Applied Research on English Language

V. 16 N. 1 2027
pp: 65-96
http://are.ui.ac.ir

DOI: 10.22108/are.2026.149268.2787
Document Type: Research Article

Effects of a Culturally Embedded AI-Powered Speaking Application on EFL Proficiency and Retention Among Indonesian University Students

Adam Adam [*], Sri Sugiharti, Annisa Putri Melchan

Faculty of Teacher Training and Education, Universitas Riau Kepulauan, Batam, Indonesia

Received: 2026/07/11 **Accepted: 2026/10/10**

Abstract: Many Indonesian university students find speaking English difficult because of anxiety and few opportunities to practice. This study evaluated VozVibe, an Android application that gives automated feedback on learners' speech using automatic speech recognition (ASR) and a large language model (LLM), sets speaking tasks in familiar Indonesian contexts, and uses group-based gamification framed around the Indonesian principle of Gotong Royong (mutual assistance). Its design drew on the Affective Filter Hypothesis, Self-Determination Theory, and sociocultural theory. In a randomized controlled trial, all 72 third-semester English Department students taking a speaking course at an Indonesian university in 2025/2026 were randomly assigned to VozVibe (n = 36) or conventional instruction (n = 36) for 16 weeks. Speaking was tested before, immediately after, and about three months after the intervention. A linear mixed model showed larger improvement in the VozVibe group at the immediate (d = 0.60) and delayed (d = 0.96) post-tests; the later advantage came mainly from the larger gain during the intervention rather than from less forgetting. Questionnaires and interviews suggested that students found the familiar tasks less stressful and valued the group features and self-paced progression. Because the application was compared as a whole with regular classes, the effect of AI feedback cannot be separated from that of additional practice. Culturally embedded AI applications can give students more speaking practice and feedback than a conventional class, and delayed post-tests are needed to check whether such gains last.

Keywords: Adaptive Gamification, Automatic Speech Recognition, Large Language Models, Mobile-Assisted Language Learning, Skill Attrition.

---

[*] **Corresponding Author.**

Authors' Email Address:
[1] **Adam Adam** (harahapadam123@gmail.com) [2] **Sri Sugiharti** (srisugihartiwismono@gmail.com) [3] **Annisa Putri Melchan** (annisaputrimelchan@gmail.com)

2322-5343 © University of Isfahan
This is an open access article under the CC BY-NC-ND/4.0/ License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Indonesian universities have expanded the role of English, including English-medium instruction, partly to make their graduates more competitive (Bolton et al., 2023), yet speaking remains one of the most difficult skills for Indonesian EFL students to develop (Francy & Ramli, 2022).

Two problems are commonly reported. Many students feel anxious when speaking English and fear making mistakes in front of classmates (Fauzi & Asi, 2023; Horwitz et al., 1986; MacIntyre & Gardner, 1994; Mustamir, 2024), and they have few chances to speak English outside class, where communication takes place mainly in Indonesian or local languages (Melansari et al., 2025). Speaking skills learned in class may therefore decline once instruction ends (Khasanah et al., 2024), yet whether gains from technology-enhanced language learning (TELL) last beyond an intervention has received little attention (Golonka et al., 2014; Yang et al., 2023).

Regular classes offer limited time for each student to speak and receive individual feedback, which takes teachers considerable time to provide (Aldino et al., 2024; Yugafiati et al., 2025). Materials and feedback are also often generic, although work on culturally responsive teaching and assessment argues that instruction should build on students' cultural knowledge (Kim, 2024; Steele et al., 2024).

AI-powered applications can offer more individual practice and immediate feedback. Automatic speech recognition (ASR) can support pronunciation learning (Ngo et al., 2024), and large language models (LLMs) can now score and comment on learners' spoken responses (Shankar et al., 2025). Most widely used applications, however, are designed for a general international audience, and research has paid limited attention to how learners' cultural backgrounds shape their use of these tools (Godwin-Jones, 2024; Huertas-Abril & Palacios-Hidalgo, 2023). Indonesian students differ by region, language background, and exposure to global media, but familiar content may still make speaking practice less threatening for many of them. To our knowledge, few studies have combined ASR- and LLM-based feedback with culturally embedded content for Indonesian learners, and fewer have tested whether the gains last.

This study evaluates VozVibe, an Android application developed by the research team for this purpose. Students practice speaking tasks set in familiar Indonesian situations and receive automated feedback on their pronunciation, fluency, grammar, and vocabulary. They can also talk with a chatbot tutor, Vivi. Progression through topics is adjusted to each

student's performance, and points and group-based leaderboards are framed around the Indonesian principle of Gotong Royong (mutual assistance).

In a randomized controlled trial, this study compares VozVibe with conventional instruction in terms of Indonesian EFL learners' speaking proficiency immediately after a 16-week intervention and at a delayed post-test. It addresses the following three research questions:

1. Is there a statistically significant difference in immediate speaking proficiency improvement between the experimental group using VozVibe and the control group receiving conventional speaking instruction?
2. Is there a statistically significant difference between the two groups in speaking proficiency at the delayed post-test, and in the amount of improvement lost between the immediate and delayed post-tests?
3. How do students who used VozVibe perceive its culturally embedded gamification and AI feedback in relation to their motivation and learning, as reported in questionnaires and interviews?

Literature Review

Theoretical Frameworks in Second Language Acquisition (SLA)

The design of VozVibe drew on three frameworks: the Affective Filter Hypothesis, Self-Determination Theory, and sociocultural theory.

Krashen's (1982) Affective Filter Hypothesis proposes that anxiety, low confidence, and low motivation can stop learners from taking in available input. The hypothesis has been criticized as difficult to test and for neglecting learners' output (Gregg, 1984; Swain, 1985), but its central idea is consistent with evidence that anxiety takes up working memory needed for language processing (Eysenck et al., 2007; Wen et al., 2022). Fear of negative evaluation is a well-documented source of speaking anxiety (Horwitz et al., 1986), including among Indonesian students (Mustamir, 2024). VozVibe aimed to reduce this pressure by giving feedback privately and by setting tasks in familiar situations.

Self-Determination Theory (SDT), proposed by Deci and Ryan (2000), holds that intrinsic motivation depends on satisfying the needs for autonomy, competence, and relatedness, and studies suggest that gamification can support motivation and learning in language education when it meets these needs (Chan & Lo, 2024; Luo, 2023; Shen et al., 2024). VozVibe was designed to support autonomy through self-paced progression,

competence through feedback and progression matched to each student's performance, and relatedness through group leaderboards, with rewards adjusted to individual performance (Alberts et al., 2026).

Sociocultural theory (Vygotsky, 1978) sees language development as a social process in which learners can do more with help than alone, within their zone of proximal development (ZPD). Because the ZPD was defined in terms of help from more capable people, the support an AI system provides is better described as scaffolding (Wood et al., 1976), that is, help adjusted to the learner's level, which generative AI can provide (Godwin-Jones, 2024; Law, 2024).

Because VozVibe delivered all of these features together, this study examines their combined effect.

Multimodal AI in Language Acquisition

Earlier ASR systems often had difficulty with non-native speech (Bashori et al., 2024; Ngo et al., 2024). Whisper, trained on a very large multilingual dataset, copes better with accented speech (Radford et al., 2023), but it is still less accurate for non-native than for native speakers, and its accuracy varies with speakers' first language and proficiency (Graham & Roll, 2024); little is known about its performance with Indonesian-accented English. VozVibe used Whisper as its main ASR engine, with an untuned Faster Whisper model as a backup.

