frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 18 September 2025
DOI 10.3389/feduc.2025.1641638

OPEN ACCESS

EDITED BY
Hassan Banaruee,
University of Education Weingarten, Germany

REVIEWED BY
Mauro Ocaña,
University of the Armed Forces (ESPE), Ecuador
Paul Sevigny,
Ritsumeikan Asia Pacific University, Japan

*CORRESPONDENCE
Mike Minwen Zhang
p121759@siswa.ukm.edu.my

RECEIVED 05 June 2025
ACCEPTED 19 August 2025
PUBLISHED 18 September 2025

CITATION
Zhang MM, Hashim H and Yunus MM (2025) Evaluating metaverse-based L2 vocabulary learning effectiveness using a proposed metric of vocabulary forgetting percentage. *Front. Educ.* 10:1641638.
doi: 10.3389/feduc.2025.1641638

Evaluating metaverse-based L2 vocabulary learning effectiveness using a proposed metric of vocabulary forgetting percentage

Mike Minwen Zhang*, Harwati Hashim and Melor Md Yunus

Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Introduction: Vocabulary gain and retention are widely recognized as essential metrics in second language (L2) vocabulary learning. However, these traditional measures often fail to reflect the proportional loss of learned vocabulary knowledge over time, thus limiting their practical, diagnostic, and comparative value across different instructional contexts.
Methods: To address this gap, the present study proposes a percentage-based metric: Vocabulary Forgetting Percentage (VFP). To evaluate metaverse-based vocabulary learning (VL) effectiveness and also to empirically validate the VFP, a quasi-experiment was conducted, involving 50 Chinese middle school EFL learners who were assigned to either a metaverse-based group (MG) or a slides-assisted control group (SG). Over three learning sessions, participants learned equivalent vocabulary content and completed pretests, immediate post-tests, and delayed post-tests. Quantitative data were analyzed using independent-samples t-tests to compare vocabulary gains, retentions, and VFPs across groups.
Results: The MG significantly outperformed the SG in vocabulary gain and retention in each session and in mean scores. However, VFP results showed a different pattern: the MG's third-session and mean VFPs were significantly lower than those of the SG, while differences in the first and second sessions were not significant.
Discussion: The MG's late-emerging VFP difference from the SG's suggests that extended exposure to immersive environments may be required before full benefits appear. Findings also confirm the pedagogical potential of the metaverse for vocabulary learning and empirically validate VFP as a complementary metric. By proportionally quantifying vocabulary loss, VFP offers researchers and educators a more nuanced tool for evaluating learning efficiency and retention sustainability in varied L2 contexts.

KEYWORDS

vocabulary forgetting, vocabulary retention, vocabulary gain, second language acquisition, metaverse, L2 vocabulary learning, learning effectiveness metrics

Introduction

Vocabulary is one of the most critical factors in second language acquisition (SLA) (Crossley et al., 2009; Schmitt, 2000; Webb, 2005), as it significantly influences how learners comprehend and produce a foreign or second language. Learners' vocabulary knowledge greatly affects their success or failure in language learning (Afzal, 2019; Ng and Rosli, 2023). Scholars (Ghalebi et al., 2020; Arochman et al., 2023) have also argued that memorizing a large number of vocabulary items is a major challenge for foreign language learners, let alone retaining them in long-term memory. To develop autonomous vocabulary knowledge, learners

must be highly motivated to engage in a dynamic process of skill development that involves practicing various strategies and employing effective techniques (Nation, 2001). Flashcards, notebooks, dictionaries, and the use of synonyms and antonyms are among the most commonly used tools by educators and students for vocabulary instruction and acquisition (Altiner, 2019; Elgort and Nation, 2010; Oxford and Crookall, 1990; Lessard-Clouston, 2021). Despite these efforts, vocabulary acquisition and retention remain among the most daunting aspects of SLA (Milton, 2009; Schmitt, 2008).

Recent years have seen tremendous progress in the integration of technology, especially in vocabulary acquisition, which opens up new opportunities for dynamic and engaging learning experiences (Lin and Wei, 2024; Teymouri, 2024; Zhang et al., 2025). The field of vocabulary learning has been profoundly impacted by emerging technological developments, including mobile learning apps (Govindasamy et al., 2019), VR (Chen and Yuan, 2023), AR (Hung and Yeh, 2023), and, most recently, the metaverse (Wang et al., 2025).

A metaverse is a shared virtual environment where users can interact with each other and the environment in a simulated three-dimensional space using a combination of digital technologies and media such as audio, video, images, animations, 3D objects, and interactive elements like slideshows (Mystakidis, 2022). Because they are capable of incorporating many digital technologies that heighten the sensation of realism, presence, and co-presence, metaverse platforms are exceptionally immersive. The immersive and collaborative features of the metaverse have the potential to revolutionize traditional vocabulary learning practices (Çelik and Baturay, 2024; Wang et al., 2025).

To evaluate VL effectiveness in technology-enhanced learning contexts, various measures and metrics have been formulated and employed to assess learning outcomes from various perspectives. The most explicit and frequently used measures are the raw scores of vocabulary assessments, such as pretests, immediate posttests, and delayed posttests (Alfadil, 2020; Lee, 2023). Many scholars have also introduced computed measures based on simple calculations of learners' test results, including vocabulary gain and retention (Elekaei et al., 2020; Tai et al., 2022). Additionally, more complex formula-based metrics—such as forgetting count (Fukushima et al., 2024) and forgetting rate (Tabibian et al., 2019; Rahman et al., 2021; Rivera-Lares et al., 2023)—have been proposed. While these existing measures and metrics each have their own advantages in analyzing and capturing VL effectiveness, their limitations are also evident. Therefore, more nuanced, accurate, and comprehensive metrics are still warranted, particularly in the realm of technology-enhanced L2 vocabulary learning.

Literature review

Vocabulary knowledge is widely acknowledged as an inherently complex construct, and for pedagogical as well as assessment purposes, broad terms such as "vocabulary knowledge" are considered too imprecise for effective operationalization (Stewart et al., 2024). Within the context of L2 vocabulary knowledge, the "form-meaning link" is recognized as its core component. In addition to form and meaning, a distinction can be made between recognition, in which a learner is presented with an L2 word form and is expected to activate its meaning(s), and recall, in which the learner is given some kind of stimulus that prompts the activation of the L2 word form from memory (Read, 2000). Schmitt (2010) classified overall vocabulary knowledge into four sub-constructs: form recognition, form recall, meaning recognition, and meaning recall. The recognition–recall distinction may carry significant implications for achieving full word mastery. Therefore, in this study, the design of the vocabulary learning sessions and vocabulary tests incorporated the four sub-constructs to enhance the accuracy and comprehensiveness of the measurement.

Vocabulary assessments

In L2 vocabulary learning (L2VL) research, how to effectively measure the learning outcomes of vocabulary knowledge has always warranted more investigations and better solutions. In the previous studies, vocabulary learning effectiveness has been widely assessed by various vocabulary tests. Researchers often rely on established assessments, including the Vocabulary Levels Test (VLT) (Nation, 2001; Schmitt et al., 2001), Wesche and Paribakht's Vocabulary Knowledge Scale (VKS) (Wesche and Paribakht, 1996); British Picture Vocabulary Scales (BPVS) (Dunn et al., 1997); Nelson-Denny Reading Test – Vocabulary Subtest (Brown et al., 1993); Cambridge Assessment English – B1 Level Vocabulary Test (Cambridge Assessment English, 2020); Peabody Picture Vocabulary Test (PPVT) (Dunn and Dunn, 2007); Laufer and Nation's Vocabulary-Size Test of Controlled Productive Ability (Laufer and Nation, 1999); Oxford Young Learners Placement Test (YLPT) (Oxford University Press, 2014). Additionally, many vocabulary researchers tailor their own tests to suit the target vocabulary and proficiency level of participants (Webb, 2005).

