ReCALL (2026), 1–18
doi:10.1017/S0958344026100615

RESEARCH ARTICLE

AI-assisted apps versus retrieval-based digital flashcards: Effects on EFL learners' receptive vocabulary development

Derya Tuzcu-Eken
School of Foreign Languages, Kırklareli University, Türkiye (derya.tuzcu@klu.edu.tr)

Abstract

Retrieval-based digital flashcards have proved their effectiveness in L2 vocabulary learning. However, recent developments in AI technology have introduced AI-assisted vocabulary learning apps that integrate multimedia input and AI-supported output activities. Considering these technological advances, this study compared the effects of an AI-assisted multimedia app (WordUp) and a retrieval-based flashcard app (Quizlet) on university students' receptive vocabulary. A total of 128 students from three EFL classes were assigned to two experimental groups and one control group. Each experimental group used its assigned app to study target words. The control group did not receive any intervention. Students' learning gains were measured with a vocabulary test administered as an immediate and a delayed posttest. Semi-structured interviews with 20 participants provided insights into learners' experiences. The findings indicated that both WordUp and Quizlet notably enhanced receptive vocabulary knowledge. However, the WordUp group demonstrated significantly higher vocabulary learning gains than Quizlet in the immediate and delayed posttests. The control group showed negligible improvement. Interviews revealed that learners valued Quizlet's simple interface and structured review but also expressed a need for greater interactivity and multimedia support. In contrast, WordUp was valued for its contextualized multimedia resources and AI-assisted productive activities, though learners reported challenges related to interface and content complexity. Overall, the findings suggest that effective mobile-assisted vocabulary learning depends not only on the technological features available in learning apps but also on balancing rich multimedia and interactive learning opportunities with usability and learners' cognitive readiness.

Keywords: AI-assisted apps; digital flashcards; mobile-assisted vocabulary learning; multimedia; receptive vocabulary; retrieval-based practice

1. Introduction

Vocabulary knowledge is central to linguistic ability and communication in foreign language (FL) learning (Webb & Nation, 2017). To comprehend spoken input, learners typically need knowledge of around 6,000–7,000 word families, which increases to 8,000–9,000 word families for written input (Nation, 2006). Given the limited classroom time available for vocabulary instruction (Tuzcu, 2023), out-of-class learning has become essential for lexical development (Nation, 2022). In this context, mobile-assisted vocabulary learning (MAVL) apps have emerged as practical solutions to support vocabulary learning beyond the classroom (Lin & Lin, 2019).

Email: derya.tuzcu@klu.edu.tr

Cite this article: Tuzcu-Eken, D. (2026). AI-assisted apps versus retrieval-based digital flashcards: Effects on EFL learners' receptive vocabulary development. *ReCALL* FirstView, 1–18. https://doi.org/10.1017/S0958344026100615

© The Author(s), 2026. Published by Cambridge University Press on behalf of EUROCALL, the European Association for Computer-Assisted Language Learning. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

Within MAVL, retrieval-based digital flashcard apps have received considerable attention. These apps commonly rely on paired-associate formats and repeated retrieval practice (Karpicke, 2012; Nakata, 2011). A growing body of research shows that they outperform traditional paper-based methods (e.g., Li & Hafner, 2022; Rahmani et al., 2022; Tuzcu-Eken & Tuzcu, 2026), making retrieval-based flashcards a dominant approach in mobile vocabulary learning.

On the other hand, research grounded in the cognitive theory of multimedia learning (CTML; Mayer, 2009) suggests that multimodal input facilitates more meaningful vocabulary learning. According to this framework, learning is enhanced when verbal and visual information are processed through dual channels and integrated into coherent mental representations (Mayer, 2009; Mayer & Moreno, 2003). In this sense, inputs such as images, audio, and video can support richer lexical processing than isolated word presentations (Chun & Plass, 1996).

Recent advances in AI and mobile technologies have further expanded these possibilities by introducing multimodal and interactive features such as guided writing and conversational practice (Anantrasirichai & Bull, 2022; Leong et al., 2024). These features may support more contextualized and generative engagement with vocabulary, aligning with generative learning theory (Fiorella & Mayer, 2016). However, they may also increase cognitive load or reduce retrieval effort due to extensive scaffolding, potentially weakening the benefits of effortful retrieval (Lin & Atkinson, 2011; Mayer, 2009).

Although interest in AI-assisted vocabulary learning has grown rapidly, the empirical evidence remains limited and inconclusive. Recent studies have explored a range of AI-assisted tools, including vocabulary applications, intelligent personal assistants, and conversational AI, with generally promising but inconsistent findings (e.g., Abdelhalim & Alsehibany, 2025; Alrajhi, 2025; Polyzi & Moussiades, 2023). While several studies have reported improvements in vocabulary learning through AI-assisted tools (e.g., Abdelhalim & Alsehibany, 2025; Iravi & Malmir, 2023), others have highlighted both the potential and the limitations of multimodal and generative AI features (e.g., Chen et al., 2024) or reported mixed outcomes for vocabulary retention and productive vocabulary use (e.g., Peng et al., 2023). Consequently, it remains unclear how different instructional features embedded in AI-assisted and retrieval-based vocabulary apps influence vocabulary learning.

To address this gap, the present study adopts a theory-driven experimental design comparing an AI-assisted app (WordUp), a retrieval-based flashcard app (Quizlet), and a no-treatment control group. Drawing on retrieval-based learning theory and CTML, while also considering AI-assisted interactive practice, the study examines whether multimodal enrichment and AI-supported interactive features lead to greater immediate and delayed receptive vocabulary gains than learning with retrieval-based digital flashcard apps.

2. Literature review

2.1 Retrieval-based vocabulary learning

Retrieval-based learning refers to the active recall of previously studied information from memory, which strengthens long-term retention more effectively than repeated study alone (Karpicke, 2012; Roediger & Karpicke, 2006). Digital flashcards are one of the most common applications of retrieval-based learning in mobile-assisted language learning. These apps use paired-associate formats in which learners repeatedly retrieve word meanings, definitions, or translations (Kornell & Vaughn, 2016; Nakata, 2011), often supported by spaced repetition algorithms designed to optimize long-term retention (Cepeda et al., 2006).

Empirical studies consistently support their effectiveness. Rahmani et al. (2022) reported significant improvements in learners' high-frequency vocabulary knowledge when digital flashcards were integrated into instruction. Zakian et al. (2022) found that mobile flashcards outperformed traditional word lists, despite some decline in long-term retention. Similarly, Li and

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

Hafner (2022) showed that digital flashcards led to greater receptive and productive vocabulary gains than paper-based formats.