Because a transcript does not show how something was said, VozVibe also used SpeechBrain (Ravanelli et al., 2024) to extract acoustic features related to pronunciation, speech rate, and hesitation. Such features are closely related to perceived fluency (Suzuki et al., 2021) and have improved the automatic prediction of expert ratings (Dong et al., 2024). In VozVibe, they served only as a basis for formative feedback and were not used to score the speaking tests; their validity as proficiency measures for Indonesian learners has yet to be examined.

LLMs can take meaning and context into account when scoring and giving feedback. In studies of writing, LLM scores have agreed reasonably well with human scores (Mizumoto & Eguchi, 2023), and LLM feedback has been of somewhat lower quality than that of trained human evaluators on most criteria (Steiss et al., 2024), but less is known about LLM feedback on speech. The prompt for VozVibe's chatbot asked the LLM to take local communicative norms into account, an approach that has so far received little research attention.

The Sociocultural Context in EFL and Adaptive Gamification

Most AI language-learning tools rely on general learner models, with little attention to learners' communicative norms and motivations (Godwin-Jones, 2024; Law, 2024). At the national level, Indonesia has been described as collectivist and high in power distance (Hofstede, 2001), but such dimensions describe national averages, have been criticized for treating national cultures as uniform (McSweeney, 2002), and do not necessarily describe individual learners. In this study, they provided general background for the design.

VozVibe builds Indonesian cultural content into its scenarios and into the prompt that guides its chatbot, an approach we call localized prompt engineering. The scenarios draw on everyday Indonesian situations, local places, food, and legends, and the chatbot is asked to use local forms of address and to bring Batam culture into the conversation. Schema theory holds that prior knowledge, including culturally learned knowledge of familiar situations, supports comprehension and memory (Anderson & Pearson, 1984; Bartlett, 1932). Familiar content may therefore leave learners more attention for the language itself.

Many commercial platforms rely on individual rankings and rewards, but the effects of such features depend on their design: leaderboards have produced mixed results (Li et al., 2024), gamification can have negative effects (Almeida et al., 2023), and expected rewards can reduce intrinsic motivation (Deci et al., 1999). Evidence on cooperative designs is also mixed; with secondary school EFL students, competitive gamification led to better reading outcomes than cooperative gamification (Qiao et al., 2024). VozVibe combined individual progress with group leaderboards framed around Gotong Royong, drawing on the value placed on cooperation in collectivist settings (Triandis, 1995) and on SDT's concept of relatedness.

Skill Attrition and Retention in Technology-Enhanced Language Learning

Few TELL studies examine whether gains last after an intervention ends (Pan et al., 2025; Yang et al., 2023), although this is important in Indonesia, where many students have little contact with English outside class (Zein et al., 2020). Speaking is thought to be more vulnerable to attrition than receptive knowledge because it depends on regular use and fast word retrieval (Schmid, 2023). Loss is not inevitable, however: the level learners reach before instruction ends is one of the main factors in how much they retain (Bardovi-Harlig & Stringer, 2010), so stronger initial learning may also lead to better retention.

Research Gap and the Present Study

In sum, AI-based tools can give learners more speaking practice and immediate feedback, but few have been built around learners' cultural context, and few studies have tested whether the gains last after instruction ends. The present study addresses both points at an Indonesian university. It compares VozVibe with conventional instruction at an immediate and a delayed post-test and examines how students experienced the application. We expected the VozVibe group to improve more by the immediate post-test and to score higher at the delayed post-test.

Methodology

Research Design

This study was a randomized controlled trial with a pre-test, an immediate post-test, and a delayed post-test (Table 1). All participants took the Computer-Assisted Speaking Assessment (CASA) as a pre-test on 7 May 2025 and were randomly assigned to the experimental or control group the following day. The intervention ran for 16 weeks, from 2 June to 15 September 2025, and both groups took the immediate post-test on 16 September. The delayed post-test was administered on 10 December 2025, about three months after the immediate post-test. This interval fell between the students' third and fourth semesters, and no classes were held during it. The VozVibe server was shut down when the intervention ended, so the application could not be used during the interval.

Table 1. Research Design and Timeline

Group	Pre-test (7 May 2025)	Assignment (8 May 2025)	Intervention (16 weeks, 2 June–15 September 2025)	Immediate post-test (16 September 2025)	Interval (about 3 months)	Delayed post-test (10 December 2025)
Experimental (*n* = 36)	CASA	Random	VozVibe application (14 sessions)	CASA	No classes	CASA-P
Control (*n* = 36)	CASA	Random	Conventional instruction (14 sessions)	CASA	No classes	CASA-P

Note. CASA = Computer-Assisted Speaking Assessment. A different set of prompts with the same structure was used at each test point; CASA-P is the set used for the delayed post-test.

Participants and Setting

The participants were all 72 third-semester English Department students at UNRIKA, Batam, Indonesia, enrolled in the Communicative Speaking course, which is taught in the third semester, in the 2025/2026 academic year. All agreed to take part, so the sample was the whole course cohort. Their mean age was 20 years; 58 were women, and 14 were men, and none had used an AI-based language learning platform before. After the pre-test, students were assigned to the groups by simple randomization, conducted by two seventh-semester research assistants who were not involved in teaching the course or recruiting participants. They wrote the anonymized ID codes of all 72 students on identical slips, folded and mixed them in a container, and assigned the first 36 slips drawn to the experimental group and the rest to the control group. The course instructor and the researcher had no role in the allocation. All tests were taken in a proctored language laboratory with individual workstations and noise-attenuating headsets, using a separate computer-based testing application rather than VozVibe.

The VozVibe Intervention and System Architecture

Both groups took the Communicative Speaking course, which follows UNRIKA's Outcome-Based Education (OBE) curriculum and had 14 weekly 60-minute sessions over 16 weeks, with a midterm examination in week 8 and a final examination in week 16. It was the only speaking class for both groups. Both groups were taught by the same English lecturer, who held a master's degree in English Education, had 11 years of teaching experience, was independent of the research team, and was not told the study hypotheses.

In the experimental group, VozVibe replaced conventional instruction in all 14 sessions. Sessions were held in a computer lab because of its good headsets, but students practiced on their own Android phones, on which the application had been installed, with a tutorial the day before the first session. Students could log in only during the scheduled session, so the application could not be used outside class. After attendance was taken, students worked individually and at their own pace through the Speaking Journey and could also talk with the chatbot, Vivi. The instructor did not assign topics, and the application had more topics than any student could complete. The instructor gave additional feedback when students asked for it, monitored their practice, and helped with technical problems. The researcher supervised the sessions but did not take part in the learning activities.

In the control group, the lecturer followed the regular procedure for the course. Each session began with the lecturer introducing the topic and the phrases used for it, followed by whole-class and pair practice and video examples, such as clips from YouTube, chosen by the lecturer within the OBE curriculum. The lecturer gave individual feedback, mostly while students were speaking, and drew the class's attention to it. No digital applications or technology-mediated feedback were used. Attendance was usually taken within the 60 minutes, so practice time was slightly shorter than in the experimental group. Class time was thus similar in both groups but used differently: in the experimental group, each student practiced individually for most of the session with automated feedback throughout, whereas in the control group, speaking time and the lecturer's feedback were shared among 36 students.

Participants were not told the study hypotheses, and control-group students were not shown the features of the VozVibe application. Because the two groups were taught as separate classes, students knew which condition they were in and could not be blinded to it. To reduce the risk of contamination, control-group students were not given accounts, and the application could be used only during the experimental sessions.