In empirical L2VL studies, the aforementioned standardized or researcher-developed vocabulary tests are often administered multiple times at different points of a VL intervention to measure learning gain and retention over time and also to capture the variability of a learner's learning performances (Schmitt, 2010). Data on pretest, immediate posttest, and delayed posttest scores are commonly gathered at three distinct time points (Nation, 2001). 1. Pretest Score: The baseline measure of vocabulary knowledge before the learning intervention. 2. Immediate Posttest Score: The measure of vocabulary knowledge immediately after the learning activity. 3. Delayed Posttest Score: The measure of vocabulary knowledge after one or many designated time delays (Webb, 2005).

The pretest is administered before the learning intervention to establish learners' baseline vocabulary knowledge. It ensures that participants have not previously mastered the target vocabulary items, and it provides a reference point for measuring subsequent gains. The immediate posttest is conducted immediately after the VL intervention (e.g., reading activity, multimedia instruction, app-based learning, or game-based practice). It is designed to capture the vocabulary gain, or the amount of knowledge acquired during the learning phase. The item types and test formats in the immediate posttest usually mirror those of the pretest to ensure measurement consistency (Read, 2000). Consistency in format is crucial for isolating the learning effect (Nation, 2001). The delayed posttest(s) are administered after a defined period—commonly one week, two weeks, or even several months after the immediate posttest—to assess vocabulary retention and measure the durability of learning (Barcroft, 2009). Like the pretest and immediate posttest, the delayed posttest(s) use consistent formats to allow direct score comparison. The time intervals are selected based on the research objective: shorter intervals assess short-term retention, while longer delays provide insights into long-term memory

consolidation. In many cases, delayed posttest(s) are administered without prior warning to avoid rehearsal effects (Zhou, 2010).

Vocabulary gain and retention

Two metrics—"vocabulary gain" and "vocabulary retention"—are commonly used by researchers as key indicators of VL effectiveness (Faramarzi et al., 2014; Okyar and Çakır, 2019), particularly in technology-assisted contexts (Elekaei et al., 2020; Lee, 2023; Tai et al., 2022). Concurrently, the three terms "vocabulary gain," "vocabulary acquisition" (Chen and Yuan, 2023; Ersanli, 2023), and "vocabulary learning" (Alfadil, 2020; Sahinler, 2023) are often used interchangeably in technology-enhanced vocabulary learning research. While "vocabulary acquisition" and "vocabulary learning" also refer to the overall VL process, they can cause conceptual confusion in many cases. To ensure terminological clarity, "vocabulary gain" is opted for in this study to refer to the immediate VL effectiveness.

Vocabulary gain (VG) was defined as the short-term memory retrieval of vocabulary knowledge, measured by subtracting the vocabulary pretest from the immediate post-test score (Nation, 2001; Webb, 2007; Lai and Chen, 2023; Reynolds et al., 2022), and refers to the increase in the number of words and expressions that an individual learns and integrates into their active lexical memory immediately after the VL interventions. VG is commonly calculated as:

$$Vocabulary\ Gain\left(VG\right) = Immediate\ Posttest\ Score - Pretest\ Score$$

Vocabulary retention (VRe), on the other hand, is often viewed as a more intricate cognitive process of memory incorporating memorization or acquisition, recall, and recognition (Suleiman, 2009). VRe was defined by Richards and Schmidt (2002) as "the ability to recall or remember things after an interval of time" or long-term memory retrieval of vocabulary knowledge. Mohammed (2009) defines VRe as "the ability to retain the acquired vocabulary and retrieve it after a period of time following a certain learning intervention." Therefore, VRe is commonly measured by the difference between a delayed post-test score and a vocabulary pretest score (Barcroft, 2004; Nation, 2001; Webb, 2007; Zhong, 2018).

$$Vocabulary\ Retention\left(VRe\right) = Delayed\ Posttest\ Score - Pretest\ Score$$

Unlike the immediate posttest and the delayed posttest, vocabulary gain and retention emphasize the changes in vocabulary knowledge resulting from specific learning interventions. Thus, learners' initial vocabulary proficiency should be controlled to account for individual variation when calculating vocabulary gain and retention.

While both VG and VRe are valuable indicators of VL performances, studies have long questioned their adequacy in fully capturing the learning effectiveness (Milton, 2009; Schmitt, 2010), especially with the increasing complexity of digital and immersive learning environments (Feng and Ng, 2024; Weng et al., 2024). VG and VRe, as raw scores, are heavily influenced by learners' baseline proficiency and the absolute difficulty of the vocabulary items, potentially skewing interpretations of effectiveness. Moreover, VG and VRe are limited in their interpretability for comparative or longitudinal research. Vocabulary gain may overestimate effectiveness by capturing short-term memorization rather than durable learning. Similarly, retention scores alone may mask inefficient learning processes, particularly if learners retained very little relative to what they initially gained.

Several recent empirical studies echo the need for more nuanced effectiveness measures. Lai and Chen (2023) emphasize that retention should not be viewed in isolation but rather in tandem with acquisition metrics. Likewise, Elekaei et al. (2020) and Zhang et al. (2025) highlight that a higher initial gain followed by a steep forgetting curve may suggest superficial or ineffective learning strategies. Consequently, there is a growing consensus that evaluating the forgotten vocabulary knowledge—rather than solely what is gained or retained—can provide a more complete picture of L2 students' learning effectiveness in varying contexts (Bahrick et al., 1993; Kornmeier et al., 2022; Sense et al., 2018).

Review of previous metrics for vocabulary forgetting

There are several existing metrics to measure forgetting or memory loss in the prior literature. Three formulas previously proposed specifically for vocabulary knowledge forgetting emerged in the L2 learning domain:

In Fukushima et al.'s (2024) study, the vocabulary forgetting measure, inconsistently referred to as "forgetting rate" and "forgetting count," was calculated as "subtraction of the immediate posttest score from the one-week delayed posttest score."

$$Forgetting\ Count = Immediate\ Posttest\ Score - Delayed\ Posttest\ Score$$

Notably, the term "forgetting rate," which was mostly used in the article, was not an accurate phrasing in light of the nature of the rate, which is a time unit and supposedly evaluates how fast the vocabulary knowledge is forgotten. While the simple, raw-score-based formula solely captures the absolute number of vocabulary items forgotten during a given time interval, it lacks sensitivity to initial learning performance and cannot adequately support comparative or inferential analyses across different learners, groups, or instructional settings.

Also, adapted from an empirical formula of forgetting rate proposed by Tabibian et al. (2019), Rahman et al. (2021), in their study concerning reducing forgetting rate in EFL students using a spaced repetition-powered digital game-based learning application, proposed a revised version of the forgetting rate metric devoted to EFL education. This empirical forgetting rate is a measure of how fast the memory of an item decays after a single exposure at a certain point in time.

$$\hat{n}_{0,(u,i)} = \frac{-\log(\hat{m}\left(t_{(u,i),2}\right)}{t_{(u,i),2} - t_{(u,i),1}}$$

In this formula, $\hat{n}_{0,(u,i)}$ = initial forgetting rate; $\hat{m}\left(t_{(u,i),2}\right)$ = a single word item learning performance tested by a single question in binary value of the second attempt; $t_{(u,i),2} - t_{(u,i),1}$ = time interval between the second attempt and the preceding one.