Beyond vocabulary learning outcomes, research has also examined learners' experiences with retrieval-based digital flashcard apps. Zung et al. (2022) found that although university students widely used digital flashcards for autonomous learning, they did not always employ them in ways that fully reflected evidence-based retrieval practices. Yaşar and Koçoğlu (2025) reported that learners generally perceived learner-created digital flashcards as practical, accessible, and engaging, while appreciating their flexibility, personalization, and support for autonomous learning. Their study also highlighted the role of metacognitive strategies and emotional regulation in vocabulary learning. Together, these findings suggest that retrieval-based digital flashcard apps influence not only vocabulary learning outcomes but also learners' experiences and self-regulated learning processes, providing an important rationale for examining learners' perceptions alongside vocabulary gains.

Overall, while research highlights the value of retrieval practice for vocabulary learning, most studies have compared retrieval-based digital flashcards with paper-based vocabulary learning materials or classroom vocabulary instruction. Less is known about how retrieval-based learning compares with AI-assisted vocabulary environments in terms of promoting both immediate and delayed gains, as well as supporting learners' experiences, which indicates a need for further investigation.

2.2 Multimedia vocabulary learning

An alternative line of research relevant to vocabulary learning is grounded in CTML (Mayer, 2009). CTML proposes that learning is enhanced when information is processed through dual channels (visual and auditory) and actively integrated into coherent mental representations, based on limited working memory capacity and active cognitive processing (Mayer & Moreno, 2003). In vocabulary learning, multimedia inputs such as images, audio pronunciation, and contextualized video can support the integration of form and meaning by engaging both verbal and visual channels. By situating lexical items in meaningful contexts rather than presenting them as isolated pairs, such input may enrich lexical representations and strengthen memory traces through multiple retrieval pathways (Chun & Plass, 1996; Mayer, 1997).

However, multimedia integration does not always produce positive learning outcomes. Cognitive load theory suggests that poorly designed multimedia environments may impose extraneous load that hinders learning (Mayer, 2009; Sweller et al., 2019). Complex visuals or misaligned audiovisual elements can increase visual search demands and reduce resources available for vocabulary processing (Lin & Atkinson, 2011). In addition, the expertise reversal effect indicates that instructional support beneficial for novices may become unnecessary or even detrimental for more advanced learners (Kalyuga, 2021).

2.3 AI-assisted vocabulary learning

Recent advances in AI have enabled the development of multimodal and interactive vocabulary learning environments. AI-assisted apps can generate contextualized content, adapt learning materials, and support interactive tasks based on learner performance and preferences (Leong et al., 2024).

Emerging empirical studies provide promising but mixed evidence regarding the effectiveness of AI-assisted vocabulary tools and applications. Iravi and Malmir (2023) reported that learners using WordUp achieved greater vocabulary gains than those using other digital vocabulary learning applications. Similarly, Polyzi and Moussiades (2023) reported that learners using an AI-based vocabulary learning assistant incorporating chatbot interaction, games, and quizzes outperformed those learning the same vocabulary through textbook-based instruction. Beyond

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

vocabulary applications, Alrajhi (2025) demonstrated that learners using Google Assistant as an intelligent personal assistant achieved significant gains in vocabulary knowledge and reported positive perceptions of its usefulness for self-access vocabulary learning. Likewise, Abdelhalim and Alsehibany (2025) found that integrating ChatGPT into classroom vocabulary instruction significantly enhanced learners' productive vocabulary knowledge and overall vocabulary performance. Studies on prototype systems further highlight both the potential and limitations of AI-assisted vocabulary learning. For example, Chen et al. (2024) showed that RetAssist supported lexical usage through generative images in story-retelling tasks. In contrast, Storyfier, which generates contextual stories containing target vocabulary, produced mixed results: although learners reported positive perceptions, they performed worse in recalling and using target words than those using a non-AI baseline app (Peng et al., 2023). Taken together, these findings suggest that AI-assisted tools can support vocabulary learning through multimodal, adaptive, and generative features; however, their effects on vocabulary retention and learning outcomes remain inconsistent.

Overall, existing research consistently supports the effectiveness of retrieval-based digital flashcards, whereas evidence regarding AI-assisted multimedia apps remains promising but limited. Few experimental studies have directly compared AI-assisted multimodal vocabulary apps with retrieval-based digital flashcard systems while examining the relative contribution of multimodal enrichment and AI-assisted interaction to receptive vocabulary learning and retention. Consequently, a critical gap emerges. Addressing this gap is important for understanding how these instructional features influence vocabulary outcomes. The present study therefore adopts a theory-driven approach to examine the relative effectiveness of these two approaches in promoting receptive vocabulary retention while also exploring learners' perceptions of learning with each app. Accordingly, the research questions of the study are as follows:

1. Do learners using an AI-assisted multimedia vocabulary app achieve greater immediate and delayed receptive vocabulary gains than those using a retrieval-based flashcard app or no-treatment condition?
2. How do learners perceive the learning experience and cognitive engagement afforded by each app?

3. Method

This study employed a quasi-experimental mixed-methods design. Three EFL classes at a university in Türkiye were randomly assigned to three groups: two experimental and one control. To minimize cross-group contamination and social influence, each class was deliberately selected from different campuses.

3.1 Participants

Participants were 128 freshmen EFL students with a mean age of 19.85. Using their assigned classes as a guide, the students were divided into three groups: two experimental groups and one control group. The experimental groups were named after their assigned mobile apps: WordUp and Quizlet. There were 50 students in the WordUp group, 40 students in the Quizlet group, and 38 in the control group.

At the time of data collection, all three groups were enrolled in a weekly two-hour A2-level General English course taught by the author. To minimize potential biases associated with this dual role, two precautions were implemented. First, students were explicitly informed that their participation or non-participation would not influence their grades. Second, although the author was the course instructor, all data were collected by the author outside the regular teaching

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

sessions to minimize the potential influence of the instructor–student relationship on participants' responses.

To account for differences in participants' educational backgrounds and establish a baseline for their vocabulary knowledge, the Updated Vocabulary Level Test (UVLT), developed by Webb et al. (2017), was administered to all three groups. The test results provided insight into the students' existing vocabulary levels and helped identify the words that still needed to be learned. Accordingly, the majority of students were found to be at Level 0 or Level 1, with only eight students demonstrating Level 2 proficiency. Two of the Level 2 students were eliminated from the dataset, as they had already mastered over 80% of the target vocabulary. Moreover, a one-way ANOVA was conducted to compare the UVLT scores among the three groups. The results indicated that there was no statistically significant difference in the UVLT scores between the groups, $F(2, 125) = 0.156$, $p = 0.856$.

In addition, 20 students from the experimental groups (10 WordUp and 10 Quizlet) participated in interviews. These participants were selected on a voluntary basis.

3.2 Target words

The target words were selected from the BNC/COCA headwords of the 2nd and 3rd thousand, based on Nation's (2012) classification. Each BNC/COCA headword list comprises 1,000 words. To ensure a balanced representation of word forms, I followed the principle outlined by Webb et al. (2017), who based their UVLT on the occurrence percentages of word forms in English: nouns (50%), verbs (30%), and adjectives (20%).