Treatment exposure was recorded automatically by the application from the moment a student opened it until the student closed it or access ended at the close of the session; idle time was not excluded. Because students opened the application right after attendance and access ended automatically, recorded times were close to the full session length, although some students closed it a few minutes early when there was not enough time for another topic. Averaged across the 14 sessions, students used the application for 57.44 minutes per session (SD = 0.99), or 95.7% of the 60 minutes. This shows how long the application was open, not how intensively students practiced, and fidelity of delivery was not formally assessed in either group beyond the instructor's monitoring.

Figure 1 summarizes the system. VozVibe is an Android application connected to a Django server. Each spoken response was transcribed with the OpenAI Whisper API, with an untuned local Faster Whisper model as a fallback; SpeechBrain extracted acoustic features related to pronunciation, speech rate, and hesitation; and the transcript and acoustic features were sent to the Gemini API (Gemini 2.5 Flash) to generate feedback. Further technical details are given in Supplementary Material S1.

The Speaking Journey consisted of topics based on the course curriculum, most of them set in Indonesian, mainly Batam, contexts, such as local places, food, and legends, and a few

in general contexts. Each topic presented model phrases, a model conversation, and a vocabulary list, followed by practice activities for fluency, pronunciation, vocabulary, grammar, and listening. In fluency practice, students recorded a timed answer to a question on the topic (e.g., "What's your favorite sport? How often do you play or watch it?") and then saw scores for aspects of fluency with written feedback from the LLM, which commented on their pace, pauses, repetitions, and filler words, sometimes quoted part of their answer, and ended with two or three practice tips. Although the prompt did not ask for it, the feedback usually also praised the student's effort and encouraged further practice. In pronunciation practice, students recorded phrases from the topic and received a score, with mispronounced parts underlined in red. Scores appeared on the topic screen and the leaderboard.

Students could also talk with Vivi, a chatbot built on Gemini 2.5 Flash (temperature 0.7), by voice or in writing; voice conversations used the Gemini Live API (gemini-live-2.5-flash-preview). Vivi was presented as a friendly young tutor from Batam and was instructed to keep its replies short, to bring Batam food, places, and culture into the conversation, to address students with local forms of address such as *Kak* and *Bang*, and occasionally to use *pantun*, a traditional Malay verse form. Its full prompt is given in Supplementary Material S1.

Progression was adaptive. A student unlocked the next topic after attempting the pronunciation, fluency, vocabulary, and grammar activities with an average score of at least 75% and no score below 60%; otherwise, the student continued with the current topic. In addition, an LLM-based coach (Gemini 2.5 Flash, temperature 0.3) used each student's recent and earlier scores and transcripts to suggest which skill and topic to practice next and at what level of difficulty; when the model was unavailable or its output was invalid, a rule-based procedure gave these suggestions instead. The prompts were the same for all students. The scoring rules, model settings, prompts, and error handling are described in Supplementary Material S1.

On registering, each student chose one of several groups named after landmarks and historical places in Batam, such as Barelang and Hang Nadim, and stayed in it throughout. Members could see and chat with one another. Points for completed tasks, pronunciation milestones, and practice streaks counted toward each student's progress and toward the group's score, and groups were ranked against one another on the leaderboard, so every member's practice contributed to the group's position, in line with the principle of Gotong Royong (mutual assistance). Before the study, the application was reviewed in a focus group

discussion with six linguistics experts and ten education practitioners, but it was not piloted with students.

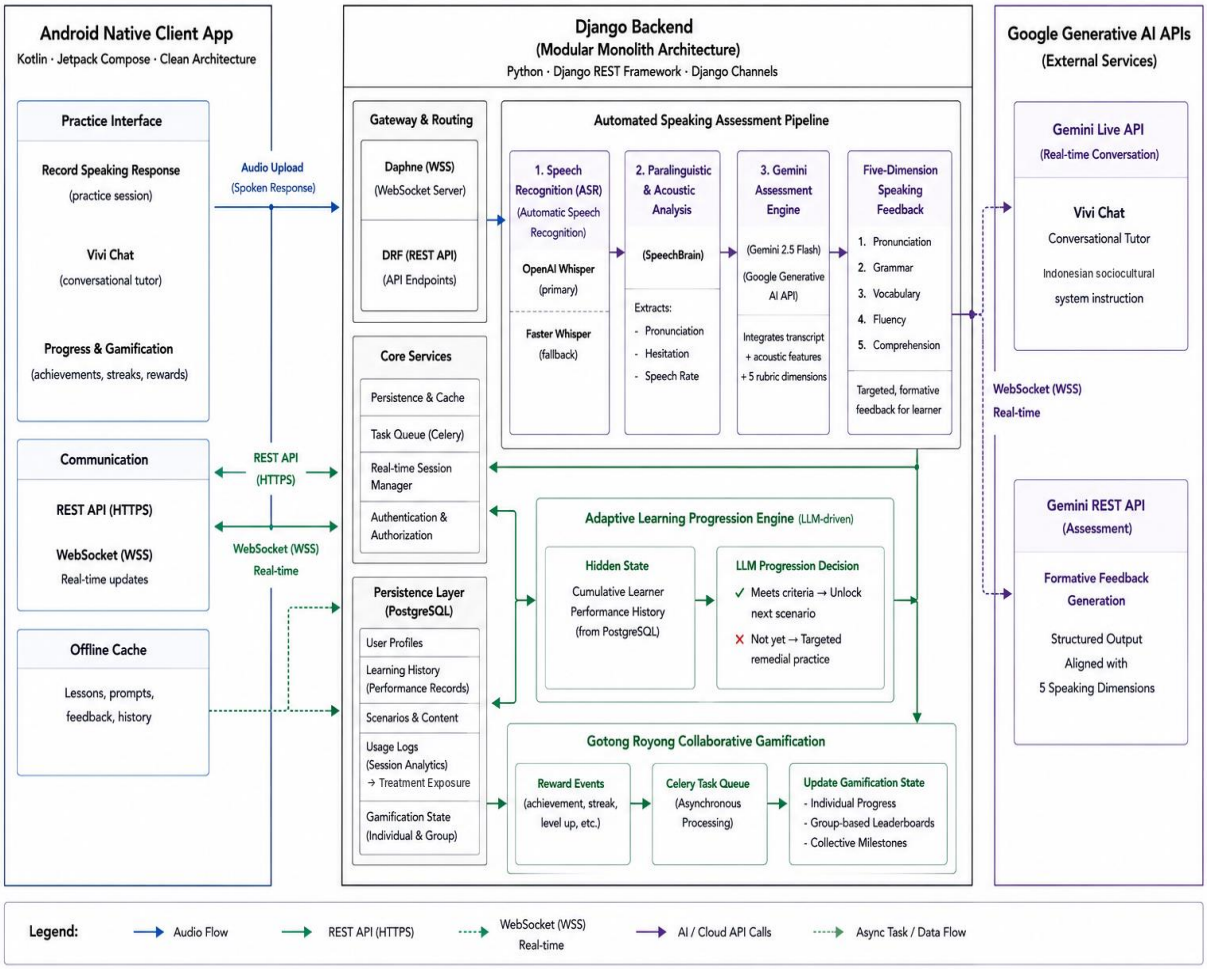


Figure 1. VozVibe System Architecture

Data Collection Instruments

Speaking was measured with the Computer-Assisted Speaking Assessment (CASA), developed by the researcher based on principles of communicative language assessment (Bachman & Palmer, 2010). Different sets of 12 prompts and pictures were used at the three test points, all with four parts: personal experience (four prompts), picture description (one), functional communication (two), and an extended discussion (five), which concerned a personal hobby in the pre-test, a festival or tradition in the immediate post-test, and a community issue in the delayed post-test (CASA-P). The prompts were reviewed by three experts with a PhD in Applied Linguistics, an M.Pd. in English Education, and an M.A. in Psychometrics. Each response was scored from 0 to 10 on fluency and coherence,

pronunciation and intonation, grammar, lexical resource, and pragmatic and sociocultural appropriateness, using a rubric adapted from IELTS and CEFR descriptors and reviewed by two specialists with M.Pd. degrees. The 12 scores were averaged for each dimension, and the overall score was the unweighted mean of the five dimensions (0–10). In a pilot with 10 students outside the sample, scores on two of the sets correlated at r = .80.

