Language Ecology and Application
WISDOM ACADEMIC https://doi.org/10.63808/lea.v2i1.471 Volume 2 , Issue 1

Research Article

From Smartphone to Smart Learning: A Quasi-Experimental Study of an AI-Enabled Mobile Application for English Vocabulary Learning

Liyun Chen[1], Mohamad Jafre Zainol Abidin[1,*]
[1]Faculty of Education & Liberal Science, City University Malaysia, Petaling Jaya 46100, Malaysia
[*]**Correspondence:** Mohamad Jafre Zainol Abidin, mohamad.jafre@city.edu.my

© The Author(s) 2026.
This article is published by Wisdom Academic Press Ltd. under a Creative Commons Attribution 4.0 International License. The license permits use, sharing, adaptation, distribution, and reproduction in any medium or format, provided that appropriate credit is given to the original author(s) and the source, a link to the license is supplied, and any modifications are indicated. Unless otherwise stated, all images and third-party materials included in this article are also covered by the same license. For any material not covered by this license, if the intended use is neither permitted by applicable law nor allowed under the license terms, direct permission must be obtained from the copyright holder. To view a copy of this license, please visit http://creativecommons.org/licenses/by/4.0/.

Abstract

Mobile vocabulary applications may support out-of-class learning, yet access alone does not resolve irregular review, weak personalization, or limited monitoring of practice. This 12-week quasi-experimental teaching case compared a bundled AI-enabled vocabulary application with non-adaptive printed and digital word lists among 100 second-year non-English-major students in eight intact tutorial classes at a university in eastern China. Four matched class pairs were randomized at the class level, and complete-case analyses included 46 students per condition. Vocabulary performance was assessed with a 100-point same-item test administered in counterbalanced orders; application logs, comparison-condition study diaries, attendance records, and a four-dimensional attitude questionnaire were collected. A linear mixed model with a random intercept for class estimated a 6.6-point adjusted posttest advantage for the application condition, 95% CI [1.3, 11.9], $p = .022$, with an intraclass correlation of .06. A conventional student-level ANCOVA yielded a 7.0-point difference, $F(1, 89) = 18.42$, $p < .001$, partial eta squared = .17. Active study days were associated with vocabulary gain in the unadjusted analysis, $r = .46$, $p = .001$, and remained associated after adjustment for pretest score and course attendance, $B = 0.21$, $p = .014$; the adjusted association for review completion was weaker. Perceived usefulness received the highest rating ($M = 4.12$), whereas continuance intention was lower ($M = 3.61$). The findings illustrate a plausible short-term advantage for the complete instructional package, while the small number of classes, same-item testing, self-reported comparison-condition study time, and bundled intervention prevent feature-specific causal conclusions.

Keywords

Mobile-assisted language learning, Artificial intelligence, English vocabulary learning, Spaced repetition, Quasi-experiment

1. Introduction

1.1. Research background and practical problem

Mobile-assisted language learning has become a routine extension of university English instruction because smartphones allow repeated access to short learning activities beyond the classroom. A systematic review of 687 Web of Science articles shows that vocabulary learning and learner perceptions remain central themes in this field (Okumuş Dağdeler, 2023), while bibliometric evidence identifies artificial-intelligence-supported personalization as an expanding line of inquiry (Karakaya & Bozkurt, 2022). These developments have broadened the range of available tools, but they do not show that mobile access alone produces durable vocabulary growth.

Everyday vocabulary practice is often weakened by irregular review, repeated attention to already familiar items, and limited feedback on productive use. Static word lists can provide broad coverage, yet they usually do not adjust review timing to individual recall history. Mobile applications may address part of this problem through adaptive scheduling, error-based recommendations, contextual examples, and pronunciation feedback. Their educational value nevertheless depends on how consistently learners use the available functions and on whether the comparison condition receives equivalent exposure and study support.

1.2. Research problem

Evidence for mobile language learning is heterogeneous in design and quality. Burston and Giannakou (2022) identified 84 methodologically adequate experiments from 1,144 citations published between 1994 and 2019 and noted recurring limitations involving short interventions, limited verification of tool use, and reliance on researcher-developed outcome measures. A further difficulty is that applications commonly combine mobile delivery, adaptive algorithms, generative dialogue, and speech feedback. A comparison between a complete application and a conventional list condition can estimate the performance difference between packages, but it cannot assign that difference to a single AI mechanism.

A semester-length design that brings achievement, process logs, and learner attitudes into one analytical frame can improve ecological relevance. Such a design still requires bounded interpretation. A small number of randomized classes, repeated use of the same test items, reliance on self-reported study time in the comparison condition, and simultaneous delivery of several application functions can all affect the precision and meaning of the estimated difference.

1.3. Study aims and contribution

The study had three connected aims. It estimated the class-adjusted posttest difference between students using a bundled AI-enabled vocabulary application and students using non-adaptive word lists, examined whether logged study persistence was associated with vocabulary gain within the application condition, and described learner ratings of usefulness, ease of use, satisfaction, and continuance intention. The contribution lies in modeling transparent reporting for a twelve-week Chinese university EFL evaluation by linking outcome data to recorded use, documenting the intervention and comparison procedures, and distinguishing the estimate for the complete package from claims about individual features. The design does not isolate adaptive scheduling, chatbot support, or speech recognition as independent causes.

2. Literature review and theoretical framework

2.1. Development of mobile-assisted vocabulary learning

Research on mobile-assisted vocabulary learning has moved from simple digital delivery toward increasingly interactive systems. In a three-wave study with pretest, posttest, and delayed assessment, digital flashcards supported stronger word learning and retention than paper word lists (Zakian et al., 2022). Mobile delivery can therefore add value even without complex AI functions, possibly because it increases accessibility and supports repeated retrieval. More elaborate applications combine multimedia and interaction. Chen et al. (2025), for example, reported stronger vocabulary outcomes and more favorable technology perceptions for an augmented-reality application with a pedagogical agent than for card-based and conventional conditions. These studies demonstrate that mobile learning encompasses materially different interventions rather than a single method.