A total of 100 words from the BNC/COCA headwords of the 2nd and 3rd thousand were randomly selected while adhering to these percentages: 50 nouns, 30 verbs, and 20 adjectives. These 100 words were then given to the students, who were instructed to provide the Turkish equivalents, allowing an assessment of their prior knowledge. The students' knowledge of each word was checked against a key containing all possible meanings of the words. Students received 1 point for each target word if they correctly provided its Turkish equivalent and 0 points for incorrect or missing responses.

The filtering process followed two main criteria to ensure the selection of unfamiliar words: (1) each word was known by no more than 10 students, and (2) the selected words represented a diverse range of word forms. The threshold of 10 students was set to ensure that no more than approximately one-quarter of participants in any group recognized a given word. A higher threshold would have included more commonly known words, whereas a lower threshold would have overly restricted the word pool.

The final list consisted of 40 target words that met these criteria. However, due to the higher accuracy rates for adjectives and nouns, the final word form distribution was adjusted: 25 verbs (62.5%), 11 nouns (27.5%), and 4 adjectives (10%). The complete list of the target words, the target word selection procedure, and the full scoring table are provided in the supplementary material (Appendices A, B, and C).

3.3 Learning apps

The vocabulary learning apps utilized in the learning sessions were Quizlet and WordUp. The basic version of Quizlet and the trial version of WordUp were used. Each group practiced the target words as an out-of-class activity, except for the control group.

Quizlet was utilized as a retrieval-based digital flashcard app. I created flashcards containing each target word, its English explanation, and Turkish translation, while pronunciation was provided by the app. Although the platform supports images, none were included. Students first reviewed each card by viewing the word and recalling its meaning from the reverse side. They then practiced the set using the Learn, Test, and Match features. The Learn mode checks knowledge of

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

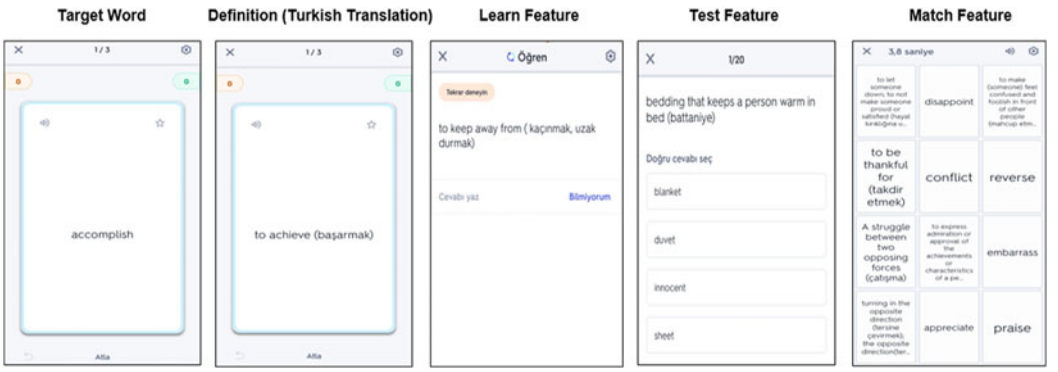


Figure 1. Example screenshots demonstrating different features of the Quizlet app.

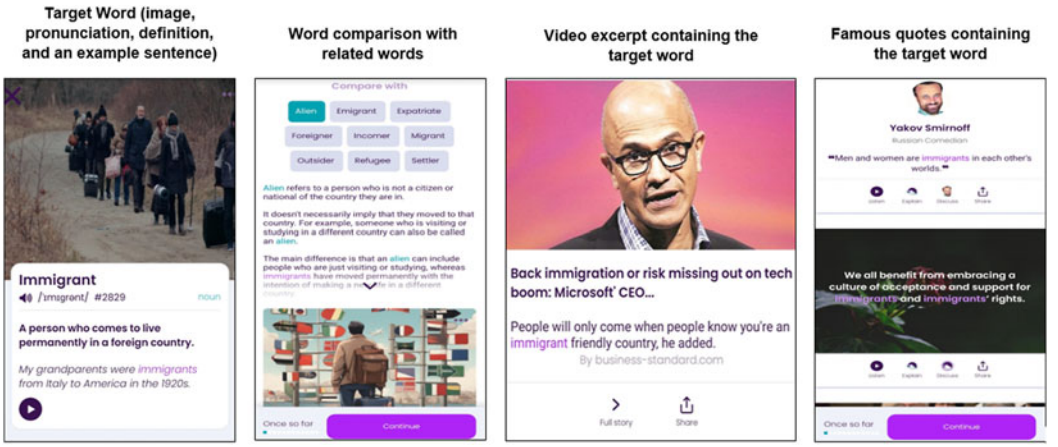


Figure 2. Example screenshots demonstrating multimedia features of the WordUp app.

individual items, the Test mode includes multiple-choice questions to support active recall, and the Match mode offers a timed activity to strengthen word-meaning associations through rapid recognition. Figure 1 presents the main features of the Quizlet app.

To track students' engagement with the app, a Quizlet Plus teacher account was used to monitor completion of the learning activities. However, the platform did not provide information on the amount of time students spent using the app. Therefore, prior to the study, students were instructed to record their daily study time. At the immediate posttest, participants in the Quizlet group reported the time they had spent studying the target words each day. These self-reports were cross-checked against the recorded completion of the activities in Quizlet. Reported study time ranged from 15 to 25 minutes per day.

WordUp, on the other hand, integrates both multimedia and AI-assisted features. Its multimedia features include images, example sentences, animated visuals, and authentic video excerpts, which are part of the app's existing instructional content. Figure 2 presents the multimedia features.

The AI-assisted features of the app include interactive writing and speaking activities, where students engage with the AI. Through these activities, students can either initiate a conversation or

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

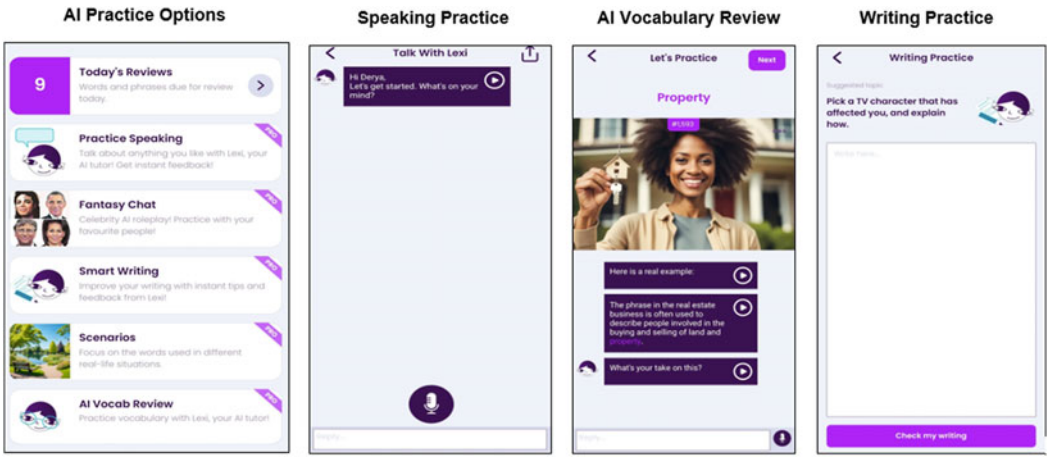


Figure 3. Example screenshots demonstrating AI-assisted practice features of the WordUp app.

allow the AI to guide the dialogue based on the information it has gathered about them, such as their interests indicated during the app registration process. Figure 3 presents the AI-assisted practice features.