For RQ3, the experimental group completed a questionnaire and six students were interviewed. The questionnaire, in English and administered through Google Forms after the delayed post-test, had 28 statements rated on a 5-point scale (1 = strongly disagree, 5 = strongly agree) and one multiple-choice item. Seventeen statements formed five subscales: autonomy, competence, and relatedness (three items each), adapted from the Intrinsic Motivation Inventory (Ryan, 1982), and collectivism and power distance (four items each), written by the research team. Eleven statements concerned the usefulness of the pronunciation, fluency, vocabulary, and grammar practice (two items each) and of Vivi (three items), and the final item asked which features had helped students' speaking most, allowing several answers. There were no open-ended items. The questionnaire was submitted 37 times, one more than the number of students in the experimental group. Because the questionnaire did not ask for names or ID codes, the extra response could not be identified, so all 37 responses were analyzed; removing any single response changed no subscale mean by more than 0.07 and did not change the order of the feature preferences. Cronbach's alpha was .73 for collectivism, .75 for power distance, .71 for autonomy, .75 for competence, .85 for relatedness, and .63–.86 for the feature scales. On the afternoon of the delayed post-test, the researcher interviewed six participants, two each with high, middle, and low immediate post-test scores, in English, using seven questions on cultural relevance, motivation, and the AI feedback; each interview lasted 8–10 minutes. The questionnaire items, the interview questions, and the CASA prompts are given in Supplementary Material S3–S5.

Validity and Reliability

Two raters scored the recordings independently: the researcher and a UNRIKA English lecturer who was independent of the research team and had no contact with the participants. Before scoring, both raters studied the rubric descriptors and practiced by scoring a few sample recordings. Recordings were identified only by ID codes and presented in random order across groups in unlabeled sets, and the key linking codes to groups was held by the research assistants. Because the researcher had supervised the experimental sessions, some

voices may have been recognizable. Inter-rater reliability was estimated with a two-way mixed, absolute-agreement intraclass correlation (ICC), because the same two raters scored every recording and were the only raters of interest. The ICC was .87, indicating good reliability (Koo & Li, 2016). Scores differing by more than one point were resolved through discussion.

Data Analysis Procedures

Data were analyzed in Python. Speaking scores were analyzed with a linear mixed model with Group, Time (pre-test, immediate post-test, delayed post-test), and the Group × Time interaction as fixed effects, using an unstructured covariance matrix because sphericity was not met (Greenhouse–Geisser $\varepsilon = .76$). Three planned contrasts compared the groups' change from pre-test to immediate post-test (RQ1), from immediate to delayed post-test, and from pre-test to delayed post-test (RQ2), with Holm-adjusted p values. Retention was assessed primarily as the raw score loss between the two post-tests, with 95% confidence intervals; the group difference at the delayed post-test was not treated as a measure of retention because it also includes the difference in gains during the intervention. Because the loss scores were not normally distributed, their comparison was also checked with a Mann–Whitney U test and a bootstrap confidence interval (20,000 resamples). Sensitivity analyses comprised a random-intercept mixed model and ANCOVAs on the two post-tests with pre-test scores as the covariate, which increases power in randomized pre-test/post-test designs (Van Breukelen, 2006); normality of residuals (Shapiro–Wilk), homogeneity of variance (Levene's test), and homogeneity of regression slopes were checked. Baseline comparability was described with means, standard deviations, and a standardized mean difference. Tests were two-sided with $\alpha = .05$, and differences were calculated from unrounded scores, so they may differ slightly from those calculated from the rounded means in the tables. The sample size was set by course enrolment rather than an a priori power analysis; with 36 students per group, $\alpha = .05$, and 80% power, differences of $d = 0.67$ or larger could be reliably detected.

For RQ3, a convergent parallel mixed-methods design was used. Questionnaire responses were analyzed descriptively. The researcher transcribed the interviews verbatim and analyzed them with Braun and Clarke's (2006) thematic analysis, combining categories based on the three areas of interest with codes developed from the data; no second coder or member checking was used. Because the researcher had supervised the sessions, participants

may have been reluctant to criticize the application. The two strands were analyzed separately and then compared in a joint display (Table 6).

Ethical Considerations

Ethical approval was granted by the Research Ethics Committee (Komite Etik Penelitian) of the Faculty of Teacher Training and Education, Universitas Riau Kepulauan, Indonesia, after a review based on the Declaration of Helsinki (Approval No. 70/D-FKIP/P/UNRIKA/V/2025, 5 May 2025). All 72 participants signed a consent form before the pre-test stating that participation was voluntary, that they could withdraw at any time without penalty, and that their data would be kept confidential, used only for research and academic purposes, and reported without revealing their identity. The form did not mention the external AI services; in a briefing before the first session, the experimental group was told that the application recorded their speech, that the recordings would be used for research, and that the application used AI services from Google (Gemini) and OpenAI.

In practice activities, students' recordings were sent to the application's server, transcribed with the OpenAI Whisper API (or the local fallback), and analyzed with SpeechBrain, and the transcript and acoustic features were sent to the Gemini API to generate feedback. Input to Vivi, by voice or in writing, was sent to the Gemini API, and Vivi saved no audio. Both services were used through paid accounts whose terms do not allow the providers to use the data to train their models. Recordings from the pronunciation and fluency activities were saved only on students' own phones for replay, and the server kept no copies. Data were transmitted over encrypted (TLS) connections, and token-based (JWT) authentication limited each participant's access to their own records. The CASA recordings, made with the separate testing application, were not processed by these services; they were stored in a Supabase storage bucket under ID codes and deleted after scoring. Questionnaire responses were collected through Google Forms without names or participant ID codes. Test scores were stored under participant ID codes on an institutional server accessible only to the research team.

Control-group participants took the Communicative Speaking course in its usual form, with the same instructor and schedule as the experimental group. They were not given access to VozVibe after the study, because the application's server was shut down when the intervention ended.

Findings

Participant Flow

All 72 students completed the intervention and all three tests, so there was no attrition in either group. According to the course attendance records kept by the instructor, no student missed a session.

RQ1: Immediate Speaking Proficiency Improvement

Tables 2 and 3 show the scores of both groups at the three time points. The Group × Time interaction was significant, $F(2, 69) = 9.12$, $p < .001$, indicating that the two groups changed differently over time. A model with a random intercept only gave the same result, $F(2, 140) = 10.18$, $p < .001$.

Table 2. Descriptive Statistics for Pre-Test and Immediate Post-Test Speaking Proficiency by Group

Group	Measure	*M*	*SD*	*n*
Control	Pre-test	6.67	0.56	36
Control	Immediate post-test	7.36	0.54	36
Experimental	Pre-test	6.62	0.60	36
Experimental	Immediate post-test	7.55	0.65	36

The two groups had similar pre-test scores (experimental: M = 6.62, SD = 0.60; control: M = 6.67, SD = 0.56; d = 0.08). From pre-test to immediate post-test, the experimental group improved by 0.93 points (SD = 0.41) and the control group by 0.69 points (SD = 0.36), so the experimental group's gain was 0.23 points larger, 95% CI [0.05, 0.41], $t(70) = 2.56$, Holm-adjusted $p = .025$, $d = 0.60$.

