Lag effects for foreign language vocabulary learning through Quizlet

Jonathan Serfaty, Universitat de Barcelona, Université Côte d’Azur
Raquel Serrano, Universitat de Barcelona

Abstract

Digital flashcard apps allow students to learn and practice foreign language vocabulary independently and efficiently, leaving more classroom time for communicative activities. However, words learned this way may be forgotten. Previous lab studies have shown that vocabulary retrieval practice can be optimized for long-term memory by employing longer intersession intervals, but this lag effect has not been shown in classroom conditions. The present study investigated the optimal gap between two Quizlet sessions for retaining new vocabulary. Secondary-school students (N = 96, mean age = 13.44) learned 16 novel words in an unknown language with either a 1-day or 1-week interval. Their productive and receptive knowledge was tested after seven or 28 days. Results showed that longer spacing was beneficial for vocabulary retention, contrary to previous findings reported with school-aged learners using other types of training. The effect was small, but significantly larger on receptive tests, suggesting that the lag effect depends upon the kind of knowledge being tested.

Keywords: Lag Effect, Vocabulary, Retrieval, Productive, Receptive Knowledge

Language(s) Learned in This Study: Hebrew


Introduction

One challenge in learning a foreign language (FL) is that the number of hours of exposure tends to be limited and often restricted to the classroom (Lightbown, 2014). Moreover, FL vocabulary is susceptible to forgetting if not sufficiently practiced (Pavlik & Anderson, 2005). Considering this, it is crucial to investigate how to optimize this limited time for the best long-term retention of knowledge.

Research has shown that vocabulary retrieval practice is an efficient way to learn new words (Elgort, 2011; Fitzpatrick et al., 2008; Nation, 2001; Webb, 2009). This can be accomplished using flashcards: Learners see a cue, such as a first language (L1) translation, and attempt to retrieve the target word from memory, or vice versa. In contrast to traditional paper flashcard drills, digital flashcards offer a wide range of features to foster deeper processing. For example, they can be used to elicit written output with tailored feedback, test items until they have been produced correctly within a session, provide audio to clarify pronunciation, and motivate students through gamification. Flashcard sets can also be assigned as homework to reduce classroom time devoted to vocabulary teaching, facilitating fluency and comprehension in subsequent classroom activities that require the target words.

Even when engaging in this type of activity, students are likely to forget words learned in a single session. However, research has shown that repeating sessions on multiple days has a powerful effect on long-term memory, even when controlling for the amount of time on task (Rawson et al., 2018), and that the optimal distribution of these relearning sessions can enhance retention further (Gerbier & Toppino, 2015). Several studies from cognitive psychology have shown that longer intervals between sessions promote long-term retention more than shorter intervals (Cepeda et al., 2006; Cepeda et al., 2009), known as the lag effect.
However, this lag effect has not been consistently found in FL research. Some studies involving grammar learning (Kasprówicz et al., 2019; Suzuki, 2017; Suzuki & DeKeyser, 2017) and vocabulary learning in the classroom with children (Rogers & Cheung, 2020, 2021) and teenagers (Küpper-Tetzel et al., 2014; Serrano & Huang, 2018, 2021) have reported no advantage to a longer interval. On the other hand, the lag effect has been reported for vocabulary learning from lab studies involving the retrieval of paired associates (Bahrick, 1979; Li & DeKeyser, 2019), which is the method employed by digital flashcard apps. Therefore, it is feasible that vocabulary learning through digital flashcards would also be optimized with longer lags between sessions. However, no previous study has investigated whether paired-associate retrieval is subject to lag effects in a classroom environment.

In order to shed light on this issue, we conducted a study in which secondary-school students learned novel vocabulary pairs through Quizlet, a popular flashcard app already commonly used in classrooms. Words were learned over two sessions, spaced either one day or one week apart, and tested after either 7 or 28 days. The findings are expected to deepen our understanding of the conditions under which the lag effect applies while also providing guidance on the optimal gap between sessions of vocabulary flashcard practice. This paper further contributes to the field by examining the difference in lag effects on productive and receptive knowledge. While many studies have shown that the former is acquired more slowly than the latter (Laufer & Goldstein, 2004), the potentially differential effect of lag on these two dimensions of knowledge has yet to be explored.