Firstly, this formula is designed for single-item forgetting; although it provides nuanced variations across items, it is not optimal

for analyzing the learning outcomes of a list of vocabulary or inter-group learning performance analysis. Secondly, using a binary (0 or 1) score for the second attempt oversimplifies the actual learning performance. It does not capture partial correctness, degrees of confidence, or nuanced performance, making the forgetting rate overly sensitive to a single correct/incorrect response, especially in contexts with more complex or probabilistic learning data. Thirdly, the formula focuses only on performance at the second attempt, $t_{(u,i),2}$ but ignores how well the item was learned during the first exposure. Without accounting for initial performance or learning strength, the forgetting rate estimate may misrepresent the true rate of forgetting. Fourth, this formula, albeit innovative, is based on a decay curve (forgetting curve) primarily modeled by a logarithmic function (Ebbinghaus, 1913; Murre and Dros, 2015). Human memory retention, however, might not fully adhere to these simple decay laws, particularly for complex materials or diverse learning settings, which does not necessarily reflect the reality (Cepeda et al., 2006; Soderstrom and Bjork, 2015; Wixted, 2004). Lastly, this logarithmic and time-normalized formula can be too sensitive to measurement errors. When analyzing a large sample or comparing across different groups, the errors can be magnified, and the results can be distorted.

Additionally, Rivera-Lares et al. (2023) implicitly indicated the term "forgetting rate" in a line graph by illustrating the means of the number of correct responses at three different points in time and the linear slopes by connecting every two adjacent time points. Hence, the forgetting rate in this study can be calculated by dividing the difference between two successive test results by the time interval between them. The formula is illustrated as follows:

$$Forgetting\ \ Rate = \frac{S_{t_1} - S_{t_2}}{\Delta t}$$

There are also two drawbacks to this formula. First, this study did not involve VL but sentence learning in its experiment design, though forgetting rate was investigated; second, no baseline test or pretest was conducted to homogenize the initial knowledge levels of learners. Therefore, this formula cannot be directly applied to L2 vocabulary acquisition.

Overall, a more interpretable, diagnostic, and comparative metric for vocabulary forgetting can be regarded as a necessary complement to VG and VRe, especially in technology-enhanced vocabulary learning. However, how to rigorously and scientifically calculate forgotten vocabulary knowledge remains unexplored in L2VL. Therefore, the current study aims to propose a new formula for vocabulary forgetting as a measure to evaluate L2 vocabulary effectiveness; subsequently, the researcher attempts to test the validity of the newly proposed formula in two different learning environments.

Research questions

1 How can the VFP be calculated to measure L2VL effectiveness?
2 To what extent does the metaverse-based learning approach affect the VL effectiveness among middle school EFL learners compared to the traditional slide-assisted counterpart?
3 What are the differences in capturing L2VL effectiveness between the proposed formula of VFP and the existing measures of VG and VRe in the metaverse-based learning context?

Vocabulary forgetting percentage formula formulation

Dissatisfied with widely used metrics like VG and VRe and also inspired by the aforementioned formula-based metrics for vocabulary forgetting, the current research is an effort to provide a metric for vocabulary retention loss to more rigorously evaluate the VL effectiveness across various L2 instructional settings.

A new metric, the Vocabulary Forgetting Percentage (VFP), is proposed in the current study. This metric aims to quantify the proportion of initially gained vocabulary that is subsequently forgotten over time, thereby offering a standardized and comparative indicator of long-term VL effectiveness. VFP is defined as the normalized percentage of vocabulary forgotten within a given time interval relative to the initial vocabulary gain immediately after the learning intervention. It offers a measure of retention loss across different groups in different learning environments. The formula for calculating the VFP is expressed as:

$$\text{Vocabulary Forgetting Percentage}\left(\text{VFP}\right) = \frac{\text{Retention Loss}}{\text{Vocabulary Gain}} \times 100\%$$

Explanation of the proposed formula

The numerator of the proposed formula, Retention Loss (RL), reflects the absolute number of vocabulary items forgotten between the immediate and delayed posttests, which can be represented as:

$$\text{Retention Loss}\left(\text{RL}\right) = \text{Immediate Posttest Score} - \text{Delayed Posttest Score}$$

Since the pretest score is typically fixed and consistent for each learner, vocabulary gain is measured by subtracting the pretest score from the immediate posttest score, and vocabulary retention is measured by subtracting the pretest score from the delayed posttest score, we can also express RL as the difference between VG and VRe.

$$\begin{aligned}\text{RL} = &\left(\text{Immediate Posttest Score} - \text{Pretest Score}\right) - \\ &\left(\text{Delayed Posttest Score} - \text{Pretest Score}\right) = \\ &\text{Vocabulary Gain} - \text{Vocabulary Retention}\end{aligned}$$

However, this raw loss can be misleading if considered in isolation, as it does not account for how much was learned initially from the learning intervention. To address this, the denominator, or VG, represents the total number of vocabulary items learned as a result of the intervention (from pretest to immediate posttest). Notably, by introducing VG instead of the Immediate Posttest Score in the denominator, the variation in initial vocabulary proficiency can

be controlled for when comparing VFP values across different learners.

By dividing RL by VG, the formula computes the proportion of learned vocabulary that was lost and then multiplies it by 100 to express the result as a percentage. This yields a forgetting percentage that is both normalized and interpretable, providing meaningful comparisons across learners, instructional approaches, and research studies.

Meanwhile, based on the breakdown of the concepts above, it can also be written as:

$$\text{VFP} = \frac{\text{Immediate Posttest Score} - \text{Delayed Posttest Score}}{\text{Immediate Posttest Score} - \text{Pretest Score}} \times 100\%$$

or

$$\text{VFP} = \frac{\text{Vocabulary Gain} - \text{Vocabulary Retention}}{\text{Vocabulary Gain}} \times 100\%$$

This formula calculates the proportion of vocabulary forgotten after a specified retention interval, relative to the vocabulary gain in a certain learning intervention. The immediate posttest score represents learners' short-term retention following instructional intervention, while the delayed posttest score reflects longer-term retention after a period of time has passed. By expressing the loss as a percentage of the vocabulary gain rather than a difference between the immediate posttest score and the delayed posttest score (Fukushima et al., 2024), the formula normalizes retention loss across different performance levels, thus offering a standardized and interpretable metric for both intra-group and inter-group comparisons.

Illustrative example

If intending to calculate and compare the VFPs of two vocabulary learners using two different learning methods, their pretest, posttest and one-week delayed test raw scores need to be first collected, respectively. Their scores are listed as follows:

	Pretest	Posttest	One-Week Delayed Posttest
Student A	20	50	35
Student B	25	60	35

Based on the formula:

$$\text{Vocabulary Forgetting Percentage}\left(\text{VFP}\right) = \frac{\text{Retention Loss}}{\text{Vocabulary Gain}} \times 100\%$$

Hence, the respective VG, RL, and VFP for students A and B are listed below.

	VG	RL	VFP
Student A	30	15	50%
Student B	35	25	71.4%

This means that, despite the same scores on the one-week delayed posttest, student A still showed a lower VFP than student B.