The WordUp group used both features of the app. After enrolling in an in-app class, the system enabled tracking of students' engagement in terms of daily and weekly study time. Based on this feature, students' app usage time was monitored. Reported daily study time ranged from 20 to 25 minutes per student.

3.4 Instruments

3.4.1 Vocabulary test

To assess the effectiveness of the vocabulary learning sessions, a meaning recognition test based on the 40 target words was developed. The test, administered as both an immediate and delayed posttest, consisted of two sections: gap-fill and multiple-choice. Internal consistency was assessed using Cronbach's alpha, which indicated high reliability for both the immediate ($\alpha = .94$) and delayed posttests ($\alpha = .92$).

In the gap-fill section, students completed two separate exercises, each consisting of 10 gapped sentences and 10 different target words to fill in the gaps. The words were chosen randomly from the target word list. The multiple-choice section included 20 items, each with four options. All sentences were carefully designed to avoid vocabulary beyond the learners' proficiency level. The full test is provided in the supplementary material (Appendix D).

3.4.2 Interview questions

A semi-structured interview was developed and conducted with 20 students to extend the quantitative results and gain a deeper understanding of participants' learning experiences. The interview questions focused on the process the students underwent while studying the words from the apps. The questions addressed the app's ease of use, the challenges students faced when learning new words, and the features of the app they found appealing. The interview questions are given in the supplementary material (Appendix E).

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

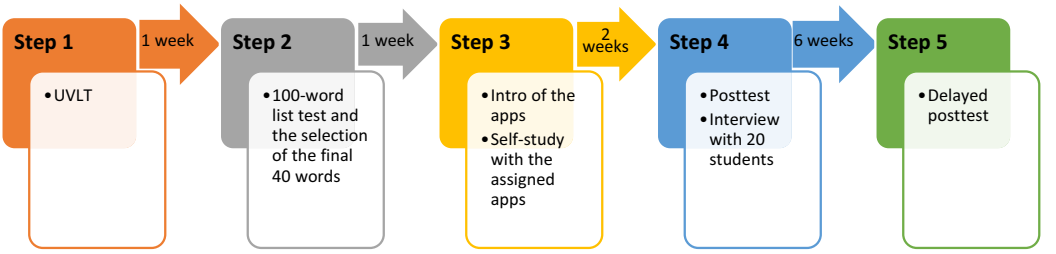


Figure 4. Data collection process.

3.5 Data collection

To collect data for the study, students first completed a consent form. The UVLT was then administered to all groups prior to the learning sessions to estimate their vocabulary levels and identify the words they still needed to learn. The test took approximately 30 minutes to complete.

One week after the UVLT, 100 randomly selected words from the 2nd and 3rd thousand BNC/COCA headwords were administered to the students. Participants were asked to provide the Turkish equivalents of the words they knew and leave unfamiliar ones blank. The task took approximately 20 minutes to complete.

After finalizing the list of 40 words, the experimental phase began. The WordUp group selected the target words from the available list and added them to their "to be learned" section. An in-app class was also created, and students were invited to join to track their progress. For the Quizlet group, an in-app class and flashcards for the 40 words were created, and students accessed the materials through the class platform. Both groups initially used the apps under guidance until they became familiar with the system.

Once ready, participants were instructed to study the words independently for 15 to 20 minutes daily over two weeks. Progress was monitored through in-app data. Since Quizlet does not record individual study duration (only activity type), students in this group self-reported their daily study time, which was later collected in the immediate posttest. The control group continued regular classroom instruction without explicit focus on the target words and only participated in the paper-based vocabulary tests alongside the experimental groups.

After the two-week learning period, all groups completed the vocabulary test as an immediate posttest within 40 minutes. Additionally, 20 students (10 WordUp and 10 Quizlet) participated in an interview with the researcher to gain further insights into their learning experiences. The interviews were conducted in Turkish to avoid language barriers. Each interview lasted 10 to 15 minutes; interviews were audio-recorded and transcribed for analysis.

Six weeks later, all groups took the same vocabulary test again as a delayed posttest to examine retention of the target words. They were given 40 minutes to complete it. The overall procedure is illustrated in Figure 4.

3.6 Data analysis

To analyze the data of the study, mixed-effects modeling was performed using R Version 4.0.2 (R Core Team, 2020). The outcome variable in the analysis was binary (1 for correct responses and 0 for incorrect responses), so I conducted a mixed-logistic regression (Jaeger, 2008) using the *glmer* function in the lme4 package (Bates et al., 2015). The primary predictors in the model were group (control, Quizlet, and WordUp), test time (posttest vs. delayed posttest), and their interactions. Additionally, question type (gap-fill vs. multiple-choice) on the posttests, students' pretest scores, their UVLT scores, and the word frequency level of the target items were added to the model as covariates. Item and student random intercepts were also incorporated into the

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

Table 1. Performance of the three groups on the immediate and delayed posttests (Max = 20)

	Gap-fill						Multiple-choice					
	Immediate posttest			Delayed posttest			Immediate posttest			Delayed posttest		
Group	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Control	3.50	2.24	[2.69, 4.31]	3.53	2.41	[2.77, 4.28]	3.86	2.83	[2.90, 4.82]	3.83	2.81	[2.88, 4.79]
Quizlet	9.90	5.27	[8.21, 11.60]	7.50	5.19	[5.84, 9.16]	10.07	4.84	[8.53, 11.60]	8.97	4.47	[7.56, 10.40]
WordUp	11.72	5.45	[10.20, 13.20]	9.86	4.75	[8.51, 11.20]	11.44	4.49	[10.20, 12.70]	9.30	4.48	[8.03, 10.60]



Figure 5. Interview coding process.