**Literature Review**

**Digital Flashcards for Vocabulary Learning**

Flashcards have traditionally been paper cards designed for self-testing. Retrieving information through testing is known to build memory more than re-reading the same information (Barcroft, 2007; Carrier & Pashler, 1992; Kang, 2010; Kang et al., 2013; Kornell & Vaughn, 2016). Online flashcard apps include useful features, such as smart feedback that highlights the users’ errors, helping them to notice the difference between their attempts and the target (Izumi & Bigelow, 2000; Izumi et al., 1999; Zalbidea, 2019). Flashcard software commonly employs criterion learning, repeating items in a cycle until they are answered correctly. Consequently, more practice is automatically allotted to the difficult items. Survey data has shown that digital flashcards are popular with students in educational contexts (Altiner, 2019; Stroud, 2014; Zung et al., 2022), and Quizlet in particular is widely used by both teachers and researchers (Franciosi et al., 2016; Korlu & Mede, 2018; Sanosi, 2018; Serfaty & Serrano, 2022; Stroud, 2014).

The goal of flashcard assignments might be to familiarize students with useful words. For example, in order to comprehend a text without assistance, most words should already be known (98%; see Hu & Nation, 2000). Teachers could assign content-specific vocabulary in preparation for an upcoming reading or listening passage (Webb, 2009), allowing more classroom time for comprehension activities or communicative practice. Alternatively, learners could focus on the most frequently occurring words of the language to increase their general listening and reading comprehension (Li & Zhang, 2019; Nation, 2006). Similarly, students could study target academic words to prepare for a course taught in a second language (Coxhead, 2000).

A second reason for using flashcards would be to collect and practice previously encountered words. Without maintenance, declarative knowledge, such as FL vocabulary, decays quickly (Kim et al., 2013; Ullman & Lovelett, 2018), and many curricula do not adequately recycle vocabulary (Tschichold, 2012). Using digital flashcard software, learners or teachers could cumulatively add new words to sets. Regular practice of these sets would prevent forgetting under-used vocabulary items (Nakata et al., 2021).

A third use for digital flashcards would be to provide individualized work to students based on their abilities and interests. Flashcards could be assigned to faster students while the teacher focuses their attention on other students. In cases where teachers cannot be present, for example, during the recent
COVID-19 school closures, students could still engage in output practice with reliable feedback.

Research suggests that productive recall, which involves retrieving the target word from an L1 cue (cue-target), would be best for developing productive knowledge (Nakata, 2016). Although the cue could be something different, like a synonym or an image, L1 translations seem to be more effective for vocabulary learning (Joyce, 2018; Laufer & Shmueli, 1997; Lotto & de Groot, 1998). The reverse direction, receptive recall (target-cue), is better in terms of words learned per minute, but productive recall is best for overall gains, especially when knowledge is measured with productive recall tests (Griffin & Harley, 1996; Nakata & Webb, 2016; Webb, 2005, 2009). Productive practice can also reduce forgetting and retraining time over repeated sessions (Schneider et al., 2002). This may be related to the high levels of effort (Pyc & Rawson, 2009) or user-involvement (Hulstijn & Laufer, 2001) in productive practice. By requiring users to produce the target word with perfect orthography, flashcard training guarantees a significant level of attention to each item, which is a crucial step toward long-term acquisition (Leow, 2015; Schmidt, 1990, 2010).

The Lag Effect in Paired-Associate Learning Under Lab Conditions

As with other domains of learning, it has been shown that distributing vocabulary practice over multiple sessions (spaced) is better for long-term memory than the same amount of practice in a single uninterrupted session (massed), known as the spacing effect. This effect has been reported for a wide variety of knowledge and skills (Cepeda et al., 2006; Donovan & Radosovich, 1999), including for FL vocabulary (Koval, 2019; Nakata, 2015; Nakata & Elgort, 2021) and grammar (Miles, 2014). Within a single session, more spacing between repetitions of the same item has led to better scores in a posttest (Nakata & Webb, 2016). However, whether more spacing between sessions (i.e., a longer intersession interval [ISI]) leads to longer retention is less clear.