Evaluation of metaverse-based VL effectiveness and validation of the proposed formula

Methodology

The second phase of this study adopts a quasi-experimental study design. This phase aims to examine the L2 learners' VL effectiveness in the metaverse-based learning environment and to compare the VFP measure with traditional metrics of VG and VRe. By doing so, it also seeks to determine whether the VFP formula provides a more comprehensive and sensitive assessment of VL effectiveness across different instructional modalities.

Participants and sampling

The subjects of the study were Grade 8 students in a public middle school in Mainland China. Fifty students (26 males and 24 females, aged 13 to 15, M = 14.64) consented to participate in the study. Of these, 25 students utilized a metaverse platform, *Spatial*, to engage in three metaverse-based learning sessions (metaverse-based group, or MG), learning 20 words or expressions per lesson. The remaining 25 students in the control group attended PowerPoint slide-assisted learning sessions in a traditional classroom to acquire the same vocabulary knowledge (slide-assisted group, or SG). A *t*-test (two-tailed $p = 0.425$, $p > 0.05$) indicated that the initial vocabulary proficiency level of the MG and the SG was not statistically significant. All participants are native Chinese speakers with no prior experience utilizing any metaverse-based platforms or tools for English language learning before this study. Both groups were instructed by one English teacher appointed by the study's researchers. The approval from the middle school was obtained prior to the commencement of the formal data-gathering process. All participants and their legal guardians were informed of the study's purpose, procedures, and their right to withdraw at any time without penalty. Written informed consent was obtained from all participants and their guardians prior to participation.

Metaverse platform--*Spatial*

This study utilized a metaverse platform called *Spatial.io* (henceforth referred to as *Spatial*). *Spatial*[1] is a free-access, open-source metaverse platform. *Spatial* allows multiple people to engage in existing virtual environments or construct new ones for gaming, meetings, chatting, collaborative learning, and content sharing, accessible via regular web browsers or in real-time with VR headsets. *Spatial* provides high-fidelity, rich-content learning experiences, allowing users to engage more immersively with the material and fellow participants in both the paid and free versions. The present study utilized the free version, which offers the same

1 https://www.spatial.io

level of customizability, enabling users to construct individualized metaverses from the ground up or to configure tailored VR rooms and avatars by modifying templates and incorporating an extensive array of intricate 3D content, embellishments, and toolkits provided on the platform.

Learning instruments and materials

Metaverse-based and slide-assisted VL sessions

The researcher prepared three consecutive and coherent VL sessions for both the MG and the SG in accordance with the English Curriculum Standards and the themes outlined in the Grade 8 middle school English textbook published by People's Education Press. In alignment with the Grade 8 English curriculum, the exact three sessions (Animal Kingdom, Treasure Island and A Trip to Thailand) were developed for both the MG and SG, with both groups acquiring the same sets of vocabulary. Each learning session involved the acquisition of 20 words (see Appendix 1). The majority of the vocabulary acquired was sourced from the Grade 8 textbook's word list, with additional terms and idioms incorporated based on their pertinence to current virtual environments and their frequency of use. Three experts were requested to assess the validity of the curriculum design for Metaverse-based VL sessions and the vocabulary list acquired (Figures 1–4).

English vocabulary proficiency test (EVPT)

The English Vocabulary Proficiency Test (EVPT) was adapted based on the Vocabulary Size Test (bilingual Mandarin version) developed by Nation and Beglar (2007). The original assessment has 14,000 words and includes 140 multiple-choice questions, with 10 questions derived from each 1,000-word family level. EVPT aims to test L2 learners' receptive vocabulary knowledge (Nation and Beglar, 2007). The total score of a student must be multiplied by 100 to determine their overall receptive vocabulary. According to the English Curriculum Standards for Nine-Year Full-Time Compulsory Education (the Ministry of Education of the People's Republic of China 2022), however, the average required vocabulary size for middle school graduates is approximately 1,600 words, including idioms and collocations, and the maximum requirement is approximately 2000 words and expressions. Presumably, in case some Grade 8 students' vocabulary sizes are superior to the average level, it is reasonable to only keep the multiple-choice items between the 1st and the 30th item in the Vocabulary Size Test. Hence, the items beyond the 3,000-word level are unnecessary to be included in the EVPT (see Appendix 2).

Vocabulary tests

The effectiveness of students' VL was assessed using test sets from three metaverse-based sessions and three slide-assisted sessions, which included pretests, post-tests, and delayed post-tests. These were adapted from the Wesche and Paribakht (1996) and evaluated by three experts and the instructor of the learning sessions. Each vocabulary assessment package has one pretest, one posttest, and one delayed posttest. Each test comprises 20 items, each with five options, where every item assesses one word or expression acquired throughout each session (see Appendix 3). Furthermore, the vocabulary and phrases assessed were derived only from the three instructional sessions. To mitigate order effects and diminish test familiarity bias, the testing sequences of the vocabulary in pretests, posttests, and delayed posttests were randomized. The exam sets were designed to assess the receptive and productive vocabulary knowledge,

FIGURE 1
MG learners interacting with 3D animated wild animals and learning the corresponding English words under the teacher's instruction and supervision in the first learning session: Animal Kingdom. This classroom activity helps students reinforce form recognition and form recall knowledge.

FIGURE 2
Learners in MG playing a multi-player treasure hunt game in the second learning session: Treasure Island, which facilitates incidental vocabulary learning, collaborative learning and situated learning. The game aims to improve students' word recognition, word recall, meaning recognition and meaning recall knowledge.

FIGURE 3
In the third session, MG participants immersing themselves in a traditional Thai house and playing an item-seeking game, which is designed to enhance learners' meaning recognition and form recall knowledge by reinforcing the psychological links between word spellings and corresponding images.

including form recognition, meaning recognition, form recall and meaning recall, of both the MG and the SG (See Table 1). The fundamental concept of the scale is to assess incremental levels of vocabulary comprehension. Students must evaluate their familiarity with a term or expression using the provided scale and complete the corresponding blank.

Data collection procedure

The complete process required 8 weeks. A pilot study was undertaken during the initial week, resulting in adjustments to the hardware, software, and difficulty levels of the three lessons. The EVPT was administered the subsequent week, and individuals were

FIGURE 4
Students in the MG exploring and learning at a virtual gallery where the spellings of different items related to Thailand and their corresponding images are displayed on the walls, with 3D models positioned in front, which helps learners consolidate their form recognition and meaning recognition knowledge. Subsequently, students are invited to play a word-matching game in the far section, which is designed to improve their meaning recall and form recall knowledge via intentional vocabulary learning.

categorized into two groups (MG and SG) according to the EVPT outcomes. The next 2 weeks consisted of training sessions for MG and SG focused on metaverse-based and slide-assisted learning, respectively, as well as for the instructor in both instructional environments. The educational sessions occurred from the fifth to the seventh week, consisting of three separate vocabulary acquisition sessions. Virtual reality-based learning sessions occurred in a computer laboratory at the designated school, while three slide-assisted instructional sessions were conducted in the original classroom. Each session lasts 90 min, including a 10-min intermission, which is similar to two consecutive normal English lessons at middle schools in China. Before each learning session, all 50 participants were directed to complete a pretest for the forthcoming topic. At the conclusion of each session, all students were directed to promptly complete a post-test regarding the vocabulary acquired during the session. One week following the relevant learning session, a delayed post-test was administered to assess their vocabulary knowledge once more. The time interval between the immediate posttest and delayed posttest was established as one week, as a one-week delayed posttest is commonly utilized by researchers in VR-assisted VL (Lai and Chen, 2021; Fuhrman et al., 2021; Fukushima et al., 2024; Kaplan-Rakowski and Thrasher, 2024; Tai et al., 2022; Luan et al., 2024).