2.2. Distinct AI-enabled mechanisms in vocabulary learning

Adaptive scheduling and recommendation systems use learner-response histories to determine which items should reappear and when review should occur. Their main instructional function is to allocate practice according to errors and estimated forgetting rather than to generate new language. Generative chatbots serve a different purpose by producing examples, explanations, or short exchanges. Experimental evidence suggests that chatbot-supported practice may benefit receptive and productive vocabulary performance (Zhang & Huang, 2024) and may support incidental and collocational learning in self-regulated settings (Jalambo et al., 2025). These findings concern generative interaction

and should not be treated as evidence for scheduling algorithms.

Automatic speech recognition addresses spoken production by comparing learner speech with a recognition model and returning pronunciation-related feedback. Mixed-methods evidence indicates that speech-recognition-supported practice can contribute to speaking development under some instructional arrangements (Sun, 2023). AI-enabled mobile assignments that combine automated feedback with individualized guidance have also been associated with stronger performance than general messaging-based practice (Ma et al., 2025). The mechanisms, data requirements, and risks of these technologies differ. A bundled application comparison can evaluate their combined instructional use, but feature-level effects require component-specific experimental conditions.

2.3. Theoretical rationale

The spacing effect provides a rationale for distributing retrieval across time. A meta-analysis of 98 effect sizes from 48 studies involving 3,411 participants found that spaced practice produced better second-language learning than massed practice (Kim & Webb, 2022). This evidence supports the inclusion of a review scheduler, although the present study did not manipulate spacing intervals independently or test spacing as a mediator.

Dual-coding and cognitive-load perspectives are relevant when vocabulary is presented through verbal, visual, and auditory channels. Li et al. (2022) showed that multimodal input can influence EFL vocabulary outcomes, with benefits depending on how information is coordinated and how much processing burden is created. In the application examined here, contextual examples and pronunciation feedback provided additional channels. The theoretical constructs were not measured directly, so they function as design rationales rather than tested explanatory variables.

2.4. Research gap and study position

Existing studies provide growing evidence for digital flashcards, adaptive practice, chatbots, and speech recognition, yet process-level data remain less common than outcome and perception measures. Research on mobile academic vocabulary learning suggests that digital flashcards can influence self-regulatory capacity as well as test performance (Boroughani et al., 2023), which underscores the importance of recorded study behavior. The present study addresses this gap by combining adjusted achievement comparisons with usage logs and attitude measures. Its ecological strength is the evaluation of a complete application package under ordinary teaching conditions; its corresponding limitation is the inability to identify which component, if any, accounts for the observed difference.

2.5. Analytical expectations

The literature supported a directional expectation that the application condition would obtain a higher adjusted posttest vocabulary score than the list condition. Within the application condition, more active study days and a higher review-completion rate were expected to be associated with larger gains. Learner ratings were expected to remain above the scale midpoint, with perceived usefulness likely to exceed continuance intention because structured course use does not necessarily translate into voluntary long-term adoption.

3. Method

3.1. Design, setting, participants, and allocation

A twelve-week quasi-experimental comparison was conducted from March to June 2025 at a provincial comprehensive university in eastern China. Participants were second-year students who were not majoring in English and had passed the College English Test Band 4. The instructional context represented intermediate university EFL study rather than beginner vocabulary learning. The participating cohort was organized into eight intact tutorial classes attached to the same required College English course, with 12 or 13 students in each class before attrition. Eligibility required enrollment in the course, adult consent, and access to a smartphone capable of running the teaching prototype.

One hundred students entered the study, with 50 allocated to each condition. Ninety-two completed the required measures, leaving 46 complete cases in each condition. Four students were lost from the application condition: one withdrew from the course, one discontinued after recurring login instability following a phone replacement, and two were absent from the posttest because of illness or an internship conflict. Four students were lost from the list condition: one withdrew from the course and three were absent from the posttest because of illness or timetable conflict. The mean baseline score among the eight non-completers was 60.9 (SD = 8.2), compared with 62.0 (SD = 8.0) among completers. This small descriptive difference does not rule out attrition bias, particularly because one withdrawal in the application condition was related to access difficulty.

Assignment occurred at the intact-class level rather than at the individual level. The eight classes were arranged into four pairs according to baseline class mean and teaching time, with morning and afternoon sections balanced where possible. Within each pair, a research assistant who was not involved in teaching used a computer-generated random sequence to assign one class to the application condition and the other to the list condition. Four classes were therefore allocated to each condition. Initial class sizes were 12 or 13, and complete-case

class sizes ranged from 11 to 12. The same instructor taught all eight tutorial classes with a common syllabus and lesson plan, while pretest and posttest administration was supervised by two assistants who were not informed of the class allocation during scoring. Because only eight clusters were available, class-adjusted estimates remain imprecise even though the analysis accounts for nesting.

The target set comprised 480 academic words drawn from the Academic Word List and balanced across frequency bands and major word classes. A preliminary familiarity screening removed items correctly recognized by more than 80% of the cohort, after which replacements were drawn from the same frequency band. Both conditions worked with the same target set, definitions, and core example sentences. Table 1 reports the characteristics of the 92 complete cases and the class structure retained for analysis.