Very few studies have tested different ISIs for studying FL vocabulary over multiple days. In a landmark study by Bahrick (1979), participants with no previous exposure to Spanish studied English-Spanish vocabulary pairs to criterion through oral productive recall at ISI-1, ISI-7, or ISI-30 (n = 10 per condition). The shorter interval led to a faster rate of learning over six sessions. The same number of participants performed a second experiment using ISI-1 or ISI-30, over either three or six sessions, with the addition of a 30-day retention interval (RI) after the final learning session. Although the shorter ISI facilitated better retention between sessions again, the ISI-30 group showed considerably higher retention on a final productive recall test. Though limited in terms of sample size, Bahrick’s experiments indicated that longer lags between sessions of criterion learning may be advantageous for retention. Li and DeKeyser (2019) also demonstrated a between-session lag effect for vocabulary retention, including 40 participants per group and more pedagogically relevant intervals. Studying Mandarin words with no previous exposure at ISI-1 (daily) or ISI-7 (weekly) over three sessions through a variety of tasks, retention was similar for both ISIs when tested seven days after training (RI-7), but at RI-28, more words were remembered from the ISI-7 condition.

In contrast, studies that have compared lags at a proportionately shorter RI have not found this effect. Bahrick and Hall (2005) found no difference between ISI-1 and ISI-14 at RI-14. Cepeda et al. (2009), who used a range of ISIs from 0 to 14 days, found no significant differences between ISIs of one day or more for participants tested at RI-10. It has been claimed that the advantage of a longer lag only emerges at a suitably long RI (e.g., Bird, 2010) and that the ISI should be around 10-30% of the RI (Cepeda et al., 2008).

Conflicting findings among previous studies may also be related to the training and testing formats (Edmonds et al., 2021), with some studies using productive recall (e.g., Bahrick, 1979), while others used receptive recall (e.g., Bahrick & Hall, 2005). To our knowledge, no study has specifically examined the differences between a productive and receptive test after different ISIs, which is one of the gaps the present study aims to fill.

The between-session lag effect has been explained by different theoretical accounts. The reminding
account holds that more time between encounters makes retrieval more effortful (Pyc & Rawson, 2009; Koval, 2022). This effort provides desirable difficulty and enhances learning (Bjork, 1994; Suzuki et al., 2019). Alternatively, the reconsolidation account (Smith & Scarf, 2017) focuses specifically on multi-day ISIs and explains the advantage of a longer lag through a greater degree of consolidation. When retrieved, a more consolidated memory trace is more effectively reconsolidated. Both of these accounts hold that if an item is completely forgotten, knowledge cannot be reinforced. It is therefore desirable to schedule a second session with the longest possible ISI before an item cannot be retrieved. A shorter ISI would allow more items to be retrieved, but a longer ISI makes retrievable items more durable.

The Lag Effect in Vocabulary Classroom Studies

Limited research has also addressed lag effects for FL vocabulary learning in the classroom, and, to our knowledge, no advantage to a longer lag has been reported. Examining assisted repeated reading for English vocabulary among 16-year-olds in Taiwan, Serrano and Huang (2018) found similar results from ISI-1 and ISI-7 on incidental vocabulary learning and an advantage to ISI-1 in a partial replication involving intentional learning (Serrano & Huang, 2021). Küpper-Tetzel et al. (2014) found ISI-1 and ISI-10 to both be better than massed learning for 11-to-13-year-olds studying English vocabulary in Germany, with no significant differences between the two ISI conditions at the delayed posttest. Rogers and Cheung (2020, 2021) examined the learning of English vocabulary among 8-9-year-old children in Hong Kong. The studies found no benefit for the longer lag (ISI-8), with a slight advantage to the shorter lag (ISI-1) in one study. All of these classroom studies used ISIs within 10-30% of the RI, so a lag effect could have been expected.

There are elements of the above-mentioned classroom studies’ methodologies that differ significantly from digital flashcard learning. Firstly, flashcards employ criterion learning. Incorrectly answered items remain in the cycle to be attempted again in a subsequent round of retrieval attempts. The session only ends when all items have been retrieved successfully. Therefore, a repeated session serves to remind learners of already-learned knowledge. In contrast, the classroom studies controlled for the amount of practice time but not for the achievement of the learner within a session. Some learners may not have fully learned all of the words in their first session, making it difficult to classify their second session as a relearning event. Secondly, the classroom studies were interactive, involving multiple learners and an instructor, as opposed to online flashcards that involve one learner guided by software. The classroom studies also used a variety of training tasks, even within studies (e.g., picture quizzes, animations) as well as a variety of testing formats (e.g., vocabulary matching test, crossword puzzles). These methodological factors may lead to less experimental control and less comparability between studies. For all of these reasons, it would be interesting to conduct a lab-like methodology within a classroom setting. Flashcard software, which is already commonly used in FL classrooms, provides this opportunity.