An ANCOVA on the immediate post-test scores gave a similar adjusted difference of 0.22 points (adjusted means 7.57 and 7.34; Figure 2), 95% CI [0.05, 0.40], $F(1, 69) = 6.47$, $p = .013$, partial $\eta^2 = .086$. Its assumptions were met; the test statistics are reported in Supplementary Material S2.

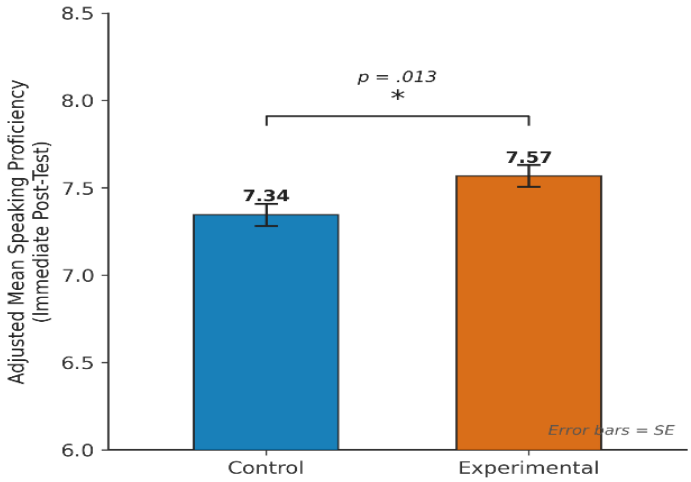


Figure 2. Adjusted Mean Immediate Post-Test Speaking Proficiency by Group

Note. Values are ANCOVA-adjusted means at immediate post-test, controlling for pre-test scores (covariate grand mean = 6.65).

RQ2: Speaking Proficiency at the Delayed Post-Test

Table 3. Descriptive Statistics for Delayed Post-Test Speaking Proficiency by Group

Group	Measure	*M*	*SD*	*n*
Control	Delayed post-test	6.99	0.57	36
Experimental	Delayed post-test	7.29	0.64	36

Note. Delayed post-test = CASA-P, administered about three months after the immediate post-test.

At the delayed post-test (Table 3), the experimental group's mean (M = 7.29, SD = 0.64) exceeded the control group's (M = 6.99, SD = 0.57). Both groups lost part of their gains between the post-tests (Figure 3): 0.26 points in the experimental group, 95% CI [0.20, 0.32], and 0.37 points in the control group, 95% CI [0.29, 0.45], both $p < .001$. The experimental group lost 0.11 points less, 95% CI [0.01, 0.21], $t(70) = 2.12$, Holm-adjusted $p = .038$, $d = 0.50$. Because the loss scores were not normally distributed, this comparison was also checked with a Mann–Whitney U test ($p = .045$) and a bootstrap confidence interval, which gave similar results (Supplementary Material S2). The difference was small in raw terms, however, and was not significant in the random-intercept model ($p = .158$), so it should be interpreted with caution.

Both groups remained above their pre-test scores by 0.67 points in the experimental group, 95% CI [0.51, 0.82], and 0.32 points in the control group, 95% CI [0.25, 0.40]. The improvement from pre-test to delayed post-test was 0.34 points larger in the experimental group, 95% CI [0.17, 0.51], $t(70) = 4.07$, Holm-adjusted $p < .001$, $d = 0.96$, and the ANCOVA gave a similar adjusted difference of 0.33 points, 95% CI [0.17, 0.50], $F(1, 69) = 16.54$, $p < .001$, partial $\eta^2 = .193$. The nonsignificant Group × Pre-test interaction at both post-tests gave no indication that the effect depended on initial proficiency (Supplementary Material S2). Most of the experimental group's advantage at the delayed post-test came from its larger gain during the intervention (0.23 of the 0.34 points) rather than from slower loss afterward (0.11 points).

As a secondary, descriptive measure, a retention index was calculated from the group means as the proportion of the immediate gain still present at the delayed post-test: (delayed post-test − pre-test) / (immediate post-test − pre-test) × 100. The index was 71.9% in the experimental group and 46.7% in the control group. Because the index depends on the size of the initial gain, it is reported only to describe the pattern and was not tested statistically.

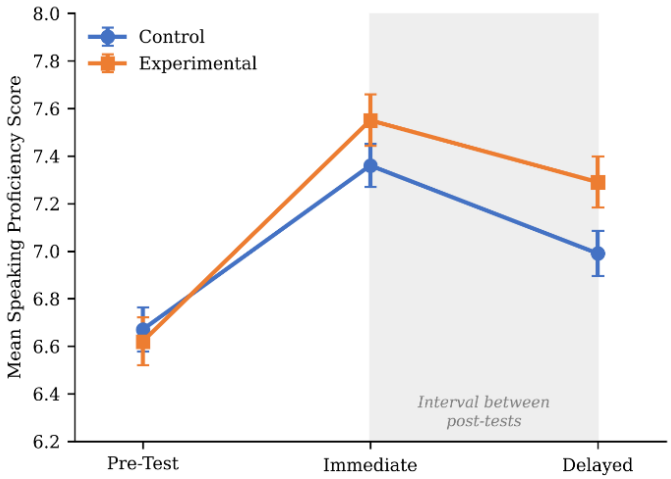


Figure 3. Mean Speaking Proficiency Across Pre-Test, Immediate Post-Test, and Delayed Post-Test by Group

Note. Error bars represent standard errors of the mean (SEM). Shaded regions indicate the interval between the immediate and delayed post-tests.

RQ3: Learner Perceptions of Culturally Embedded Gamification and Multimodal AI Feedback

Quantitative Survey Results

Table 4. Descriptive Statistics for Cultural and Motivational Profiles (37 Responses)

Construct	Sub-Scale	*M*	*SD*
Cultural	Collectivism	3.99	0.80
	Power Distance	3.89	0.84
Motivational Profile	Autonomy	4.00	0.74
	Competence	3.94	0.72
	Relatedness	4.08	0.77

As Table 4 shows, the mean subscale scores were between 3.89 and 4.08 on the 5-point scale, so most students agreed with the statements, although scores this close to the top of the scale may partly reflect a ceiling effect or a wish to answer positively. Relatedness had the highest mean (M = 4.08), followed by autonomy (M = 4.00) and competence (M = 3.94). The collectivism (M = 3.99) and power distance (M = 3.89) scores describe the students' own views of group work and of status in communication. They are reported as background and are not taken as evidence that the application matched Indonesian culture.

Table 5. Perceived Usefulness and Preference for AI Features

AI Feature	Perceived Usefulness (*M*)	Preference Vote (*n*)	Preference Vote (%)
Vocabulary Practice	4.12	19	51.4%
Grammar Practice	4.04	12	32.4%
Fluency Practice	4.03	11	29.7%
Pronunciation Practice	3.96	22	59.5%
Chatbot (Vivi)	3.81	9	24.3%

Note. N = 37 responses. Perceived usefulness was rated on a 1–5 scale. Participants could vote for more than one feature, so preference percentages do not sum to 100%.

Table 5 shows that vocabulary practice received the highest usefulness rating (M = 4.12), whereas pronunciation practice was chosen most often as the most helpful feature (59.5%, n = 22) despite a lower rating (M = 3.96). The ratings and the choices may reflect different things: how useful each feature seemed in general and which feature students valued most for their own speaking. Grammar practice (M = 4.04, 32.4%) and fluency practice (M = 4.03, 29.7%) were in between, and the Vivi chatbot had the lowest rating and was chosen least often (M = 3.81, 24.3%).