Table 1

Participant characteristics and group composition

Characteristic	Application condition (n = 46)	List condition (n = 46)	Baseline comparison
Female, n (%)	27 (58.7)	25 (54.3)	chi-square = 0.18, p = .67
Age, M (SD)	19.6 (0.8)	19.5 (0.7)	t = 0.62, p = .54
Prior English study, years, M (SD)	8.4 (1.1)	8.3 (1.2)	t = 0.39, p = .70
Pretest vocabulary score, M (SD)	62.3 (8.1)	61.7 (7.9)	t(90) = 0.41, p = .68
Allocated students, n	50	50	Class-level allocation
Participating intact classes, n	4	4	Eight classes in total
Complete-case class size, median (range)	11.5 (11-12)	11.5 (11-12)	Comparable cluster sizes

Note. Group sizes refer to complete cases. Four intact tutorial classes were assigned to each condition through matched-pair class randomization; complete-case class sizes ranged from 11 to 12.

3.2. Intervention and comparison condition

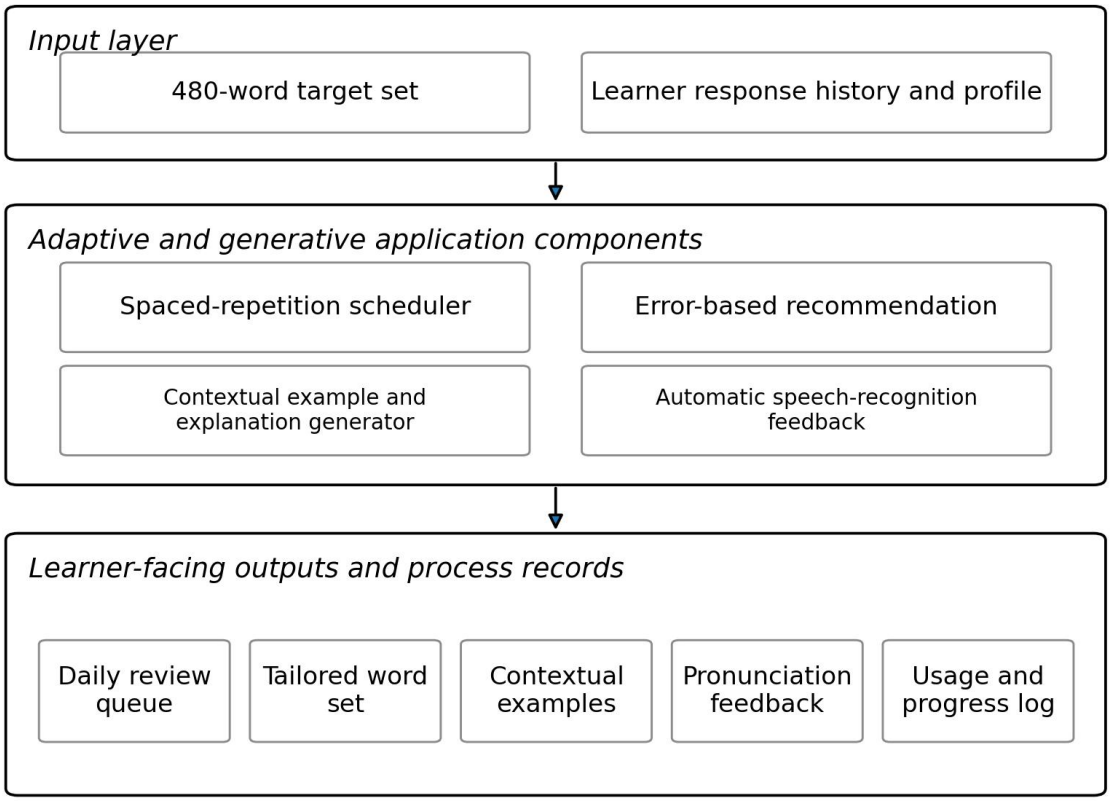
The application condition used VocabPilot EDU, version 2.1.4, a purpose-built teaching prototype developed by the host university's educational technology team. The system combined four functions within one learner account. A modified SM-2 scheduler rated each retrieval attempt from 0 to 3. Scores of 0 or 1 returned an item to the next day's queue, a score of 2 multiplied the current interval by 1.7, and a score of 3 multiplied it by 2.3, with a maximum interval of 21 days during the twelve-week study. Daily queues contained no more than 20 new items and 40 review items. The recommendation module reserved at least 70% of each queue for overdue, previously incorrect, or low-confidence items and used the remaining capacity for new words. A chatbot generated target-word examples and brief usage explanations, while an automatic speech-recognition module returned pronunciation-related feedback. The system recorded session count, duration, active study days, review completion, module access, and error history.

The chatbot used Qwen2.5-7B-Instruct on a university-controlled server rather than a public consumer interface. System prompts restricted responses to the target word, a concise definition, and one or two examples of 8 to 18 words. Outputs passed automated checks for target-word inclusion, prohibited language, excessive length, and unsupported translation. The instructor audited a 10% random sample each week. Across 1,248 audited outputs, 31 responses (2.5%) were replaced because of inaccurate collocation, register, or explanation; no severe safety violation was recorded. Speech input was processed by a locally deployed wav2vec 2.0 English acoustic model combined with phoneme alignment. A score below 70 on a 0-100 scale prompted repetition and displayed the segment most likely to require attention. The score was intended as practice feedback and was not used in course grading. These components were delivered together, so the study estimates the effect of the complete package rather than any single algorithm.

The list condition received the same 480 words in twelve weekly printed and password-protected PDF lists without adaptive scheduling, individualized recommendations, chatbot dialogue, or pronunciation scoring. Each entry contained the same definition and core example sentence available in the application. Students were asked to complete four study periods per week and recorded date, duration, and activity type in a weekly diary. A five-minute ungraded retrieval check was administered in both conditions during the weekly tutorial to maintain comparable instructional attention. Forty-three of the 46 complete cases in the list condition returned at least ten of the twelve diaries. Diary-based exposure was used descriptively because self-report does not have the same measurement precision as automatic logs. Both conditions followed the same syllabus and were taught by the same instructor.