The scheduling of FL vocabulary learning through software under classroom conditions has been investigated for university students (Schuetze, 2015; Schuetze & Weimer-Stuckmann, 2011). However, the software in these studies prompted students to copy words rather than to retrieve them, and the schedule differed only in terms of an expanding versus a uniform ISI, rather than the length of the ISI itself. No studies have yet provided insights into the optimal ISI between different sessions of vocabulary flashcard training using common educational software with secondary-school students using productive recall. The only previous study in this area used full-sentence items as flashcards with the aim of learning grammatical accuracy in English (Serfaty & Serrano, 2022). In this case, the longer ISI-7 was only better for students with high English proficiency and fast completion times, adding desirable difficulty for learners whose abilities were well-suited to this task. ISI-1, however, was better for participants who found the task more challenging, as evidenced by their lower English proficiency and longer completion times. While grammar learning involves applying the same target rule repeatedly through each item, vocabulary learning involves the independent practice of multiple unrelated target items. This difference in the relationship between items means that, within one set, the grammar target is practiced repeatedly, whereas vocabulary targets are retrieved only once. Therefore, it is unclear to what extent findings from
grammar learning would apply to vocabulary learning.

**The Present Study**

The present study explored lag effects for FL vocabulary learning through digital flashcards. Participants retrieved new words over two sessions, with either one or seven days between sessions, and their retention was assessed after one week or one month. The study represents an important contribution to the field for several reasons. First, the treatment consists of paired-associate learning of vocabulary from an unknown language studied individually, as in typical lab studies with adults. However, the methodological design approximated a classroom environment in the following ways: (a) the participants were secondary school students, (b) the experiment was conducted during normal classes, in their normal classroom environment, and supervised by their regular teacher (first author), and (c) the tool used was Quizlet, a familiar and popular app already regularly used by these students. This design will clarify whether the lag effect applies to this demographic in this setting while still controlling for prior knowledge, extra-experimental exposure, and interactional factors. Second, in contrast to previous research in this area, the present study used both a productive and a receptive recall test. The former taps the ability to generate the target word in speaking or writing, whereas the latter only tests the ability to comprehend the target word when it is encountered through listening or reading. Receptive knowledge is known to develop before productive knowledge and therefore represents a lower level of mastery of the target word (González-Fernández & Schmitt, 2020). In order to disentangle the effect of this potentially confounding factor, it is imperative to explore how lag effects could vary between these two kinds of knowledge. Finally, although not a specific research question in this paper, our experiment used the same design as a grammar experiment (Serfaty & Serrano, 2022) in terms of the participants, setting, tool, ISI, RI, and number of items. Consequently, a valid comparison of lag effects for grammar and vocabulary learning can be made without the confounds of task differences. Our research questions are as follows:

1. Is there a lag effect for vocabulary learning through digital flashcards for secondary school students?
2. Is the lag effect different for productive and receptive vocabulary knowledge?

For the first research question, we hypothesized that a lag effect would be found, despite it not being found in previous classroom vocabulary studies, because paired-associate learning to criterion has produced a lag effect under lab conditions with the intervals used in this study. No hypothesis was made for the second research question since this issue has not been previously explored.

**Methodology**

**Participants**

Participants came from an English-language international school in Cambodia. All students aged 11-18 were recruited for the training phase of this study on a voluntary basis, as part of a project to teach students how to study independently. In particular, these students took monthly exams in all subjects for their local curriculum and spent a lot of time reading copied passages from textbooks in order to memorize information for these exams. It was hoped that by learning more efficient studying techniques, including the use of flashcard apps, their revision time and stress could be reduced. This project was conducted during English classes in their international curriculum.

All students initially chose to participate, but around half missed at least one session due to COVID-19 closures and were therefore excluded from the analysis. Any students that failed to document their learning as required were also excluded, leaving a total of 96 participants (51 female). The distribution of ages was as follows: \( n_{11} = 20; n_{12} = 16; n_{13} = 15; n_{14} = 13; n_{15} = 16; n_{16} = 12; n_{17-18} = 4. \)
Experimental Design

Target words. The priority in this experiment was to use target words that were previously unknown to all participants and to be sure that gains could be solely attributed to the experimental training; hence English was discarded as the target language. Hebrew was chosen because it contains many words with two syllables and with phonology common to English (the students’ school language) and Khmer (their L1). The categories of animals and food were chosen for their high imageability and familiarity in English for these students, whose reading was more advanced in English than in Khmer. Each category included eight nouns of five or six letters (e.g., kelev - dog), transliterated into the Latin alphabet (Appendix A), each word with three consonants and two vowels. All words had a CVCVC structure, with two slight deviations (glida/ice cream: CCVCV & arnav/rabbit: VCCVC) included to add a variety of meanings.