models. Step-by-step backward selection of random slopes was used in the model selection process. Models were compared using Akaike Information Criterion (AIC) values, and the final model was the most parsimonious one with the lowest AIC values. Model diagnostics were also conducted on the final model. Multicollinearity was examined using variance inflation factors (VIFs) computed from a fixed-effects logistic regression, and residual diagnostics were performed using the DHARMa package (Version 0.4.7; Hartig, 2024). I also calculated the odds ratios and effect sizes (Cohen's *d*) for each model's coefficient estimate to improve the interpretability of the results. The standardized effects sizes were calculated by dividing the log odds ratios by 1.81, following Chinn (2000). Finally, *emmeans* function in the emmeans package (Version 2.0.5; Lenth & Piaskowski, 2026) was used to conduct pairwise comparisons to explore the interactions between group and test time. The data and R codes are publicly available on OSF at https://osf.io/j8eup/overview?view_only=0a0d2eec150f4cf29f2badff9a2fd525

The interview data were analyzed using conventional content analysis. Figure 5 illustrates the coding process. First, the transcripts were read carefully, and initial notes were taken to identify potential categories. Two main categories were then developed, each comprising several themes (see Table 4). Clear definitions of categories and themes were established to ensure coding consistency.

The data were then coded systematically under the relevant categories and themes by the author. To ensure intra-rater reliability (Cohen, 2017), the author repeated the coding process after a three-month interval by reexamining and recoding all transcripts. All transcripts were anonymized to ensure participant confidentiality in line with ethical guidelines. Reliability was assessed using Cohen's kappa, which indicated high agreement for both groups (Quizlet: $\kappa = 0.93$; WordUp: $\kappa = 0.91$; Brown et al., 2004).

4. Results

4.1 Descriptive analyses

The mean prior knowledge of the target words was 17% for the control group ($SD = 0.37$, 95% CI [0.15, 0.19]), whereas the mean prior knowledge for the Quizlet and WordUp groups was 11% ($SD = 0.31$, 95% CI [0.09, 0.12]) and 7% ($SD = 0.26$, 95% CI [0.06, 0.08]), respectively. As reflected

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

Table 2. The mixed logit model results

Predictors	*b*	*OR*	*SE*	95% CI for *OR*	*z*	*p*
Intercept	−2.61	0.07	0.02	[0.05, 0.11]	−11.42	<.001
Group (Quizlet)	2.24	9.38	2.02	[6.16, 14.30]	10.41	<.001
Group (WordUp)	2.75	15.62	3.22	[10.43, 23.39]	13.34	<.001
Test time (Delayed)	−0.00	1.00	0.11	[0.80, 1.25]	0.00	1.000
Type (Multiple-choice)	0.04	1.04	0.23	[0.67, 1.62]	0.17	.869
UVLT scores	1.27	3.56	0.62	[2.53, 5.00]	7.33	<.001
Pretest scores	2.00	7.39	0.73	[6.09, 8.97]	20.21	<.001
Group (Quizlet) * Test (Delayed)	−0.50	0.61	0.09	[0.46, 0.80]	−3.51	<.001
Group (WordUp) * Test (Delayed)	−0.55	0.58	0.08	[0.44, 0.75]	−4.06	<.001
Random effects	Variance	*SD*				
1|Participant	0.66	0.81				
1|Item	0.49	0.70				
$N_{Participant}$	126					
N_{Number}	40					
Observations	10080					
Marginal R^2/Conditional R^2				0.296/0.478		
Model:	Accuracy ~ Group * Test + Question_Type + UVLT + Pretest + (1 | Participant) + (1 | Item)					

Note. CI = confidence interval; UVLT = Updated Vocabulary Level Test.

Table 3. Pairwise comparisons of three groups with Tukey adjustment $p < .05$

	Immediate posttest					Delayed posttest				
Contrasts	*b*	*SE*	*z*	*p*	*d*	*b*	*SE*	*z*	*p*	*d*
Control – Quizlet	−2.24	0.22	−10.41	<.001	1.24	−1.74	0.22	−8.11	<.001	0.96
Control – WordUp	−2.75	0.21	−13.34	<.001	1.52	−2.20	0.21	−10.70	<.001	1.22
Quizlet – WordUp	−0.51	0.19	−2.67	.021	0.28	−0.46	0.19	−2.38	.045	0.25

Table 4. Themes identified from learners' interview responses

App	Favorable aspects	*n*	Areas for enhancement	*n*
Quizlet	Flashcard repetition	7	Lack of videos	4
	User-friendliness	8	Lack of visuals	8
	Test and practice feature	6	Lack of gamified features	7
WordUp	AI-assisted skills	7	Complex interface	7
	Video support	5	Complex example sentences	5
	Visual support	10	Complex language in videos	7
	Contextualized vocabulary use	6		

in the 95% confidence intervals (CIs), initially, all three groups differed from each other in their mean prior knowledge of the target words, with the control group showing a higher level of prior knowledge (6.8 words) compared to the Quizlet (4.4 words) and WordUp (2.8 words) groups. However, since no treatment was applied to the control group, this situation is considered to have had no direct impact on the results of the study.

The means, standard deviations, and 95% CIs of posttest and delayed posttest scores for all student groups on the two different types of tests are provided in Table 1. In the gap-fill exercise, the control group's mean accuracy was 18% on the posttest and remained the same on the delayed posttest. The mean accuracy of the Quizlet group was 50% on the posttest, which decreased to 38% on the delayed posttest. The mean accuracy of the WordUp group was 59% on the posttest and decreased to 49% on the delayed posttest. Despite having a lower mean prior knowledge of the target words at the beginning of the experiment, the WordUp group's learning gains were the highest, followed by the Quizlet group and the control group, respectively.

As for multiple-choice questions, the mean accuracy of the control group was 19% on both the immediate and delayed posttests. The Quizlet group's accuracy was 50% on the posttest and 45% on the delayed posttest. The WordUp group, on the other hand, achieved the highest scores compared to the other two groups. The posttest accuracy for the WordUp group was 57%, and the delayed posttest accuracy was 47%. The mean accuracy of the WordUp group was higher than that of the Quizlet and control groups on both the posttest and delayed posttest across all types of questions.

4.2 Main analyses

The results of the final mixed logit model are given in Table 2. Model diagnostics showed no evidence for multicollinearity (all VIFs were below 2) and overdispersion of residuals

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

(dispersion = 0.99, $p = .59$), no outliers ($p = .915$), and no deviation from residual uniformity (Kolmogorov–Smirnov test; $p = .29$), suggesting good model fit.

The results demonstrated a statistically significant effect of group on students' response accuracy. The odds of obtaining a correct answer for the Quizlet group were 9.38 times the odds for the control group ($d = 1.24$). The odds of obtaining a correct response for the WordUp group were 15.62 times the odds for the control group ($d = 1.52$). Moreover, the nonsignificant main effect of test showed that the scores of the control group on the posttest and delayed posttest were similar. However, the significant interactions of test with the groups indicated a statistically significant decrease in response accuracy in delayed posttest for the Quizlet and WordUp groups.