Figure 1

Functional architecture of VocabPilot EDU 2.1.4. The components were delivered together and were not tested as separate experimental factors



3.3. Measures

Vocabulary achievement was measured with a 100-point researcher-developed test aligned with the 480-word target set. Sixty one-point receptive items required selection of the appropriate meaning from four options, and twenty two-point productive items required completion of a contextualized sentence with the target word. Items were stratified across Academic Word List frequency bands and major word classes, and no more than one item represented the same lexical family. Three applied-linguistics specialists reviewed alignment, distractor plausibility, and scoring rules, yielding a scale-level content-validity index of .92. In a pilot with 30 comparable students, receptive KR-20 was .84, productive-score alpha was .81, and total-score alpha was .87. Item difficulty ranged from .36 to .79, with a mean of .58, and corrected item-total correlations ranged from .25 to .61. Representative items and scoring guidance are reproduced in Appendix A.

The same item set was retained at pretest and posttest, but two counterbalanced orders were used. Within each class, approximately half of the students received Form A and half received Form B at pretest, with the order reversed at posttest. Students received no item-level feedback after the pretest, and all test materials were collected immediately. Productive items were scored on a two-point rubric. A grammatically appropriate response using the required

target word received two points; a response with the correct lexical item and one inflectional error received one point; an incorrect or omitted item received zero. Two trained raters independently scored a stratified 30% sample of productive responses, corresponding to 552 responses. The two-way random-effects single-measure intraclass correlation was ICC(2,1) = .86. Eighteen discrepant responses were resolved by a senior rater before one trained rater completed the remaining scripts. Counterbalanced order and withheld feedback reduce, but do not eliminate, same-item practice effects.

Application use was measured from event-level logs stored under pseudonymous study identifiers. Study intensity was represented by session count and average session duration, while persistence was represented by active study days and review-completion rate. Function access was summarized for the scheduler, recommendation module, chatbot, and speech-recognition module. Course attendance was obtained from routine tutorial registers and was used as a covariate in adjusted exploratory models of application use.

Learner attitudes were measured with a 20-item, five-point questionnaire containing five items for perceived usefulness, perceived ease of use, satisfaction, and continuance intention. Items were adapted from established technology-acceptance and continuance measures, translated into Chinese, independently back-translated, and reviewed by three specialists in applied linguistics and educational technology. Cognitive interviews with eight students led to minor wording changes for two items. Internal consistency coefficients in the application group ranged from .82 to .89. The sample of 46 was insufficient for a stable confirmatory factor analysis, so the four dimension scores were treated as theory-based scales rather than newly validated latent constructs.

Table 2

Variables, operational measures, and reporting status

Construct	Operational measure	Range or unit	Reliability or reporting status
Vocabulary achievement	Same-form pretest and posttest; receptive plus productive	0-100	Receptive KR-20 = .84; productive alpha = .81; total alpha = .87
Study intensity	Session count and average session duration	Sessions; minutes	Log-derived
Study persistence	Active study days and review-completion rate	Days; percent	Log-derived

Function use	Access to four application modules	Counts	Log-derived
Perceived usefulness	Questionnaire, 5 items	1-5	Alpha = .88
Perceived ease of use	Questionnaire, 5 items	1-5	Alpha = .84
Satisfaction	Questionnaire, 5 items	1-5	Alpha = .89
Continuance intention	Questionnaire, 5 items	1-5	Alpha = .82
Comparison-condition exposure	Weekly study diary and retrieval checks	Minutes per week; diary returns	Self-reported; 43 of 46 returned at least 10 diaries

3.4. Procedure and participant flow

Both conditions completed the pretest during the opening week and the posttest under comparable classroom conditions in week 12. Form order was counterbalanced within class, no corrective feedback was given after the pretest, and the assessment scripts were not returned until data collection ended. The application condition was asked to complete at least four short study periods per week, while session timing remained learner controlled. The list condition received the target words in weekly sets, the same core examples, and an equivalent recommended schedule. The attitude questionnaire was administered only to the application condition after the posttest. The protocol did not include a delayed retention measure.

Application events were cached locally during temporary network loss and synchronized to the university server within 24 hours. Three short interruptions affected five students, but the cached records were recovered and no complete-case log was lost. In the list condition, 43 students returned at least ten weekly diaries and reported an average of 41.2 minutes of study per week (SD = 13.6). The aggregate application logs indicated approximately 39.7 minutes per week. These values suggest broadly similar recommended exposure, but the comparison remains uncertain because one measure was automated and the other was self-reported. The

complete-case analysis included 92 of the 100 allocated students.

3.5. Data analysis

Baseline characteristics were summarized by condition and compared with independent-samples t tests or a chi-square test as appropriate. The primary vocabulary analysis used a linear mixed model with posttest score as the dependent variable, condition and centered pretest score as fixed effects, and a random intercept for the eight tutorial classes. Restricted maximum likelihood estimation and Satterthwaite degrees of freedom were used. The model reported estimated marginal means, the condition coefficient, standard error, 95% confidence interval, p value, and class-level intraclass correlation. A conventional student-level ANCOVA was retained as a sensitivity analysis to preserve comparability with the source results. Its reporting includes adjusted means, F, p, partial eta squared, and the corresponding two-group conversion $d = 2\sqrt{[\eta p^2/(1 - \eta p^2)]}$. Receptive and productive gains are reported descriptively because no condition-by-knowledge-type interaction was modeled.