Qualitative Interview Results

Three themes were identified in the six interviews: (1) cultural familiarity and comfort in speaking, (2) points, progress, and peers, and (3) low pronunciation scores as a spur to practice. The interviews were conducted in English; participants could switch to Indonesian at any point, but all six responded in English throughout. Quotations are presented verbatim, with ellipses marking omitted words.

Theme 1: Cultural Familiarity and Comfort in Speaking

Participants linked the Batam-specific scenarios, local food, and legends to easier and more comfortable speaking. S20 noted that "the conversation reflects the everyday Indonesian habits... the lesson is easier to understand," which may reflect familiarity with the content as well as reduced anxiety, while S32 described a more personal response:

> "The app is so nice and I can feel like I am in Indonesia because when first time I used it I asked the app about Batam it talked about Batam and also about Batu Aji."

Practicing with the application also seemed to make speaking less stressful. S01 reported feeling able to "mess up without feeling embarrassed," and S32 drew a sharper contrast with people:

> "Compared to when I talked to lecturers or friends they always like judged me but the app just helped me gave me advice."

These comments match the fear of negative evaluation reported among Indonesian and other EFL learners (Fauzi & Asi, 2023; Horwitz et al., 1986). S20 also said that the confidence gained with the application carried over to speaking outside class, becoming "brave enough to speak English by myself, with close friends, and with my family."

Theme 2: Points, Progress, and Peers

Students mentioned three things that kept them practicing: points, the leaderboard, and progress through the Speaking Journey. Points were mentioned most often. S32 "felt excited" earning them and "wanted to try more," while S20 named the competitive side: "the competition with friends made me want to use the app."

Others described the leaderboard mainly as a way to compare themselves with peers. As S07 put it:

> "I like to see my progress in the leaderboard... I can compare my skills with my friends and classmates."

Comments like this suggest that the leaderboard connected students' own progress with that of their classmates, which relates to the SDT need for relatedness (Deci & Ryan, 2000), although S20's comment shows that it also encouraged competition. The Speaking Journey, though mentioned less often, was important for S01, who called it "the only thing I open regularly."

Most participants described their use of the application as voluntary rather than required. S01 said the app "didn't feel like homework," S32 said it "feels like I learn for myself," and S20 gave a mixed view: "I felt both... learning for myself and also doing homework." These comments are consistent with intrinsic motivation, although they were given to the researcher who had supervised the sessions.

Theme 3: Low Pronunciation Scores as a Spur to Practice

Several participants said that low pronunciation scores made them want to practice more. S32 gave the clearest account, referring to the difference between sea and see:

> "I often got a low score that makes me feel frustrated but that helps me try to motivate myself to be better and better."

S07 similarly said that unexpected scores "motivated me to practice more," while S11 described more mixed feelings: "sometimes it's frustrating but the notification really make us want to continue." Not all comments were positive. S01 said that the application "could mess up" at times, which suggests that low scores motivate students only when they trust them.

Comments on the Vivi chatbot were positive but brief. S11 likened it to "talking with a guide," and S32 valued its attention to politeness. When asked which single feature they would keep, most participants chose pronunciation practice; S32 explained that "good pronunciation is very important for us to speak confidently."

Triangulation

Table 6 compares the questionnaire and interview results. The two sources agreed on relatedness, autonomy, and the lower value of the chatbot, and the interviews helped explain why pronunciation practice was chosen most often although vocabulary practice was rated more useful. They differed in one respect: the group features were intended to be cooperative, but several interviewees described the leaderboard in competitive terms. Cultural relevance appeared only in the interviews, as no subscale measured it directly. Because the interviews involved only six students, they were used to explain the questionnaire results rather than to confirm them.

Table 6. Joint Display of Questionnaire and Interview Results

Area	Questionnaire (37 responses)	Interviews (n = 6)	Relation
Relatedness and peers	Relatedness highest subscale (M = 4.08)	Leaderboard used to compare progress with classmates (S07); "the competition with friends made me want to use the app" (S20)	Partly converge: connection with peers, but also competition
Autonomy	Autonomy M = 4.00	Practice did not feel like homework (S01, S32); mixed view (S20)	Converge
Pronunciation feedback	Lowest-rated practice feature (M = 3.96) but	Low scores motivated more practice (S07, S11, S32);	Expand: interviews explain

Area	Questionnaire (37 responses)	Interviews (n = 6)	Relation
	chosen most often (59.5%)	pronunciation linked to confidence (S32); doubts about accuracy (S01)	the choices
Cultural content	No subscale measured cultural relevance directly	Familiar Batam scenarios made lessons easier and speaking more comfortable (S01, S20, S32)	Interviews only
Vivi chatbot	Lowest rating (M = 3.81) and fewest choices (24.3%)	Positive but brief comments (S11, S32)	Converge

Note. M = mean on a 5-point scale. Percentages are the proportion of responses that chose the feature as one of the most helpful.

Discussion

Students who used VozVibe improved their speaking more than students in conventional classes, and this advantage was still present at the delayed post-test. One possible explanation for the larger immediate gain (0.23 points, d = 0.60), consistent with the Affective Filter Hypothesis (Krashen, 1982), is that receiving feedback privately from the application, rather than speaking in front of classmates, reduced anxiety and left more attention for the language itself (Eysenck et al., 2007; Wen et al., 2022). Several interviewees contrasted the application's supportive feedback with the judgment they felt from people, which fits this explanation, but anxiety was not measured, so it remains a possibility.

By the benchmarks proposed for L2 research (Plonsky & Oswald, 2014), the immediate effect falls between small and medium, and the difference at the delayed post-test (d = 0.96) is close to large. In raw terms, however, the immediate advantage of 0.23 points on the 10-point scale is less than a quarter of one rubric point, so its practical value for individual learners is modest.

At the delayed post-test, the experimental group still scored higher (its improvement from pre-test was 0.34 points larger, d = 0.96). Both groups lost part of their gains, as is common for productive skills in EFL settings with little authentic use (Bardovi-Harlig & Stringer, 2010; Schmid, 2023). The experimental group lost slightly less, but the difference

was only 0.11 points and was not significant in every analysis, so its advantage at the delayed post-test mainly reflects its larger gain during the intervention. The results therefore show that VozVibe's advantage outlasted the intervention but offer only weak evidence that it protected learners against forgetting. Culturally meaningful scenarios may have supported deeper initial learning (Anderson & Pearson, 1984; Bartlett, 1932), and the adaptive progression may have helped by keeping tasks at a suitable level of challenge, a form of scaffolding (Wood et al., 1976), although the larger gain may also reflect the greater amount of individual practice and feedback in the experimental group (see Limitations).

The questionnaire and interviews show how students experienced the application. The high relatedness ratings (M = 4.08) and students' comments about comparing their progress with classmates suggest that the group features gave them a sense of connection with their peers (Deci & Ryan, 2000). Some students, however, described the leaderboard in competitive terms, and the groups were ranked against one another, so the design combined cooperation within groups with competition between them. Whether this combination suits Indonesian learners better than individual competition, which has had mixed effects elsewhere (Almeida et al., 2023; Li et al., 2024; Qiao et al., 2024), cannot be judged from this study. Vivi was rated least useful and chosen least often (24.3%). Its use was optional and less structured than the practice activities, and students may have preferred feedback linked to scores, but the data do not show which reason applies. The power distance items raise a further question: students who feel uncomfortable making mistakes in front of a teacher may find feedback from an application less threatening, but they may also trust it less than a teacher's judgment.