The pairwise comparisons with Tukey adjustment $p < .05$ showed statistically significant differences between the learning gains of all three groups on both immediate and delayed posttests (Table 3). Overall, the results from both models showed that the students in the WordUp group had the highest learning gains, followed by the Quizlet group. The control group's learning gains were negligible.

The main effect of question type was nonsignificant, indicating that question type was not a significant predictor of learning gains. However, students' vocabulary levels, measured using the UVLT, and their pretest scores were positively associated with their learning gains. Students with higher vocabulary levels had greater learning gains than those with lower vocabulary levels.

4.3 Interview results

The content analysis of the interview data revealed two categories, each with its own set of themes: the favorable aspects of each MAVL app and areas for enhancement (Table 4).

The students interviewed in the Quizlet group highlighted flashcard repetition as one of the app's main strengths. They appreciated the opportunity to review selected words by categorizing flashcards as "I know" and "I don't know," which allowed them to focus on unfamiliar vocabulary. For example, S7 stated, "I like that I can pick words to review," emphasizing how this repetition feature supported their learning. Students also described the app as user-friendly, highlighting its simplicity and ease of use. As S2 noted, "It's really easy to use. I can just go in and start studying without any trouble." The app's test and practice feature, including multiple-choice, matching, and flashcard activities, was perceived positively. For example, S4 commented, "Even after I learned the words, I enjoyed repeating the exercises."

Regarding areas for enhancement, many students expressed a desire for visual and video support. They suggested that pictures or videos illustrating word meanings and usage would make learning more engaging and facilitate understanding. For instance, S1 remarked, "It would have been helpful to see a picture or a video to really understand the word." Students also recommended incorporating more gamified features, such as competitive challenges and rankings, to increase motivation and enjoyment. As S10 commented, "It would be better if the games felt more competitive."

The students in the WordUp group, on the other hand, expressed appreciation for the app's visual and video support, which they believed facilitated vocabulary learning and retention. As S1 explained, "The videos were useful because I could see the word usage, and it made it easier to remember." They also reported positive perceptions of the AI-assisted speaking and writing activities. As S4 stated, "I liked writing and speaking practice features. It felt like a real conversation." Furthermore, students appreciated the contextualized vocabulary use provided through example sentences. S3 explained, "Seeing the word usage in a sentence helped me understand and remember it better."

Regarding areas for enhancement, students most frequently referred to the complexity of the interface, which some found difficult to navigate. For example, S9 mentioned, "I spent several minutes just trying to find the cards to study due to the interface's complexity." Students also reported that some example sentences were difficult to understand because they contained

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

unfamiliar vocabulary. As S2 explained, "The sentences were hard to understand because there were words I didn't know. It would have been better if the sentence had the Turkish translation." In addition, several students noted that the language used in the videos exceeded their proficiency level. As S1 remarked, "Some of the videos were difficult to understand because they used words I didn't know." Similarly, S5, S7, and S9 reported that they often focused only on identifying the target word because the videos were linguistically demanding.

Overall, the interview findings indicate that both apps were perceived positively, albeit for different reasons. Learners valued Quizlet for its straightforward retrieval-based practice and ease of use, whereas WordUp was appreciated for its multimedia support and AI-assisted learning opportunities, despite some challenges related to interface complexity and the linguistic difficulty of some example sentences and videos.

5. Discussion

5.1 Short-term learning gains

In terms of vocabulary learning gains, both experimental groups significantly outperformed the control group, highlighting the value of structured, intentional vocabulary learning beyond regular classroom instruction. While learners in the control group continued their regular general English classes without structured out-of-class vocabulary practice, those in both experimental groups engaged in systematic practice with the target vocabulary beyond the classroom. The substantial gap between the experimental and control groups therefore suggests that regular, deliberate engagement with target vocabulary may contribute more to vocabulary development than relying on classroom instruction alone. This interpretation lends further support to Nation's (2022) argument that classroom exposure alone is rarely sufficient for substantial lexical growth and underscores the importance of intentional out-of-class vocabulary learning.

The comparison between the two experimental conditions further suggests that different MAVL environments may facilitate vocabulary learning through different instructional mechanisms. Quizlet appears to have supported vocabulary learning primarily through repeated retrieval practice and focused review of target vocabulary, which is consistent with previous research demonstrating the effectiveness of retrieval-based digital flashcard apps in promoting vocabulary acquisition (Li & Hafner, 2022; Rahmani et al., 2022; Yaşar & Koçoğlu, 2025; Zakian et al., 2022). This interpretation is reinforced by the interview findings, as participants consistently described Quizlet as an efficient and easy-to-use environment that facilitated repeated review and focused vocabulary practice. In contrast, the significantly greater learning gains observed in the WordUp group may be explained by its integration of contextualized example sentences, visual and video support, and AI-assisted speaking and writing activities. These features may have promoted more meaningful vocabulary processing by encouraging learners to connect new lexical items with multiple representations and opportunities for productive use, consistent with the CTML (Mayer, 2009) and generative learning theory (Fiorella & Mayer, 2016). Interview findings similarly indicated that learners valued these contextualized and interactive features, frequently describing them as helping them understand, remember, and actively use newly learned vocabulary in meaningful contexts. This interpretation is also consistent with previous studies reporting positive effects of AI-assisted and multimedia-supported vocabulary learning environments (Abdelhalim & Alsehibany, 2025; Iravi & Malmir, 2023; Polyzi & Moussiades, 2023). Learners' interview responses further revealed that many of the features they perceived as missing in one application were strengths of the other, suggesting that the two apps offered complementary rather than competing learning opportunities. Rather than identifying a universally superior instructional approach, the present findings suggest that retrieval-based practice, multimedia support, and AI-assisted productive activities each contribute to vocabulary learning through distinct yet complementary instructional mechanisms. This highlights that the

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

effectiveness of MAVL depends less on whether an app is retrieval-based or AI-assisted than on how its pedagogical features support vocabulary processing and learning.