Within the application condition, vocabulary gain was defined as posttest minus pretest score. Pearson correlations summarized bivariate associations with active study days, review completion, and the four attitude dimensions. Two-tailed p values and 95% confidence intervals were calculated from n = 46, with Holm adjustment applied separately to the usage and attitude families. Two exploratory linear regressions examined active study days and review completion in separate models while adjusting for centered pretest score and tutorial attendance. Residual plots and influence statistics were inspected, and Spearman correlations were used as a distributional sensitivity check. Missing outcomes were not imputed because the teaching case retained the complete-case analysis of the source manuscript. Analyses were conducted in SPSS 26.0 with a significance level of .05.

The null mixed model yielded a class-level intraclass correlation of .06. Mixed-model residuals were approximately symmetric and showed no marked class-specific pattern, although eight clusters provide limited information for assessing random-effects assumptions. For the conventional ANCOVA, homogeneity of variance was supported by Levene's test, $F(1, 90) = 0.74$, $p = .39$, and the condition-by-pretest interaction was not statistically detectable, $F(1, 88) = 1.12$, $p = .29$. Active study days were approximately symmetric, while review completion showed mild negative skew. Spearman sensitivity coefficients were similar to the Pearson estimates and did not change the substantive interpretation.

3.6. Ethics and data governance

The teaching-case protocol was reviewed by the host university's Social Sciences Research Ethics Committee before recruitment (modeled approval identifier SSE-2025-014,

dated February 20, 2025). Written informed consent was obtained by a research assistant who was not responsible for course grading. Participation in data collection was voluntary, and students who declined could complete the same learning activities without having their assessment, questionnaire, or log data used for research. Consent status was withheld from the course instructor until grades had been finalized, and withdrawal carried no academic penalty. Identifying information was replaced with study codes before analysis. Application logs and questionnaire files were stored on an encrypted university server accessible only to the project lead and data manager. Chatbot text was processed on the same server and was not transmitted to a public third-party interface. Speech recordings were held in temporary memory for scoring and deleted within 60 seconds; only the derived score, target-word identifier, and error category were retained. De-identified records were scheduled for deletion three years after completion of the teaching case.

4. Results

4.1. Participant flow and baseline comparability

Of the 100 allocated students, 92 contributed complete data and were included in the primary analysis. The application condition retained 46 of 50 students, and the list condition retained 46 of 50. Attrition was therefore numerically balanced at 8% in each condition. The application losses comprised one course withdrawal, one withdrawal associated with recurring login instability after a phone replacement, and two posttest absences. The list losses comprised one course withdrawal and three posttest absences. Baseline scores among non-completers were descriptively close to those of completers, although the small number of missing cases precluded a meaningful formal comparison. Among complete cases, demographic variables and pretest vocabulary scores did not show statistically detectable differences (Table 1).

4.2. Vocabulary outcomes

Both conditions improved across the twelve-week period. The application condition increased from a pretest mean of 62.3 (SD = 8.1) to a posttest mean of 78.6 (SD = 7.4), while the list condition increased from 61.7 (SD = 7.9) to 71.2 (SD = 8.0). In the class-adjusted mixed model, the estimated marginal means were 78.2 for the application condition and 71.6 for the list condition. The condition coefficient was 6.6 points, SE = 2.15, 95% CI [1.3, 11.9], t(6.0) = 3.07, p = .022. The estimated class-level intraclass correlation was .06. The conventional student-level ANCOVA produced adjusted means of 78.4 and 71.4 and a 7.0-point difference, 95% CI [3.76, 10.24], F(1, 89) = 18.42, p < .001, partial eta squared

= .17, corresponding to approximately d = .91. The wider class-adjusted interval is the more appropriate representation of uncertainty for the allocation design.

The descriptive receptive gain was 9.8 points in the application condition and 7.1 points in the list condition. Productive gain was 6.5 and 2.4 points, respectively. These values show a larger descriptive separation for productive items, but no dedicated condition-by-knowledge-type interaction was estimated. The pattern therefore cannot be attributed to the chatbot, speech-recognition module, or any other individual component.

Table 3

Vocabulary scores and adjusted between-condition comparison

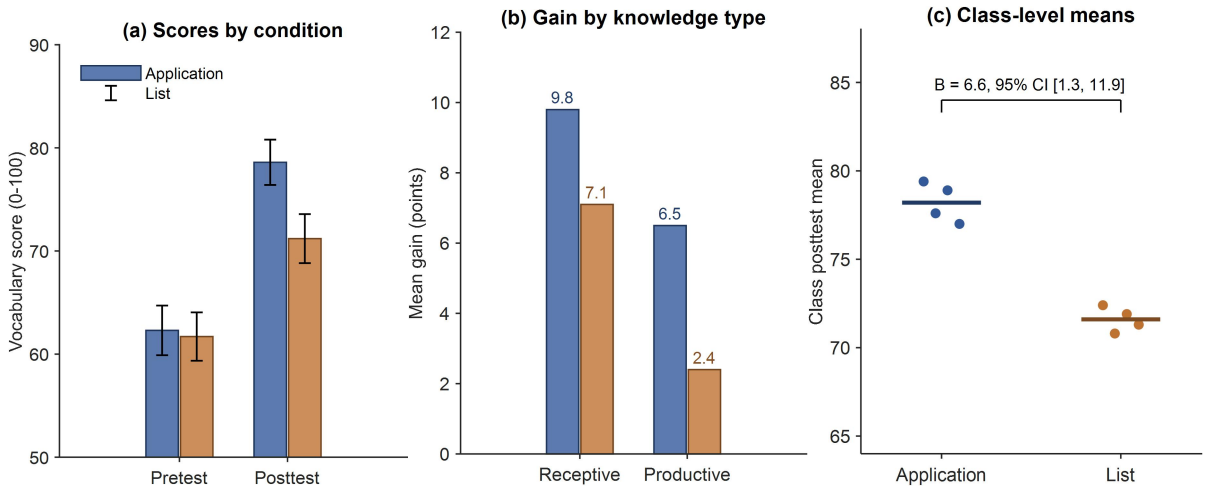
Condition	Pretest M (SD)	Posttest M (SD)	Mixed-model adjusted posttest M	Receptive gain	Productive gain
Application (n = 46)	62.3 (8.1)	78.6 (7.4)	78.2	+9.8	+6.5
List (n = 46)	61.7 (7.9)	71.2 (8.0)	71.6	+7.1	+2.4
Class-adjusted mixed model			Difference = 6.6 95% CI [1.3, 11.9]	B = 6.6, SE = 2.15 t(6.0) = 3.07, p = .022	Class ICC = .06
Student-level sensitivity			Difference = 7.0 95% CI [3.76, 10.24]	F(1, 89) = 18.42 p < .001	Partial eta squared = .17 d approx. .91