Some interviewees also described low pronunciation scores as motivating rather than discouraging. This is consistent with Hattie and Timperley's (2007) view that feedback is most useful when it shows a clear gap between current and desired performance. However, the observation comes from a few of the six interviewees, and other explanations are possible, such as competitiveness or a general belief that effort leads to improvement. It also seemed to depend on students trusting the scores, since one student noted that the application could "mess up." The observation would need to be tested with a larger sample.

The results are consistent with the reasoning behind VozVibe's design, but because the application was evaluated as a whole, the study cannot show how much each element contributed or whether the elements reinforced one another. What it adds is delayed post-test

evidence that the advantage gained during the intervention was still present after it ended, which remains relatively scarce in TELL research (Golonka et al., 2014; Yang et al., 2023).

The use of such applications also has possible drawbacks. Replacing class sessions with individual practice reduces interaction with the teacher and classmates, students may come to rely on automated feedback that is not always accurate, sending learners' speech to commercial AI services raises privacy questions, and students without suitable phones or a stable internet connection may be disadvantaged.

Limitations

Several limitations should be considered. First, VozVibe was tested as a complete package: ASR-based pronunciation feedback, LLM-generated feedback, the Vivi chatbot, adaptive progression, Indonesian scenarios, and Gotong Royong gamification were delivered together and compared with conventional instruction. The results show the effect of the application as a whole, not which components contributed, how much, or whether they strengthened each other, and the questionnaire and interviews show how students perceived the features, not how much each contributed to their scores. Dismantling or factorial designs, for example, comparing the full application with versions without the gamification or the cultural content, could identify the active components.

A related limitation concerns the amount of practice and feedback. The study compared VozVibe sessions with the regular class, not with an equal amount of practice and feedback without AI. Students in the experimental group practiced individually for most of each session with automated feedback, whereas in the control group speaking time and feedback were shared among 36 students, so the experimental group probably spoke more and received more feedback, although neither was recorded. The study therefore cannot show whether the advantage at either post-test came from the AI technology, from more practice, from more frequent feedback, or from a combination of these. It shows that replacing conventional speaking classes with VozVibe sessions led to higher scores in this setting, not that AI feedback is more effective than the same amount of practice and feedback provided in other ways. Future studies could use a control condition matched for individual speaking time and feedback, such as recorded practice with teacher feedback.

Novelty and Hawthorne effects may also have played a role. The application was new to the students, and knowing that they were using it as part of a study, under the researcher's supervision, may have increased their effort; the enthusiasm of the researcher and the

instructor may also have influenced them. The 16 weeks gave some time for novelty to wear off, but changes in interest were not examined, and such effects could also have influenced the delayed results. An active control condition that is also new to students, together with tracking of engagement, could address this.

Two further limitations concern the delayed post-test. Although no classes were held between the two post-tests, the students' contact with English outside class during this period, for example through films, social media, or work, was not recorded. Such contact may be more frequent in Batam, an industrial and trade city close to Singapore, than in many other parts of Indonesia, and it may have reduced the loss of speaking skills in both groups. Because the students were randomly assigned, there is no particular reason to expect that this exposure differed between the groups, but this cannot be confirmed. In addition, a single delayed post-test about three months after the immediate post-test shows how much of the gain was still present at that point, not how well speaking skills are retained over longer periods.

The measures also have limitations. The CASA was developed for this study: its validity evidence comes from expert review and a small pilot, and its factor structure was not examined. The three prompt sets were not statistically equated, and the pre-test discussion topic (a personal hobby) may have been easier than those in the post-tests; because both groups took the same sets, this would affect both groups' gains similarly rather than the comparison between them. The accuracy of the speech recognition and AI feedback was checked only informally before the study: digital literacy was not assessed, and anxiety, schema activation, and cognitive load were not measured. The questionnaire was given once, only to the experimental group, without open-ended items and with two or three items per feature, and its scores were close to the top of the scale, which may reflect ceiling or social desirability effects. The interviews were short and were conducted, transcribed, and coded by the researcher who had supervised the sessions, without a second coder or member checking.

Finally, the study involved English Department students at a single institution, a modest sample, and six interviewees, which limits generalizability, particularly to non-English majors. With 36 students per group, only differences of about $d = 0.67$ or larger could be reliably detected, so the immediate effect ($d = 0.60$) and the difference in loss ($d = 0.50$) are estimated imprecisely, as their wide confidence intervals show. The questionnaire and interviews took place after the intervention had ended, so students described their experience from memory. Future research should compare Gotong Royong and individualistic

gamification directly and test whether culturally localized AI applications work in other Indonesian regions and other EFL settings.

Conclusion

This study asked whether a culturally embedded AI speaking application could improve speaking more than conventional instruction and whether the advantage would last. At the immediate post-test, it did, and the experimental group still scored higher about three months later, mainly because of its larger gain during the intervention. Students' accounts suggest that the familiar scenarios made speaking practice less stressful and that the group features and self-paced progression supported their sense of relatedness and autonomy. These findings come from one institution, 72 students, and one delayed post-test, and they show the effect of the application as a whole, not of any single feature or of AI feedback compared with the same amount of practice provided in other ways. Within these limits, the results suggest that culturally embedded design is worth testing further and that TELL studies should include delayed post-tests. For teachers, applications like VozVibe offer a way to give students more speaking practice and feedback than a conventional class can provide. They require students' own Android phones, a stable internet connection, headsets, a server, paid access to AI services, and teacher time for technical support, and they may be most useful alongside interactive class activities, with teachers checking the feedback and explaining to students how their speech data are processed.

Declaration of Conflicting Interests

The authors declare no conflict of interest. VozVibe was developed by members of the research team specifically for the purposes of this study; it is not commercially available, and the authors have no financial interest in the application.

Funding

This work was supported by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemendiktisaintek) under Grant 0070/C3/AL.04/2025 (BIMA Fundamental Research Grant Program, Fiscal Year 2025).

Declaration of Generative AI Use

During the revision of this manuscript, the authors used Claude (Anthropic) to help edit the language and revise the text. The authors reviewed and edited all suggestions and take full responsibility for the content of this article.

References

Alberts, L., Lyngs, U., & Lukoff, K. (2026). Designing for sustained motivation: A review of self-determination theory in behaviour change technologies. *Interacting with Computers, 38*(3), 447–468. https://doi.org/10.1093/iwc/iwae040

Aldino, A. A., Tsai, Y.-S., Mello, R. F., Gašević, D., & Chen, G. (2024). Enhancing feedback quality at scale: Leveraging machine learning for learner-centered feedback. *Computers and Education: Artificial Intelligence*, *7*, Article 100332. https://doi.org/10.1016/j.caeai.2024.100332

Almeida, C., Kalinowski, M., Uchôa, A., & Feijó, B. (2023). Negative effects of gamification in education software: Systematic mapping and practitioner perceptions. *Information and Software Technology, 156*, Article 107142. https://doi.org/10.1016/j.infsof.2022.107142

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (Vol. 1, pp. 255–291). Longman.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bardovi-Harlig, K., & Stringer, D. (2010). Variables in second language attrition: Advancing the state of the art. *Studies in Second Language Acquisition, 32*(1), 1–45. https://doi.org/10.1017/S0272263109990246

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). I Can Speak: Improving English pronunciation through automatic speech recognition-based language learning systems. *Innovation in Language Learning and Teaching, 18*(5), 443–461. https://doi.org/10.1080/17501229.2024.2315101