Data analysis

In the present study, the Statistical Package for the Social Sciences (SPSS) version 27.0.1 was used to analyze the data from the quasi-experiment. Specifically, an independent-samples *t*-test was conducted, and both descriptive and inferential analyses were employed to compare the differences in VL effectiveness between two groups.

Findings

The quantitative findings of this study reveal significant differences in VL outcomes between students who engaged with the metaverse-based learning environment and those who participated in slide-assisted instruction. These results are discussed in terms of vocabulary gain, vocabulary retention, and vocabulary forgetting percentage, each offering a distinct perspective on the effectiveness of the respective instructional approaches.

From the descriptive analysis (see Table 2), it is shown that the mean scores of both vocabulary gain (M = 52.56) and retention (M = 49.21) in the MG are larger than those in the SG (M_{VG} = 38.91; M_{VRe} = 33.33), indicating the potential advantage of using metaverse as the VL approach. Additionally, the smaller mean value of VFP in the MG (M = 6.63%) than that in the SG (M = 14.72%) also demonstrates that the metaverse may help learners more effectively retain the learned vocabulary. However, the nuanced differences among the three effectiveness measurements and the superiority of metaverse-based VL cannot be concluded until inferential analysis is introduced.

Moreover, the boxplot graph (Figure 5) reveals that the MG exhibited a lower median VFP, suggesting a reduced vocabulary memory loss compared to the SG. Additionally, the interquartile range (IQR) for MG was narrower, indicating more consistent performance

TABLE 1 Task description and type of knowledge measured.

Level	Task Description	Scoring	Type of Knowledge Measured
1	Recognize that you have never seen the word	1 point	Neither (baseline)
2	Recognize the form but not know meaning	2 points	Form recognition (low-level receptive)
3	Recognize the form + guess/recall meaning	3 points	Meaning recognition (receptive)
4	Know the form + give correct meaning	4 points	Meaning recall (productive in the sense of meaning recall, but still mainly receptive)
5	Know meaning + use in a sentence	5 points	Productive vocabulary knowledge (form recall + meaning recall in active use)

among participants. Several outliers (cases 6, 19, 22, and 23) were observed, but overall, the variability in VFP was lower than that of SG. In contrast, the SG demonstrated a higher median VFP and a larger IQR, indicating greater variability and a tendency toward higher vocabulary forgetting. Notably, an extreme outlier (case 49; $MeanVFP_{49} = 46.84\%$) in the SG group suggests that some participants experienced significant vocabulary loss.

In order to compare two groups in vocabulary gain, retention, and VFP, an independent samples t-test was conducted (see Table 3). In terms of vocabulary gain, students in the MG demonstrated significantly greater improvement from pretests to immediate posttests in all three vocabulary lessons compared to their peers in the slide group. Specifically, Lesson 1 revealed a statistically significant gain difference ($p = 0.002$), with MG students showing a notably higher short-term acquisition of vocabulary. This difference became more pronounced in Lesson 2 ($p < 0.001$) and remained significant in Lesson 3 ($p < 0.001$). The mean vocabulary gain across all three lessons was also significantly higher for the MG than the slide-assisted group ($p < 0.001$), confirming the overall effectiveness of the metaverse-based approach in enhancing vocabulary gain. Similarly, the results for vocabulary retention also favored the MG. All three lessons demonstrated significant differences (Lesson 1: $p = 0.002$; Lesson 2: $p < 0.001$; Lesson 3: $p < 0.001$), with higher retention scores among students in the MG. The overall mean vocabulary retention was also significantly greater for the MG. The between-group comparison yielded a statistically significant result, $t\,(48) = 5.168$, $p < 0.001$.

Of particular note is the analysis of VFP, which provides a relative indicator of the sustainability of learning (See Table 3). The MG exhibited a significantly lower mean VFP than the slide-assisted group, indicating that a smaller proportion of the vocabulary learned was lost between the immediate and delayed posttests. The difference in forgetting percentages was statistically significant, $t(48) = -2.22$, $p = 0.031$. Although the mean VFP further corroborated these patterns, the analysis of VFPs across specific lessons exhibited some noteworthy differences. Specifically, while the differences in forgetting percentage between groups were not significant in the first two lessons ($p_1 = 0.424$; $p_2 = 0.476$), a significant difference emerged in the third session ($p_3 = 0.001$), which was distinctive from the previous findings in both vocabulary gain and retention, with all comparisons yielding statistically significant results ($p < 0.05$), suggesting that the benefits of metaverse may accumulate over time or with repeated exposure (Table 3).

Discussion and conclusion

This study aimed to formulate the VFP as a novel and informative metric to evaluate the effectiveness of L2VL and validate it, particularly in comparing the impact of the metaverse-based learning approach with traditional slide-assisted learning among middle school English learners. Defined as the normalized percentage of vocabulary forgotten within a given time interval relative to the initial vocabulary gain immediately after the learning intervention in the metric formulation phase, VFP offers a normalized and interpretable measure of long-term VL effectiveness. In the validation phase of this study, by analyzing learners' VG, VRe, and VFP, the study offers a more comprehensive and comparative lens for assessing both the extent and durability of VL.

Consistent with prior studies (Chen and Yuan, 2023; Lai and Chen, 2021; Sahinler et al., 2023; Tai et al., 2022), the results demonstrate that the MG significantly outperformed the traditional slide-assisted group in both vocabulary gain and vocabulary retention. Specifically, the MG achieved a higher mean vocabulary gain ($M = 52.56$) than the slide group ($M = 38.91$) and also demonstrated superior long-term retention ($M = 49.21$ vs. 33.33). These findings suggest that immersive, interactive, and collaborative environments like the metaverse facilitate both immediate vocabulary gain and durable retention, possibly due to enhanced learner engagement (Çelik and Baturay, 2024), multisensory input (Jiao et al., 2024), and contextualized usage (Taguchi and Zhao, 2025). The significant enhancement in both vocabulary gain and retention across all sessions ($p < 0.05$) reinforces the hypothesis that metaverse-based learning environments provide not just novelty but sustained engagement and cognitive support (Makransky and Petersen, 2021).

Notably, the late-emerging difference in VFP highlights the importance of extended exposure to immersive learning environments, suggesting that the metaverse-based learning may require a threshold of interaction before its full benefits become apparent. This supports the notion that short-term exposure to the XR-related technologies may not be sufficient to manifest retention advantages, but continued engagement enables the encoding and consolidation of vocabulary into long-term memory (Ersanli, 2023; Kaplan-Rakowski and Thrasher, 2024; Lai and Chen, 2023; Xie et al., 2019). Moreover, the embodied nature of metaverse-based learning—where learners interact with virtual environments—may facilitate deeper cognitive processing and memory retention. Embodied cognition theories suggest that learning is grounded in sensory and motor experiences, and the metaverse provides a platform for such embodied interactions (Johnson-Glenberg, 2018; Macedonia, 2019). This aligns with previous findings that highlight the importance of active engagement and multimodal input in second language vocabulary acquisition (Çelik and Baturay, 2024; Glenberg and Kaschak, 2002; Mayer, 2014).