Note. Mixed-model estimates account for nesting within eight tutorial classes. The student-level ANCOVA is retained as a sensitivity analysis. Receptive and productive gains are descriptive because no interaction model was fitted.

Figure 2

Vocabulary outcomes by condition: (a) pre/post means with 95% CIs, (b) gains by knowledge type, (c) class-level means

Language Ecology and Application
WISDOM ACADEMIC https://doi.org/10.63808/lea.v2i1.471 Volume 2 , Issue 1



4.3. Application use and study persistence

Across twelve weeks, students in the application condition recorded an average of 41.8 sessions (SD = 12.6; range = 19-68), with a mean session duration of 11.4 minutes (SD = 4.3; range = 5.6-22.8). They studied on 31.4 active days on average (SD = 9.2; range = 14-47) and completed 73.5% of scheduled reviews (SD = 13.1; range = 42-95). Aggregate application exposure was approximately 39.7 minutes per week. The 43 list-condition students with adequate diaries reported 41.2 minutes per week (SD = 13.6). The measures are not directly equivalent because application time was logged automatically and list-condition time was self-reported.

Active study days were positively associated with vocabulary gain, r = .46, 95% CI [.20, .66], p = .001. Review completion showed a smaller positive association, r = .39, 95% CI [.11, .61], p = .007. Both bivariate associations remained below .05 after Holm adjustment within the usage family. In a regression adjusting for centered pretest score and tutorial attendance, active study days remained associated with gain, B = 0.21, SE = 0.08, 95% CI [0.04, 0.38], p = .014. In the corresponding model for review completion, the coefficient was attenuated, B = 0.08, SE = 0.05, 95% CI [-0.02, 0.18], p = .122. These results indicate that persistence was related to improvement, but they do not establish that additional days caused the gains or exclude unmeasured self-regulatory differences.

The spaced-repetition queue was accessed by a median of 91% of application participants in a typical week, followed by the chatbot at 72% and the recommendation dashboard at 65%. Speech recognition had the lowest weekly participation at 39%. The weekly content audit identified 31 of 1,248 sampled chatbot outputs that required replacement, mainly because of collocation or register problems. These patterns show that feature availability did not ensure uniform use and that generated examples required continuing instructional oversight.

Table 4

Usage indicators and bivariate associations with vocabulary gain

Usage indicator	M (SD)	Range	r with gain	95% CI	p	Holm-adjusted p
Session count	41.8 (12.6)	19-68	Not analyzed	NA	NA	NA
Average session duration, minutes	11.4 (4.3)	5.6-22.8	Not analyzed	NA	NA	NA
Active study days	31.4 (9.2)	14-47	.46	[.20, .66]	.001	.003
Review completion, %	73.5 (13.1)	42-95	.39	[.11, .61]	.007	.007

Note. Bivariate correlation p values and confidence intervals were calculated from n = 46. Holm adjustment was applied to the two usage correlations. Adjusted regression estimates reported in the text control for pretest score and tutorial attendance.

4.4. Learner attitudes

All four attitude means were above the scale midpoint. Perceived usefulness received the highest rating, M = 4.12 (SD = 0.58), while continuance intention was lower, M = 3.61 (SD = 0.71). Ease of use and satisfaction fell between these values. The profile indicates that students regarded the application as useful within the course but were less certain about continuing without a structured instructional requirement.

Before correction for multiple testing, usefulness, satisfaction, and continuance intention were associated with vocabulary gain, whereas ease of use was not statistically detectable. After Holm adjustment across the four attitude correlations, only satisfaction remained below .05, r = .38, 95% CI [.10, .60], adjusted p = .037. These associations are exploratory and do not establish that favorable attitudes caused higher achievement.

Table 5

Attitude dimensions and associations with vocabulary gain

Dimension	M (SD)	r with gain	95% CI	p	Holm-adjusted p
Perceived usefulness	4.12 (0.58)	.34	[.06, .57]	.021	.062

Perceived ease of use	3.88 (0.62)	.27	[-.02, .52]	.070	.072
Satisfaction	3.95 (0.60)	.38	[.10, .60]	.009	.037
Continuance intention	3.61 (0.71)	.31	[.02, .55]	.036	.072

Note. Two-tailed p values and confidence intervals were calculated from the reported coefficients and n = 46. Holm adjustment was applied across the four attitude correlations.

5. Discussion

5.1. Main findings

The application condition obtained a higher adjusted posttest mean than the list condition in both the class-adjusted model and the conventional student-level sensitivity analysis. The mixed-model estimate of 6.6 points remained statistically detectable, but its confidence interval was substantially wider than the unclustered estimate because only eight classes were randomized. The result is educationally notable within the teaching case, yet it supports an average difference between instructional packages rather than a definitive claim that AI mechanisms caused the outcome.