Bolton, K., Hill, C., Bacon-Shone, J., & Peyronnin, K. (2023). EMI (English-medium instruction) in Indonesian higher education. *World Englishes, 42*, 424–446. https://doi.org/10.1111/weng.12622

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chan, S., & Lo, N. (2024). Enhancing EFL/ESL instruction through gamification: A comprehensive review of empirical evidence. *Frontiers in Education, 9*, Article 1395155. https://doi.org/10.3389/feduc.2024.1395155

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668. https://doi.org/10.1037/0033-2909.125.6.627

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Dong, W., van Hout, R., Cucchiarini, C., & Strik, H. (2024). L2 prosody assessment by combining acoustic and neural model features. In *Proceedings of Speech Prosody 2024* (pp. 165–169). https://doi.org/10.21437/SpeechProsody.2024-34

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Fauzi, I., & Asi, N. (2023). Examining speaking anxiety of Indonesian learners of English: A case of university students. *Indonesian Journal of English Language Teaching and Applied Linguistics*, *7*(2), 263–275. https://doi.org/10.21093/ijeltal.v7i2.1338

Francy, F., & Ramli, R. (2022). Problems encountered by Indonesian EFL learners in mastering speaking skills. *Pioneer: Journal of Language and Literature, 14*(1), 1–14. https://doi.org/10.36841/pioneer.v14i1.1176

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI. *Language Learning & Technology, 28*(2), 5–31. https://doi.org/10.64152/10125/73570

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning, 27*(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters, 4*(2), Article 025206. https://doi.org/10.1121/10.0024876

Gregg, K. R. (1984). Krashen's monitor and Occam's razor. *Applied Linguistics, 5*(2), 79–100. https://doi.org/10.1093/applin/5.2.79

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Sage.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125–132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Huertas-Abril, C. A., & Palacios-Hidalgo, F. J. (2023). New possibilities of artificial intelligence-assisted language learning (AIALL): Comparing visions from the East and the West. *Education Sciences, 13*(12), Article 1234. https://doi.org/10.3390/educsci13121234

Khasanah, Y. U., Widiati, U., Tresnadewi, S., & Wulyani, A. N. (2024). English language attrition levels and language maintenance efforts among Indonesian non-English department students. *Register Journal, 17*(1), 73–99. https://doi.org/10.18326/register.v17i1.73-99

Kim, H. J. (2024). Guidelines for culturally and linguistically responsive instruction in higher education. *TESOL Journal, 15*, Article e807. https://doi.org/10.1002/tesj.807

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

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open, 6*, Article 100174. https://doi.org/10.1016/j.caeo.2024.100174

Li, C., Liang, L., Fryer, L. K., & Shum, A. (2024). The use of leaderboards in education: A systematic review of empirical evidence in higher education. *Journal of Computer Assisted Learning, 40*(6), 3406–3442. https://doi.org/10.1111/jcal.13077

Luo, Z. (2023). The effectiveness of gamified tools for foreign language learning (FLL): A systematic review. *Behavioral Sciences, 13*(4), Article 331. https://doi.org/10.3390/bs13040331

MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning, 44*(2), 283–305. https://doi.org/10.1111/j.1467-1770.1994.tb01103.x

McSweeney, B. (2002). Hofstede's model of national cultural differences and their consequences: A triumph of faith—a failure of analysis. *Human Relations, 55*(1), 89–118. https://doi.org/10.1177/0018726702551004

Melansari, D., Fransisca, D., & Wahyuni, S. (2025). Indonesian student's difficulties in speaking English. *JEIT (Journal of Educational Innovations and Technologies), 1*(2), 150–167. https://doi.org/10.63324/jeit.1.2.2025.98

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mustamir, M. (2024). Indonesian EFL learners and speaking anxiety: Insights from a meta-synthetic analysis. *English Review: Journal of English Education, 12*(2), 509–518. https://doi.org/10.25134/erjee.v12i2.9950

Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL, 36*(1), 4–21. https://doi.org/10.1017/S0958344023000113

Pan, M., Lai, C., & Guo, K. (2025). Technology-enhanced self-regulated second language learning interventions: A systematic review. *System, 131*, Article 103675. https://doi.org/10.1016/j.system.2025.103675

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Qiao, S., Yeung, S. S., Shen, X. A., Leung, J. K. L., Ng, D. T. K., & Chu, S. K. W. (2024). How competitive, cooperative, and collaborative gamification impacts student learning and engagement. *Language Learning & Technology*, *28*(1), 1–19. https://doi.org/10.64152/10125/73546

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th*

International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202, pp. 28492–28518). PMLR.

Ravanelli, M., Parcollet, T., Moumen, A., de Langen, S., Subakan, C., Plantinga, P., Wang, Y., Mousavi, P., Della Libera, L., Ploujnikov, A., Paissan, F., Borra, D., Zaiem, S., Zhao, Z., Zhang, S., Karakasidis, G., Yeh, S.-L., Champion, P., Rouhe, A., . . . Estève, Y. (2024). Open-source conversational AI with SpeechBrain 1.0. *Journal of Machine Learning Research, 25*(333), 1–11. https://jmlr.org/papers/v25/24-0991.html

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*(3), 450–461. https://doi.org/10.1037/0022-3514.43.3.450

Schmid, M. S. (2023). The final frontier? Why we have been ignoring second language attrition, and why it is time we stopped. *Language Teaching*, *56*(1), 73–93. https://doi.org/10.1017/S0261444822000301

Shankar, N. B., Zhang, K., Mai, A., Shi, M., Long, A., Washington, J., Morris, R., & Alwan, A. (2025). Leveraging ASR and LLMs for automated scoring and feedback in children's spoken language assessments. In *Proceedings of the 10th Workshop on Speech and Language Technology in Education* (pp. 1–5). International Speech Communication Association. https://doi.org/10.21437/SLaTE.2025-1

Shen, Z., Lai, M., & Wang, F. (2024). Investigating the influence of gamification on motivation and learning outcomes in online language learning. *Frontiers in Psychology, 15*, Article 1295709. https://doi.org/10.3389/fpsyg.2024.1295709

Steele, C., Gower, G., & Bogachenko, T. (2024). Creating and enacting culturally responsive assessment for First Nations students in higher education settings. *Australian Journal of Education, 68*(2), 84–102. https://doi.org/10.1177/00049441241258496

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal, 105*(2), 435–463. https://doi.org/10.1111/modl.12706

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Triandis, H. C. (1995). *Individualism & collectivism*. Westview Press.

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*(9), 920–925. https://doi.org/10.1016/j.jclinepi.2006.02.007

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wen, Z., Teng, M. F., Han, L., & Zeng, Y. (2022). Working memory models and measures in language and bilingualism research: Integrating cognitive and affective perspectives. *Brain Sciences, 12*(6), Article 729. https://doi.org/10.3390/brainsci12060729

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yang, Y., Wen, Y., & Song, Y. (2023). A systematic review of technology-enhanced self-regulated language learning. *Educational Technology & Society, 26*(1), 31–44. https://doi.org/10.30191/ETS.202301_26(1).0003

Yugafiati, R., Yuliasri, I., & Widhiyanto, W. (2025). A decade of research on speaking for academic purposes (2015–2025): Methodological patterns, pedagogical tools, and emerging directions. *Proceedings of English Language Teaching, Literature, and Translation (ELTLT), 14*, 516–530. https://proceeding.unnes.ac.id/eltlt/ article/view/ 4720

Zein, S., Sukyadi, D., Hamied, F. A., & Lengkanawati, N. S. (2020). English language education in Indonesia: A review of research (2011–2019). *Language Teaching*, *53*(4), 491–523. https://doi.org/10.1017/S0261444820000208