Training. Using the Write mode of Quizlet, participants saw an English cue (e.g., dog) with an image, and typed their response in Hebrew. Since there was no presentation stage, participants needed to guess incorrectly during Round 1 in order to see the target words for the first time as feedback. They then continued through the rounds until they had typed all items correctly once (see Appendix B for screenshots).

There were three training sessions (S). Half of the participants studied animals in the first session (S1) and food in S2, while the other half did the reverse. In S3, participants studied a combined set of all the target words. S1 and S3 were separated by one week (ISI-7), while S2 and S3 were separated by one day (ISI-1). These intervals were chosen to foster comparability with previous research and due to their relevance to pedagogical schedules. Figure 1 shows the timing of each session.

Tests

Participants were tested on all 16 items using Google Forms, firstly through productive recall (English-Hebrew translation) and then through receptive recall (Hebrew-English translation), as defined by Nakata (2020). This format was similar to the training stage but without feedback. Since the possible answers for the receptive test were used as cues in the productive test, some priming was unavoidable. To mitigate these effects, a distraction round of three unrelated questions preceded the receptive test, and the order of items was randomized in both sections. In any case, being reminded of the 16 possible answers in English would not enable participants to match them to their Hebrew associates. Cronbach’s alpha showed high internal consistency for animals and food for productive (.815; .830) and receptive (.746; .792) measures.

To avoid testing effects, RI was a between-subjects variable (Suzuki, 2017). RI-7 and RI-28 were chosen based on their relevance to real school schedules and for comparability with previous research using the same intervals. Based on claims that the optimal ISI is 10-30% of the RI (Cepeda et al., 2008), ISI-1 would be optimal for RI-7, and ISI-7 would be optimal for RI-28.

Figure 1

Experimental Design

Procedure

Participants were split alphabetically within grade levels to assign categories to ISIs. RI groups were
manipulated after training so that the order of categories was equally represented at each RI, with no statistically significant differences in mean age or time taken to complete the training (see Appendix C).

All sessions were conducted under conditions in which students would normally complete classwork individually, either in a classroom or, in some cases, from home \( (n = 14) \) through a video link. All classrooms were limited to 15 students at one time, separated according to COVID-19 guidelines. Consequently, participants did not interact meaningfully with each other or with the instructor during training or testing.

Participants were already familiar with Quizlet. They recorded their completion time and added screenshots of their progress in Google Classroom (see Appendix D). Each student used their personal device, in most cases a laptop, but some used a tablet. This difference affected how students took screenshots, used Google Classroom, and how easily they could type. Nevertheless, since the present study used a within-subjects design and response times were not compared between participants, this would not have affected the results. Students completed posttests individually on their assigned days. Although students’ screens could not be watched simultaneously, it is highly unlikely that participants tried to search for answers given that the students were unaware of which language they had learned, that searching for the target words themselves does not give a Hebrew-English translation due to choices in transliteration, and that there was no incentive for them to score highly.

Analysis

Posttests were scored one point for each correct response, with a possible total of eight points. No ambiguous or partially correct responses were identified. Paired-samples \( t \)-tests showed no statistically significant differences in posttest scores between categories for productive (animals: \( M = 2.43, SD = 2.41 \); food: \( M = 2.16, SD = 2.38 \)), \( t(95) = 1.417, p = .160 \), and receptive measures (animals: \( M = 3.79, SD = 2.37 \); food: \( M = 3.61, SD = 2.47 \)), \( t(95) = 0.775, p = .440 \).

A generalized linear model with a binomial outcome was performed using SPSS 27 (IBM, 2020), which is suitable for data that is not normally distributed. Participant and item variations were included as random intercepts. Initially, individual differences in age and time needed to complete the training were included as covariates, but they had no effect and were removed. The fixed predictors were ISI, RI, and test (productive, receptive), as well as their possible interactions.

The effect size for this model is the odds ratio (\( OR \)), representing the added likelihood of a correct response in one condition over another. For example, \( OR = 2.000 \) indicates that a correct response is twice as likely from the condition with the higher mean. Significance tests were two-tailed, and the alpha was set at \( p = .05 \) with sequential Bonferroni correction.

Results

We first present training data in order to better interpret the results, followed by descriptive and inferential statistics for the posttests. All datasets and syntax can be found via Open Science Framework.