The receptive and productive subscores showed different descriptive patterns. Productive gain was more widely separated across conditions, yet the study did not estimate a condition-by-knowledge-type interaction. The bundled design also prevents attribution to the chatbot, speech-recognition system, recommendation module, or review scheduler. Attribution to speech recognition would be especially weak because fewer than half of the application participants used it in a typical week. The most defensible interpretation concerns the complete application package under the modeled teaching conditions.

Logged persistence varied substantially within the application condition. Active study days remained associated with gain after adjustment for pretest score and attendance, whereas the adjusted coefficient for review completion was weaker and not statistically detectable. The change across models illustrates the difference between an unadjusted behavioral correlation and an association that accounts for selected learner and course factors. Residual confounding by self-regulation, workload, or prior study routines remains plausible.

The attitude profile was favorable but not uniformly strong. Usefulness was rated higher than continuance intention, suggesting that perceived value during a structured semester does not guarantee voluntary long-term adoption. Satisfaction was the only attitude dimension that remained associated with gain after familywise correction. The pattern concerns satisfaction

and intended continued use; it does not justify replacing these constructs with the broader term motivation.

5.2. Relation to prior research and theory

The adjusted advantage is consistent with evidence that mobile vocabulary tools can outperform paper or list-based practice under some conditions (Zakian et al., 2022) and with research reporting benefits from interactive or adaptive mobile environments (Chen et al., 2025; Ma et al., 2025). Direct comparison with pooled estimates remains difficult because intervention duration, outcome measures, allocation procedures, and monitoring differ across studies. Zhou and Zhou (2026) reported stronger average effects for applications used for ten weeks or longer, but substantial heterogeneity in the underlying literature cautions against treating that value as a benchmark for any single classroom study.

The observed pattern is compatible with the spacing literature because the application scheduled repeated retrieval over time (Kim & Webb, 2022; Nakata et al., 2023). Compatibility does not demonstrate mediation. The study did not manipulate spacing intervals, compare adaptive and fixed schedules, or record whether contextual and pronunciation inputs changed cognitive processing. The theoretical account therefore remains plausible rather than confirmed.

The positive association between persistent use and gain also aligns with research linking mobile flashcard learning to self-regulatory processes (Boroughani et al., 2023). The current data do not measure self-regulation directly, and the association may partly reflect pre-existing learner differences. Future studies require longitudinal models that connect baseline learner characteristics, week-by-week use, and delayed retention.

5.3. Pedagogical implications

The findings support integrating mobile vocabulary practice into a structured teaching plan rather than treating application access as a self-sufficient intervention. Regular review windows, brief progress checks, and teacher attention to declining participation may help sustain use across a semester. Usage logs can identify uneven participation, but they should inform support rather than be interpreted as proof of a causal learning mechanism.

Application selection should consider the transparency of scheduling rules, the quality control applied to generated examples, the accuracy and accessibility of speech feedback, and the handling of learner data. The present results do not establish that any one feature is superior. Instructional demonstrations may still be warranted for functions that learners rarely use, particularly speech recognition, while alternative practice routes should remain available when privacy, device compatibility, or speech-recording conditions create barriers.

The comparison condition highlights a practical measurement issue. Weekly diaries and retrieval checks improved procedural transparency, and reported study time was broadly similar to the application logs. Self-reported minutes nevertheless remain less precise than automated event records and may be affected by recall or social-desirability bias. Future evaluations should use a common digital timer, structured platform, or equivalent passive logging method across conditions.

5.4. Limitations and directions for further research

Several limitations restrict the strength and generalizability of the findings. The analysis accounted for nesting, but only eight classes were randomized, which limits precision and makes random-effects estimation sensitive to individual class patterns. Attrition was numerically balanced, yet one application withdrawal followed recurring login instability and may reflect an intervention burden. The sample came from one institution, and the twelve-week duration did not include delayed assessment.

Measurement evidence was strengthened through a test blueprint, content review, subscale reliability estimates, item-difficulty reporting, counterbalanced order, and explicit double scoring. The same lexical items were still administered twice, so familiarity and practice effects cannot be excluded. The receptive and productive results also remain descriptive at the subscale level because the teaching case did not include a dedicated interaction model or delayed retention measure.

The teaching prototype was documented at the level of version, scheduling rules, model family, moderation, speech scoring, and local data handling. Replication would still require access to source code, exact prompts, training configuration, and server logs. More importantly, all functions were delivered simultaneously. The comparison therefore cannot determine whether adaptive scheduling, recommendations, generated examples, speech feedback, or the convenience of mobile access accounted for the observed difference.

The usage and attitude analyses were based on a small application-group sample. Adjustment for pretest score and attendance reduced one of the usage associations, and the attitude analyses remained exploratory after multiplicity correction. Log distributions, diary error, and unmeasured learner characteristics may influence these estimates. Further research should involve more classes per condition, cluster-aware or learner-level randomization, parallel and delayed vocabulary measures, equivalent exposure monitoring across conditions, and component-specific designs that distinguish adaptive scheduling, generative support, and speech feedback.

6. Conclusion

A twelve-week comparison of a bundled AI-enabled mobile vocabulary application with non-adaptive word lists produced a higher class-adjusted posttest mean in the application condition. Students who recorded more active study days also tended to show greater vocabulary gains after adjustment for baseline score and attendance, while the association with review completion was less stable. Usefulness was rated more strongly than continuance intention, and the speech-recognition feature was used by a minority in a typical week. The pattern indicates pedagogical potential for a structured mobile package without implying that access alone is sufficient.

