically responsive pedagogy: Examining teachers' conceptualizations of affirmative instructional practices for multilingual learners. *TESOL Journal*, 15(4), Article e881. <https://doi.org/10.1002/tesj.881>
- Xu, Q., & Peng, H. (2022). Exploring learner motivation and mobile-assisted peer feedback in a business English speaking course. *Journal of Computer Assisted Learning*, 38(4), 1033–1045. <https://doi.org/10.1111/jcal.12660>
- Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. *Education in Medicine Journal*, 11(2), 49–54. <https://doi.org/10.21315/eimj2019.11.2.6>