Another noteworthy finding is the positive correlation between learners' vocabulary size, as measured by the UVLT, and their vocabulary learning gains. Students with higher initial vocabulary knowledge achieved greater vocabulary gains than those with lower initial vocabulary knowledge. This finding aligns with the Matthew effect in language development (Coyne et al., 2019), which posits that learners with a strong foundational vocabulary benefit more from new learning opportunities, thereby widening performance differences over time. Beyond confirming this pattern, however, the present findings suggest that initial vocabulary knowledge may also serve as an important predictor of learners' responsiveness to mobile-assisted vocabulary instruction. This effect emerged across both experimental conditions, indicating that prior vocabulary knowledge remained influential despite the different instructional mechanisms provided by the two applications. Although this pattern does not fully align with the expertise reversal principle, which proposes that instructional support is generally more beneficial for learners with lower levels of prior knowledge (Kalyuga, 2021), a different pattern emerged in the present study. One possible explanation is that learners with larger existing vocabularies were better able to integrate newly encountered lexical items into their existing lexical networks, regardless of whether learning occurred through retrieval-based practice or more contextualized multimedia-supported activities. This may have enabled them to benefit more from the instructional opportunities provided by both learning environments. This interpretation also resonates with the qualitative findings, as several learners reported that some of WordUp's videos and example sentences were linguistically demanding, suggesting that richer learning resources may be more readily exploited by learners with stronger existing lexical knowledge. Taken together, these findings suggest that maximizing the effectiveness of MAVL may require greater consideration of learners' existing lexical knowledge when selecting or designing digital learning environments.

5.2 Long-term vocabulary retention

Although both experimental groups demonstrated significant vocabulary loss between the immediate and delayed posttests, the WordUp group maintained significantly higher retention than the Quizlet group. This finding suggests that while some degree of forgetting appears to be an inevitable consequence of discontinuing structured vocabulary practice, different MAVL environments may vary in their capacity to support long-term vocabulary retention. One explanation for the decline observed in both experimental groups is the absence of continued, spaced engagement with target vocabulary following the intervention. Retrieval-based learning research consistently demonstrates that repeated retrieval and distributed practice are essential for strengthening lexical memory traces and supporting long-term retention (Karpicke, 2012; Webb & Nation, 2017). Likewise, Nation (2022) argues that newly learned vocabulary is unlikely to be retained without continued opportunities for review and retrieval. From this perspective, the decline observed in both experimental groups is not unexpected, as learners no longer engaged in systematic vocabulary practice once the intervention had ended. The present findings therefore reinforce the view that vocabulary learning should be regarded as an ongoing process requiring sustained engagement rather than a onetime instructional event, consistent with Ebbinghaus's (1913) forgetting curve.

However, the superior delayed retention observed in the WordUp group suggests that long-term vocabulary retention also depends on the quality of lexical representations established during initial learning. Rather than assuming that forgetting is determined only by the frequency of subsequent review, the present findings indicate that the depth and richness of initial lexical encoding also influence how successfully vocabulary is retained over time. This interpretation is reflected in the interview findings, where many learners explained that contextualized example

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

sentences, videos, visual support, and AI-assisted speaking and writing activities helped them remember not only the meanings of target words but also how they could be used appropriately in context. Such experiences suggest that richer initial learning experiences strengthened lexical representations during vocabulary learning, thereby contributing to better delayed retention despite the absence of continued practice. In this respect, these findings extend previous research by suggesting that the durability of vocabulary learning depends not only on opportunities for later retrieval but also on the quality of the initial learning experience through which lexical knowledge is established.

The interview findings also shed light on why richer learning environments did not benefit all learners equally. Several participants reported that the videos and example sentences contained linguistically demanding vocabulary and complex sentence structures, making comprehension difficult. Others perceived the interface as overly complex, with multiple learning options sometimes reducing, rather than facilitating, learning efficiency. These observations suggest that instructional richness does not automatically translate into more effective learning. Instead, the pedagogical value of multimedia resources and AI-assisted activities appears to depend on maintaining an appropriate balance between instructional richness, cognitive accessibility, and learners' existing vocabulary knowledge. Overall, these findings highlight the importance of considering not only the instructional features incorporated into MAVL environments but also how those features align with learners' cognitive and linguistic readiness.

5.3 Pedagogical implications

This study offers several pedagogical and practical implications for language teachers and MAVL app developers. The substantial differences between the experimental and control groups indicate that meaningful vocabulary development benefits from structured, intentional engagement with target vocabulary beyond regular classroom instruction. Rather than serving merely as supplementary resources, MAVL applications can be systematically integrated into vocabulary instruction through retrieval practice, productive language use, and spaced review. Intentional out-of-class vocabulary learning should be planned as an integral component of vocabulary instruction rather than left entirely to learners' independent study. The findings further suggest that retrieval-based and AI-assisted multimedia environments support vocabulary learning through complementary instructional mechanisms. Accordingly, teachers may benefit from combining retrieval-based review with contextualized, multimedia-supported activities to promote both vocabulary consolidation and productive language use. For developers of MAVL applications, the interview findings highlight the importance of balancing instructional richness with interface simplicity. While learners valued multimedia resources, AI-assisted activities, and motivational features such as gamification, they also emphasized the need for clear interface design and manageable content complexity. Future MAVL applications should therefore prioritize learner-centered, adaptive designs that align instructional features with learners' proficiency levels and cognitive readiness rather than simply increasing the number of technological features.

5.4 Limitations and future research

Despite its contributions, this study has some limitations. First, as WordUp integrates both multimedia and AI-assisted interaction features, it is difficult to isolate the individual effects of these components on learning outcomes. Future research should examine these features independently to better understand their relative contributions. Second, reliance on self-reported study time in the Quizlet group may have introduced response bias; future studies could use more objective measures such as in-app analytics or automated usage logs. Third, the relatively small sample size and single-institution context limit the generalizability of the findings. The interview participants were volunteers, which may have introduced self-selection bias, as those who chose to

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

participate may not fully represent the experiences of all learners. Replication across different educational contexts, age groups, and proficiency levels would strengthen the external validity of the results. Finally, longitudinal research is needed to examine the sustained effects of different mobile-assisted instructional designs on vocabulary retention, productive vocabulary use, and overall language development. Future studies should also investigate how learner characteristics interact with different instructional features to optimize vocabulary learning outcomes.

6. Conclusion

This study contributes to the growing body of research on MAVL by comparing two pedagogically distinct applications and examining the instructional mechanisms that may explain their differential effects on vocabulary learning. Although both applications produced substantially greater vocabulary gains than regular classroom instruction alone, they appeared to do so through different yet complementary instructional mechanisms. Retrieval-based practice supported systematic review and repeated engagement with target vocabulary, whereas contextualized multimedia input and AI-assisted productive activities appeared to promote richer lexical processing and more durable vocabulary retention. The interview findings further supported these interpretations by highlighting the distinctive pedagogical features of each learning environment.

Beyond these findings, the study suggests that the effectiveness of MAVL should be understood primarily as a matter of instructional design rather than technological innovation. The findings indicate that the pedagogical value of technological features lies not in their mere presence, but in how they are integrated and aligned with learners' existing vocabulary knowledge, cognitive readiness, and opportunities for sustained engagement.

The findings further suggest that future MAVL research may benefit from using comparisons of individual applications to identify the instructional principles that best support meaningful, sustainable, and learner-responsive vocabulary development. Such a perspective may provide a more productive framework for both the design of vocabulary learning applications and the integration of mobile technologies into language education.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0958344026100615

Data availability statement. The data that support the findings of this study are available within the article's supplementary material and in a public repository: https://osf.io/j8eup/overview?view_only=0a0d2eec150f4cf29f2badff9a2fd525

Authorship contribution statement. Derya Tuzcu-Eken: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Validation.