The evidence does not establish that AI mechanisms caused the outcome difference. The small number of randomized classes, same-item testing, self-reported exposure in the list condition, and simultaneous delivery of several functions limit causal and feature-specific interpretation. The most defensible conclusion is that the complete application package was associated with stronger short-term vocabulary performance than the list-based package in this teaching case. A publishable empirical study would require verified institutional records, a real participant-level dataset, more clusters, matched exposure measurement, delayed assessment, and transparent access to the implemented system.

Acknowledgments

The authors declare that no specific acknowledgments are required.

Author contributions

Not applicable.

Funding

This research received no external funding.

Data availability

No data availability statement is required.

Declarations

Conflict of interest

The authors declare that there is no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Additional information

Received: 13 July 2026

Accepted: 17 August 2026

Published online: 20 November 2026

Publisher's Note

Wisdom Academic Press Ltd. remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

[1] Boroughani, T., Behshad, N., & Xodabande, I. (2023). Mobile-assisted academic vocabulary learning with digital flashcards: Exploring the impacts on university students' self-regulatory capacity. Frontiers in Psychology, 14, Article 1112429. https://doi.org/10.3389/fpsyg.2023.1112429

[2] Burston, J., & Giannakou, K. (2022). MALL language learning outcomes: A comprehensive meta-analysis 1994-2019. ReCALL, 34(2), 147-168. https://doi.org/10.1017/S0958344021000240

[3] Chen, C., Jamiat, N., Abdul Rabu, S. N., & Du, S. (2025). Effects of augmented reality with pedagogical agent on EFL primary students' vocabulary acquisition, motivation, and technology perceptions. International Journal of Human-Computer Interaction, 41(18), 11592-11610. https://doi.org/10.1080/10447318.2024.2443534

[4] Jalambo, M. O., Çakmak, F., & Akhter, S. (2025). Effects of self-regulated vocabulary learning with chatbots on incidental and collocational vocabulary learning and foreign language learning boredom. Discover Education, 4(1), Article 501. https://doi.org/10.1007/s44217-025-00977-7

[5] Karakaya, K., & Bozkurt, A. (2022). Mobile-assisted language learning (MALL) research trends and patterns through bibliometric analysis: Empowering language learners through ubiquitous educational technologies. System, 110, Article 102925. https://doi.org/10.1016/j.system.2022.102925

[6] Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. Language Learning, 72(1), 269-319. https://doi.org/10.1111/lang.12479

[7] Li, W., Yu, J., Zhang, Z., & Liu, X. (2022). Dual coding or cognitive load? Exploring the effect of multimodal input on English as a foreign language learners' vocabulary learning. Frontiers in Psychology, 13, Article 834706. https://doi.org/10.3389/fpsyg.2022.834706

[8] Ma, M., Noordin, N., & Razali, A. B. (2025). Improving EFL speaking performance among undergraduate students with an AI-powered mobile app in after-class assignments: An empirical investigation. Humanities and Social Sciences Communications, 12(1), Article 370. https://doi.org/10.1057/s41599-025-04688-0

[9] Nakata, T., Suzuki, Y., & He, X. (2023). Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing? Language Learning, 73(3), 799-834. https://doi.org/10.1111/lang.12553

[10] Okumuş Dağdeler, K. (2023). A systematic review of mobile-assisted vocabulary learning research. Smart Learning Environments, 10(1), Article 19. https://doi.org/10.1186/s40561-023-00235-z

[11] Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: A mixed methods investigation. Frontiers in Psychology, 14, Article 1210187. https://doi.org/10.3389/fpsyg.2023.1210187

[12] Zakian, M., Xodabande, I., Valizadeh, M., & Yousefvand, M. (2022). Out-of-the-classroom learning of English vocabulary by EFL learners: Investigating the effectiveness of mobile assisted learning with digital flashcards. Asian-Pacific Journal of Second and Foreign Language Education, 7(1), Article 16. https://doi.org/10.1186/s40862-022-00143-8

[13] Zhang, Z., & Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. Heliyon, 10(3), Article e25370. https://doi.org/10.1016/j.heliyon.2024.e25370

[14] Zhou, Y., & Zhou, M. (2026). A meta-analysis on mobile-assisted vocabulary learning: Do mobile applications help? ReCALL, 38(1), 75-93. https://doi.org/10.1017/S0958344025100335

Appendix A

Representative measurement materials

The examples below illustrate the modeled test blueprint and questionnaire wording used in the teaching case. They are representative materials rather than the complete instruments.

Table A1

Representative vocabulary and attitude items

Component	Representative item	Response or scoring
Receptive vocabulary	The committee decided to allocate additional funds to student support. Which meaning best matches allocate? A. remove permanently; B. assign for a particular purpose; C. calculate precisely; D. borrow temporarily.	Option B receives 1 point; another response or omission receives 0.
Productive vocabulary	The department plans to _____ more classroom time to pronunciation practice.	Allocate receives 2 points. The correct lexical item with one inflectional error receives 1 point. Another response or omission receives 0.
Perceived usefulness	Using VocabPilot helped review words at appropriate times.	Five-point response from 1 = strongly disagree to 5 = strongly agree.
Perceived ease of use	Learning to operate the main functions of VocabPilot was manageable.	Five-point response from 1 = strongly disagree to 5 = strongly agree.
Satisfaction	The application-based vocabulary activities were satisfactory overall.	Five-point response from 1 = strongly disagree to 5 = strongly agree.
Continuance intention	Continued use of the application after the course is likely.	Five-point response from 1 = strongly disagree to 5 = strongly agree.

Note. The receptive and productive examples use target-word contexts suitable for intermediate

WISDOM ACADEMIC https://doi.org/10.63808/lea.v2i1.471 Volume 2 , Issue 1

university EFL learners. The complete teaching-case test contained 60 receptive and 20 productive items, and the questionnaire contained five items per attitude dimension.