In conclusion, these findings not only validate the VFP formula as a meaningful and nuanced complement to existing evaluation metrics but also confirm the pedagogical potential of the metaverse

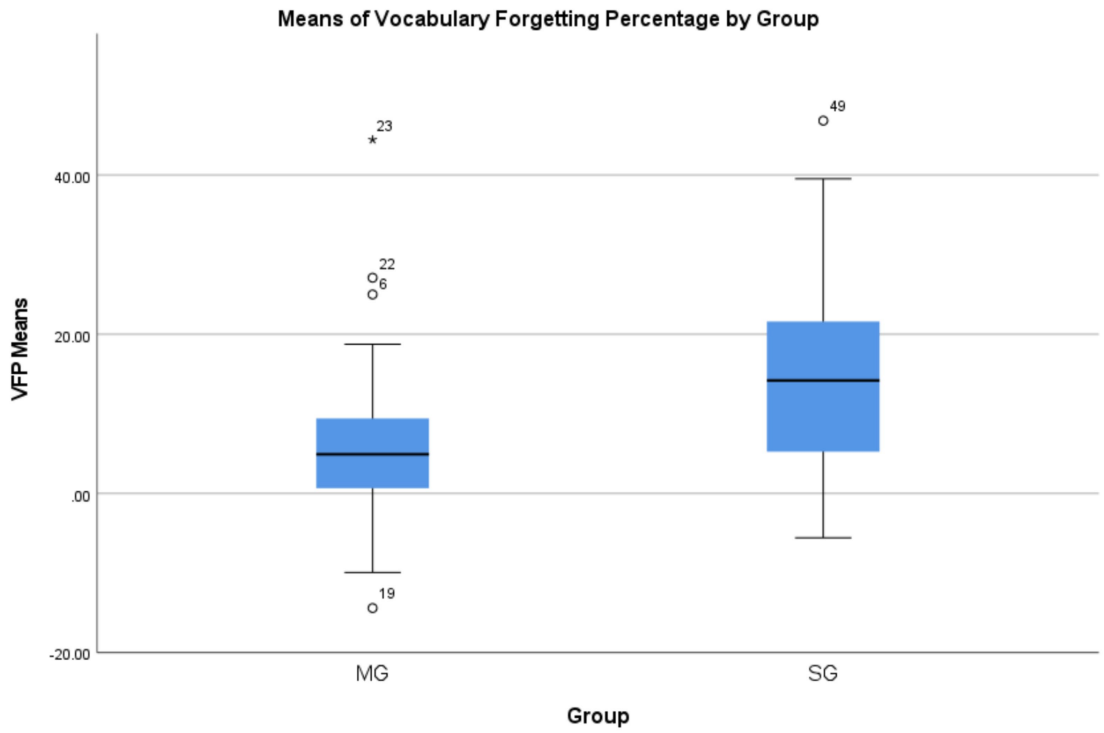

FIGURE 5
Means of vocabulary forgetting percentage by group.

TABLE 2 Descriptive statistics analysis of comparing MG and SG in vocabulary gain, retention, and VFP.

VL by metrics	Group	*N*	Mean	Std. deviation	Std. error mean
Vocabulary gain mean	MG	25	52.56	9.39894	1.87979
	SG	25	38.9067	10.63995	2.12799
Vocabulary retention mean	MG	25	49.2133	10.99153	2.19831
	SG	25	33.3333	10.73589	2.14718
VFP mean (%)	MG	25	6.6267	12.30613	2.46123
	SG	25	14.7187	13.44236	2.68847

for vocabulary instruction. Vocabulary gain and retention, which are expressed as raw scores, are independent from each other, hence making it difficult to compare across different learners due to the lack of control over individual variations of immediate vocabulary gain. The VFP, on the other hand, reflects the L2 learners' VL effectiveness from a more comparative perspective. By quantifying vocabulary loss proportionally, the proposed VFP metric enables researchers and educators to better assess learning efficiency and retention sustainability in diverse L2 instructional contexts. Additionally, the validation approach enables a robust examination of the VFP formula's reliability, generalizability, and practical applicability, ensuring that the proposed metric can serve as a useful tool for evaluating the effectiveness of various L2VL approaches from a long-term retention perspective. Also, it demonstrates that the metaverse can play a pivotal role in vocabulary acquisition, particularly when learning involves repeated exposure over time. Future research should further explore the threshold and mechanisms by which the metaverse contributes to durable vocabulary knowledge, perhaps considering variables such as interactivity, individual learner differences, and the nature of the target vocabulary. Overall, these findings contribute to the growing body of evidence supporting the integration of the metaverse in language education, offering practical insights for curriculum designers and educators aiming to foster deeper, more durable vocabulary learning.

Despite its potential, the proposed VFP metric also presents certain limitations. First, the formula is sensitive to learners' initial posttest performance; if the immediate posttest score is low, even a small amount of forgetting can result in a high VFP, potentially exaggerating retention loss. Therefore, it is advisable to apply the formula only when the initial learning outcome exceeds a predetermined threshold, and also it is recommended to ensure the number of test items is not too small to ensure that the total assessment score is not too low. Second, the VFP can only report a learner's forgetting trend at a limited number of time points. When

TABLE 3 Independent *t*-test of comparing MG and SG in vocabulary gain, retention, and VFP.

VL by metrics	*t*	*p*	Mean Difference	95% Confidence Interval	
				Lower	Upper
Vocabulary gain 1	3.358**	0.002	11.64	4.66993	18.61007
Vocabulary gain 2	5.48***	0	19.64	12.43437	26.84563
Vocabulary gain 3	3.531**	0.001	9.68	4.16832	15.19168
Vocabulary gain mean	4.809***	0	13.65333	7.94443	19.36224
Vocabulary retention 1	3.256**	0.002	13.16	5.03289	21.28711
Vocabulary retention 2	5.842***	0	19.56	12.82773	26.29227
Vocabulary retention 3	5.035***	0	14.92	8.96228	20.87772
Vocabulary retention mean	5.168***	0	15.88	9.70145	22.05855
VFP 1	−0.806	0.424	−5.38848	−18.82927	8.0523
VFP 2	0.718	0.476	9.02298	−16.2589	34.30485
VFP 3	−3.383**	0.001	−14.71851	−23.46681	−5.97021
VFP Mean	−2.22*	0.031	−8.09202	−15.42064	−0.76339

* < 0.05; ** < 0.01; *** < 0.001.

conditions permit, it is suggested to administer multiple delayed post-tests to more accurately capture the forgetting patterns and characteristics of vocabulary learners across different time intervals. Third, the length of the retention interval should also be standardized or carefully reported when comparing different studies, as varying intervals can significantly influence the degree of forgetting observed.

Data availability statement

Data and supplementary files, including appendices and tables are available online at: 10.6084/m9.figshare.29880485.

Ethics statement

The studies involving humans were approved by Research Ethics Committee, Universiti Kebangsaan Malaysia. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

MZ: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. HH: Funding acquisition, Supervision, Writing – review & editing. MM: Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. We used ChatGPT to correct grammatical and syntactical errors, ensuring clarity and precision of the language. However, the originality and innovation of the ideas presented in this study are entirely the product of our team's creativity and expertise. AI was used solely as a supportive tool to enhance the presentation of our concepts, not to generate the ideas themselves.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1641638/full#supplementary-material

References

Afzal, N. (2019). A study on vocabulary-learning problems encountered by BA English major students at the university level of education. *Arab World English J.* 10, 81–98. doi: 10.24093/awej/vol10no3.6

Alfadil, M. (2020). Effectiveness of virtual reality game in foreign language vocabulary acquisition. *Comput. Educ.* 153:103893. doi: 10.1016/j.compedu.2020.103893

Altiner, C. (2019). Integrating a computer-based flashcard program into academic vocabulary learning. *J. Lang. Linguist. Stud.* 15, 262–273.