Funding disclosure statement. This research did not receive any specific funding. However, the open-access publication fee was covered by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Competing interests statement. The author declares no competing interests.

Ethical statement. Ethical approval for this study was obtained from the author's institution. Participation was entirely voluntary, and informed consent was obtained from all participants.

GenAI use disclosure statement. ChatGPT-4 was used to improve the readability and clarity of the manuscript; however, all scholarly content and contributions are the author's own.

References

Abdelhalim, S. M., & Alsehibany, R. (2025). Integrating ChatGPT for vocabulary learning and retention: A classroom-based study of Saudi EFL learners. *Language Learning & Technology*, 29(1), 1–24. https://doi.org/10.64152/10125/73635

Alrajhi, A. S. (2025). Intelligent personal assistants in self-access L2 vocabulary learning. *Education and Information Technologies*, 30(1), 801–831. https://doi.org/10.1007/s10639-024-12967-7

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review*, 55(1), 589–656. https://doi.org/10.1007/s10462-021-10039-7

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, 9(2), 105–121. https://doi.org/10.1016/j.asw.2004.07.001

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Chen, Q., Liu, S., Huang, K., Wang, X., Ma, X., Zhu, J., & Peng, Z. (2024). RetAssist: Facilitating vocabulary learners with generative images in story retelling practices. In A. Vallgårda, L. Jönsson, J. Fritsch, S. Fdili Alaoui, & C. A. Le Dantec (Eds.), *DIS '24: Proceedings of the 2024 ACM Designing Interactive Systems Conference* (pp. 2019–2036). Association for Computing Machinery. https://doi.org/10.1145/3643834.3661581

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, 19(22), 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::aid-sim784>3.0.co;2-m

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, 80(2), 183–198. https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

Cohen, Y. (2017). Estimating the intra-rater reliability of essay raters. *Frontiers in Education*, 2, Article 49. https://doi.org/10.3389/feduc.2017.00049

Coyne, M. D., McCoach, D. B., Ware, S., Austin, C. R., Loftus-Rattan, S. M., & Baker, D. L. (2019). Racing against the vocabulary gap: Matthew effects in early vocabulary instruction and intervention. *Exceptional Children*, 85(2), 163–179. https://doi.org/10.1177/0014402918789162

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). Teachers College, Columbia University. (Original work published 1885)

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, 28(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Hartig, F. (2024). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models* (R package Version 0.4.7) [Computer software]. https://CRAN.R-project.org/package=DHARMa

Iravi, Y., & Malmir, A. (2023). The effect of lexical tools and applications on L2 vocabulary learning: A case of English academic core words. *Innovation in Language Learning and Teaching*, 17(3), 636–649. https://doi.org/10.1080/17501229.2022.2102638

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kalyuga, S. (2021). The expertise reversal principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (pp. 171–182). Cambridge University Press. https://doi.org/10.1017/9781108894333.017

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, 21(3), 157–163. https://doi.org/10.1177/0963721412443552

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation*, 65, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

Lenth, R., & Piaskowski, J. (2026). emmeans: Estimated marginal means, aka least-squares means (R package version 2.0.5). https://rvlenth.github.io/emmeans/ [Computer software].

Leong, J., Pataranutaporn, P., Danry, V., Perteneder, F., Mao, Y., & Maes, P. (2024). Putting things into context: Generative AI-enabled context personalization for vocabulary learning improves learning motivation. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, & M. L. Wilson (Eds.), *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642393

Li, Y., & Hafner, C. A. (2022). Mobile-assisted vocabulary learning: Investigating receptive and productive vocabulary knowledge of Chinese EFL learners. *ReCALL*, 34(1), 66–80. https://doi.org/10.1017/S0958344021000161

Lin, J.-J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, 32(8), 878–919. https://doi.org/10.1080/09588221.2018.1541359

Lin, L., & Atkinson, R. K. (2011). Using animations and visual cueing to support learning of scientific concepts and processes. *Computers & Education*, 56(3), 650–658. https://doi.org/10.1016/j.compedu.2010.10.007

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist*, 32(1), 1–19. https://doi.org/10.1207/s15326985ep3201_1

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://drive.google.com/file/d/1UQKDYfqoNjHuPZp62t-yckVDa-VMRXYC/view

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, 24(1), 17–38. https://doi.org/10.1080/09588221.2010.520675

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59–82. https://doi.org/10.1353/cml.2006.0049

Nation, I. S. P. (2012). The BNC/COCA word family lists. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists

Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781009093873

Peng, Z., Wang, X., Han, Q., Zhu, J., Ma, X., & Qu, H. (2023). Storyfier: Exploring vocabulary learning support with text generation models. In S. Follmer, J. Han, J. Steimle, & N. H. Riche (Eds.), *UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606786

Polyzi, P., & Moussiades, L. (2023). An artificial vocabulary learning assistant. *Education and Information Technologies*, 28(12), 16431–16455. https://doi.org/10.1007/s10639-023-11810-9

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rahmani, A., Asadi, V., & Xodabande, I. (2022). Using mobile devices for vocabulary learning outside the classroom: Improving the English as foreign language learners' knowledge of high-frequency words. *Frontiers in Psychology*, 13, Article 899885. https://doi.org/10.3389/fpsyg.2022.899885

Roediger, H. L., III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tuzcu, A. (2023). The effects of input mode in learning declarative and nondeclarative knowledge of L2 collocations. *System*, 113, Article 103006. https://doi.org/10.1016/j.system.2023.103006

Tuzcu-Eken, D., & Tuzcu, A. (2026). Enhancing technical vocabulary learning: A comparative study of a mobile multimedia app and a word list. *System*, 138, Article 103962. https://doi.org/10.1016/j.system.2025.103962

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *International Journal of Applied Linguistics*, 168(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

Yaşar, S., & Koçoğlu, Z. (2025). Enhancing vocabulary mastery: The impact of learner-created digital flashcards on L2 vocabulary learning and self-regulation. *Australasian Journal of Educational Technology*, 41(4), 17–31. https://doi.org/10.14742/ajet.10402

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, Article 16. https://doi.org/10.1186/s40862-022-00143-8

Zung, I., Imundo, M. N., & Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? *Memory*, 30(8), 923–941. https://doi.org/10.1080/09658211.2022.2058553

About the authors

Derya Tuzcu-Eken is an associate professor in the School of Foreign Languages at Kırklareli University, Türkiye. Her research interests include intercultural communication, technology-enhanced language learning, vocabulary acquisition, and L2 development.

https://doi.org/10.1017/S0958344026100615 Published online by Cambridge University Press