Arochman, T., Madani, S., Welasiyah, S., and Setiandari, R. (2023). Exploring students' difficulties in memorizing English vocabularies in higher education. *J. English Lang. Educ.* 8:2023. doi: 10.31004/jele.v8i2.430

Bahrick, H. P., Bahrick, P. O., Bahrick, L. E., and Bahrick, A. S. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychol. Sci.* 4, 316–321. doi: 10.1111/j.1467-9280.1993.tb00571.x

Barcroft, J. (2004). Second language vocabulary acquisition: a lexical input processing approach. *Foreign Lang. Ann.* 37, 200–208. doi: 10.1111/j.1944-9720.2004.tb02193.x

Barcroft, J. (2009). Effects of synonym generation on incidental and intentional L2 vocabulary learning during reading. *TESOL Q.* 43, 79–103. doi: 10.1002/j.1545-7249.2009.tb00228.x

Brown, J. I., Fishco, V. V., and Hanna, G. S. (1993). *Nelson-Denny Reading test: Manual for scoring and interpretation*: Riverside Publishing.

Cambridge Assessment English (2020). *B1 preliminary (PET) handbook for teachers*: Cambridge University Press.

Çelik, F., and Baturay, M. H. (2024). The effect of metaverse on L2 vocabulary learning, retention, student engagement, presence, and community feeling. *BMC Psychol.* 12:58. doi: 10.1186/s40359-024-01549-4

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., and Rohrer, D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychol. Bull.* 132, 354–380. doi: 10.1037/0033-2909.132.3.354

Chen, C., and Yuan, Y. (2023). Effectiveness of virtual reality on Chinese as a second language vocabulary learning: perceptions from international students. *Comput. Assist. Lang. Learn.* 1–29. doi: 10.1080/09588221.2023.2192770

Crossley, S. A., Salsbury, T., and McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymy relations. *Lang. Learn.* 59, 307–334. doi: 10.1111/j.1467-9922.2009.00508.x

Dunn, L. M., and Dunn, D. M. (2007). *Peabody picture vocabulary test. 4th* Edn: Pearson Assessments.

Dunn, L. M., Dunn, L. M., Whetton, C., and Burley, J. (1997). *British picture vocabulary scale* (2nd ed.). NFER-Nelson.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*: Teachers College, Columbia University (Original work published 1885).

Elekaei, A., Tabrızı, H. H., and Chalak, A. (2020). Evaluating learners' vocabulary gain and retention in an e-learning context using vocabulary podcasting tasks: a case study. *Turk. Online J. Distance Educ.* 21, 190–203. doi: 10.17718/tojde.728162

Elgort, I., and Nation, P. (2010). Vocabulary learning in a second language: knowledge of collocations. *TESOL Q.* 44, 391–420. doi: 10.5054/tq.2010.222223

Ersanli, C. Y. (2023). The effect of using augmented reality with storytelling on young learners' vocabulary learning and retention. *Novitas-ROYAL* 17, 62–72.

Faramarzi, S., Elekaei, A., and Koosha, M. (2014). On the impact of multimedia glosses on reading comprehension, vocabulary gain and vocabulary retention. *Int. J. Lang. Learn. Appl. Linguist. World* 6, 623–634.

Feng, B., and Ng, L.-L. (2024). The spatial influence on vocabulary acquisition in an immersive virtual reality-mediated learning environment. *Int. J. Comput.-Assist. Lang. Learn. Teach.* 14, 1–17. doi: 10.4018/IJCALLT.339903

Fuhrman, O., Eckerling, A., Friedmann, N., Tarrasch, R., and Raz, G. (2021). The moving learner: Object manipulation in VR improves vocabulary learning. *PsyArXiv*. doi: 10.31234/osf.io/wgzxu

Fukushima, S., Sakamoto, K., and Nakamura, Y. (2024). Naritan: enhancing second language vocabulary learning through non-human avatar embodiment in immersive virtual reality. *Multimodal Technol. Interact.* 8:93. doi: 10.3390/mti8100093

Ghalebi, S., Sadighi, F., and Bagheri, M. (2020). Vocabulary learning strategies: a comparative study of EFL learners. *Cogent Psychol.* 7:1824306. doi: 10.1080/23311908.2020.1824306

Glenberg, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565. doi: 10.3758/BF03196313

Govindasamy, P., Md Yunus, M., and Hashim, H. (2019). Mobile-assisted vocabulary learning: examining the effects on students' vocabulary enhancement. *Univ. J. Educ. Res.* 7, 85–92. doi: 10.13189/ujer.2019.071911

Hung, H.-T., and Yeh, H.-C. (2023). Augmented-reality-enhanced game-based learning in flipped English classrooms: effects on students' creative thinking and vocabulary acquisition. *J. Comput. Assist. Learn.* 39, 1786–1800. doi: 10.1111/jcal.12839

Jiao, L., Zhu, M., Xu, Z., Zhou, G., Schwieter, J. W., and Liu, C. (2024). An ERP study on novel word learning in an immersive virtual reality context. *Bilingualism Lang. Cogn.* 27, 128–136. doi: 10.1017/S136672892300038X

Johnson-Glenberg, M. C. (2018). Embodied education and mixed reality: how gesture and motion capture affect learning. *Educ. Psychol. Rev.* 30, 377–389. doi: 10.1007/s10648-017-9411-9

Kaplan-Rakowski, R., and Thrasher, A. (2024). The impact of high-immersion virtual reality and interactivity on vocabulary learning. *Br. J. Educ. Technol.* doi: 10.1111/bjet.13603

Kornmeier, J., Spitzer, B., and Baur, S. (2022). The effects of spacing on vocabulary learning and forgetting: evidence from an EEG study. *J. Appl. Res. Mem. Cogn.* 11, 32–44. doi: 10.1016/j.jarmac.2021.12.001

Lai, C., and Chen, G. (2021). The impact of digital flashcards on vocabulary retention among university students. *Comput. Assist. Lang. Learn.* 34, 234–250. doi: 10.1080/09588221.2020.1744663

Lai, K.-W. K., and Chen, H.-J. H. (2023). A comparative study on the effects of a VR and PC visual novel game on vocabulary learning. *Comput. Assist. Lang. Learn.* 36, 312–345. doi: 10.1080/09588221.2021.1928226

Laufer, B., and Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Lang. Test.* 16, 33–51. doi: 10.1177/026553229901600103

Lee, S.-M. (2023). Factors affecting incidental L2 vocabulary acquisition and retention in a game-enhanced learning environment. *ReCALL* 35, 274–289. doi: 10.1017/S0958344022000209

Lessard-Clouston, M. (2021). *Vocabulary and the four skills: Pedagogy, practice, and implications*: Routledge.

Lin, H., and Wei, W. (2024). A systematic review on vocabulary learning in AR and VR gamification context. *Comput. Educ. X Real.* 4:100057. doi: 10.1016/j.cexr.2024.100057

Luan, L., Hwang, G.-J., Yi, Y., Lu, Z., and Jing, B. (2024). The effects of a selfdeveloped virtual reality environment on college EFL learners' vocabulary learning. *Interact. Learn. Environ.* 33, 335–346. doi: 10.1080/10494820.2024.2344056

Macedonia, M. (2019). Embodied learning: why at school the mind needs the body. *Front. Psychol.* 10:2098. doi: 10.3389/fpsyg.2019.02098

Makransky, G., and Petersen, G. B. (2021). The cognitive affective model of immersive learning (CAMIL): a theoretical research-based model of learning in immersive virtual reality. *Educ. Psychol. Rev.* 33, 937–958. doi: 10.1007/s10648-020-09586-2

Mayer, R. E. (2014). "Cognitive theory of multimedia learning" in *The Cambridge handbook of multimedia learning*. ed. R. E. Mayer. *2nd* ed (Cambridge University Press), 43–71.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Mohammed, E. F. (2009). *The effectiveness of TPRS in vocabulary acquisition and retention of EFL students and their attitude toward English language. MA thesis*. Egypt: Marsoura University.

Murre, J. M. J., and Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLoS One* 10:e0120644. doi: 10.1371/journal.pone.0120644

Mystakidis, S. (2022). Metaverse. *Encyclopedia* 2, 486–497. doi: 10.3390/encyclopedia2010031

Nation, I., and Beglar, D. (2007). A vocabulary size test. *Lang. Teach.* 31, 9–13.

Nation, I. S. P. (2001). *Learning vocabulary in another language*: Cambridge University Press.

Ng, L. W., and Rosli, M. N. (2023). Vocabulary size and lexical richness among ESL learners in Malaysia. *Int. J. Acad. Res. Bus. Soc. Sci.* 13, 1451–1465. doi: 10.6007/IJARBSS/v13-i2/16285

Okyar, H., and Çakır, A. (2019). Effects of different reading texts on vocabulary gain, use and retention. *J. Lang. Linguist. Stud.* 15, 111–122. doi: 10.17263/jlls.547634

Oxford, R., and Crookall, D. (1990). Vocabulary learning: a critical analysis of techniques. *TESL Can. J.* 7, 9–30. doi: 10.18806/tesl.v7i2.566

Oxford University Press. (2014). Oxford Young Learners Placement Test. Oxford University Press

Rahman, F., Amalia, T., and Lutfi, M. (2021). Reducing forgetting rate in EFL students using a spaced repetition-powered digital game-based learning application. Open Science Framework

Read, J. (2000). *Assessing vocabulary*: Cambridge University Press.

Reynolds, B. L., Xie, X. S., and Pham, Q. H. P. (2022). Incidental vocabulary acquisition from listening to English teacher education lectures: A case study from Macau higher education. *Front. Psychol.* 13:993445. doi: 10.3389/fpsyg.2022.993445

Richards, C. J., and Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics*. London: Pearson Education.

Rivera-Lares, K., Sala, S. D., Baddeley, A., and Logie, R. (2023). Rate of forgetting is independent from initial degree of learning across different age groups. *Q. J. Exp. Psychol.* 76, 1672–1682. doi: 10.1177/17470218221128780

Sahinler, M. (2023). Gamifying vocabulary acquisition and retention in virtual reality. *Teaching English with Technology.* 21, 42–57. doi: 10.56297/BKAM1691/DFXC4759

Schmitt, N. (2000). *Vocabulary in language teaching*: Cambridge University Press.

Schmitt, N. (2008). Review article: instructed second language vocabulary learning. *Lang. Teach. Res.* 12, 329–363. doi: 10.1177/1362168808089921

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*: Palgrave Macmillan.

Schmitt, N., Schmitt, D., and Clapham, C. (2001). Developing and exploring the behavior of two new versions of the vocabulary levels test. *Lang. Test.* 18, 55–88. doi: 10.1177/026553220101800103

Sense, F., Behrens, F., Meijer, R. R., and van Rijn, H. (2018). An individual differences approach to the rate of forgetting. *Front. Educ.* 3:112. doi: 10.3389/feduc.2018.00112

Soderstrom, N. C., and Bjork, R. A. (2015). Learning versus performance: an integrative review. *Perspect. Psychol. Sci.* 10, 176–199. doi: 10.1177/1745691615569000

Stewart, J., Gyllstad, H., Nicklin, C., and McLean, S. (2024). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Lang. Test.* 41, 89–108. doi: 10.1177/02655322231162853

Suleiman, H. M. (2009). *Implicit and explicit vocabulary acquisition with a computerassisted hypertext reading task, comprehension, and retention*. MA thesis. Arizona: University of Arizona.

Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., and Gomez-Rodriguez, M. (2019). Enhancing human learning via spaced repetition optimization. *Proc. Natl. Acad. Sci. USA* 116, 3988–3993. doi: 10.1073/pnas.1815156116

Taguchi, N., and Zhao, S. (2025). Can virtual reality leverage technology-mediated task-based language teaching?: a research synthesis. *Digital Stud. Lang. Liter.* 2, 25–53. doi: 10.1515/dsll-2024-0028

Tai, T. Y., Chen, H. H. J., and Todd, G. (2022). The impact of a virtual reality app on adolescent EFL learners' vocabulary learning. *Comput. Assist. Lang. Learn.* 35, 892–917. doi: 10.1080/09588221.2020.1752735

Teymouri, R. (2024). Recent developments in mobile-assisted vocabulary learning: a mini review of published studies focusing on digital flashcards. *Front. Educ.* 9:1496578. doi: 10.3389/feduc.2024.1496578

Wang, Z., Zou, D., Peng, P., Wang, F. L., Lee, L. K., and Xie, H. (2025). Effects of mobile metaverse-based vocabulary learning on learners' perception and performance: a case study of Chinese EFL learners. *J. Comput. Educ.* Advance online publication. doi: 10.1007/s40692-024-00348-5

Webb, S. (2005). Receptive and productive vocabulary learning: the effects of reading and writing on word knowledge. *Stud. Second. Lang. Acquis.* 27, 33–52. doi: 10.1017/S0272263105050023

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Appl. Linguist.* 28, 46–65. doi: 10.1093/applin/aml048

Weng, Y., Schmidt, M., Huang, W., and Hao, Y. (2024). The effectiveness of immersive learning technologies in K–12 English as a second language learning: a systematic review. *ReCALL* 36, 210–229. doi: 10.1017/S0958344024000041

Wesche, M., and Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: depth versus breadth. *Can. Mod. Lang. Rev.* 53, 13–40. doi: 10.3138/cmlr.53.1.13

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.* 55, 235–269. doi: 10.1146/annurev.psych.55.090902.141555

Xie, H., Chu, H.-C., Hwang, G.-J., and Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: a systematic review of journal publications from 2007 to 2017. *Comput. Educ.* 140:103599. doi: 10.1016/j.compedu.2019.103599

Zhang, M. M., Hashim, H., and Yunus, M. M. (2025). Analyzing and comparing augmented reality and virtual reality assisted vocabulary learning: a systematic review. *Front. Virtual Real.* 6:1522380. doi: 10.3389/frvir.2025.1522380

Zhong, Q. M. (2018). Vocabulary learning through reading: the effects of three learning conditions. *TESL Can. J.* 35, 92–108. doi: 10.18806/tesl.v35i1.1280

Zhou, Y. (2010). The effect of L1 translation on L2 incidental vocabulary learning in reading. *Read. Matrix* 10, 135–151.