Training

Two measures were used to examine participants’ training performance: (1) time needed to complete the learning (S1 and S2) and relearning (S3) sessions and (2) accuracy at the beginning of S3 (Table 1).

Time to complete training was highly variable between participants, based on the \( SD \). On average, participants required less than one minute for each word. The time to complete S3, which included all words from S1 and S2, was less than the sum of the previous two sessions, indicating that relearning was faster than learning. However, very few words were typed without errors in Round 1 of S3, averaging at just over one out of eight from ISI-1 and less than one from ISI-7. This difference was statistically significant but small (\( t(89) = 3.809, p < .001, d = 0.362 \)).
Posttest Results

Table 2 and Figure 2 display the descriptive statistics for posttest scores. Overall, scores were quite low, which is unsurprising after only two sessions with relatively long RIs. As expected, receptive scores were higher than productive scores, and RI-7 scores were higher than RI-28 scores. Crucially, ISI-7 words were better remembered than ISI-1 words.

Table 1

Minutes on Task and Accuracy in Round 1 of S3 (maximum 8)

<table>
<thead>
<tr>
<th>Minutes</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S3 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ISI-1 Words</td>
</tr>
<tr>
<td>S1</td>
<td>7.86 (4.08)</td>
<td>6.53 (6.17)</td>
<td>9.66 (5.09)</td>
<td>1.25 (1.77)</td>
</tr>
</tbody>
</table>

Table 2

Posttest Results from Productive and Receptive Tests by ISI (maximum score = 8) and together (maximum score = 16) at RI-7, RI-28, and Overall

<table>
<thead>
<tr>
<th>ISI</th>
<th>Productive</th>
<th>Receptive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RI-7</td>
<td>RI-28</td>
</tr>
<tr>
<td>ISI-1</td>
<td>2.42 (2.28)</td>
<td>1.42 (2.13)</td>
</tr>
<tr>
<td>ISI-7</td>
<td>3.17 (2.63)</td>
<td>2.17 (2.24)</td>
</tr>
<tr>
<td>All Words</td>
<td>5.58 (4.60)</td>
<td>3.58 (4.02)</td>
</tr>
</tbody>
</table>

Figure 2

ISI-1 and ISI-7 Items at RI-7 and RI-28 for Productive and Receptive Tests
**Statistical Model**

Full details of the model, including all non-significant means and effect sizes, can be found in the online Appendix S1. The model produced statistically significant but small main effects for all variables. ISI-7 scores were significantly higher than ISI-1 scores \((p < .001, OR = 1.613)\), RI-7 scores were significantly higher than RI-28 scores \((p = .004, OR = 1.972)\), and receptive scores were significantly higher than productive scores \((p < .001, OR = 2.188)\).

The interaction between ISI and RI was significant. While RI-7 scores were always higher than RI-28 scores, the drop was bigger in the ISI-1 condition \((p = .001, OR = 2.425)\) but less pronounced in the ISI-7 condition \((p = .046, OR = 1.603)\). Viewed differently, the difference between the ISI conditions was smaller at RI-7 \((p = .003, OR = 1.310)\) but larger at RI-28 \((p < .001, OR = 1.984)\). Thus, retention between the two RIs was better for words learned at ISI-7.

The other two-way interactions (ISI*test and RI*test) were not statistically significant in this model, but there was a significant three-way interaction between all predictor variables. For productive scores, the drop from RI-7 to RI-28 was consistent for words from ISI-1 \((p = .017, OR = 2.012)\) and ISI-7 \((p = .030, OR = 1.764)\). However, for receptive scores, only the drop for ISI-1 words was statistically significant \((p < .001, OR = 2.924)\). For ISI-7 words, the drop was not significant \((p = .124, OR = 1.457)\). The advantage to ISI-7 words at the longer RI was therefore more pronounced in receptive scores than in productive scores.

**Additional Analysis**

Following a reviewer’s suggestion, scores for individual words were also checked to explore whether some words were more memorable than others. It seems from Figure 3 that neither category was better remembered. Two words differed from the CVCVC structure in the majority of words. One *glida*, was the most memorable word, and the other, *arnav*, was towards the middle. However, the meaning of words seems to be a better predictor of memorability. The higher-scoring words for *food* (ice cream, meat, orange) would be considered more extravagant or flavourful than the blander lower-scoring words (carrot, milk, bread). Similarly, for *animals*, the higher-scoring words (crocodile, tiger, eagle) are more dangerous and exotic than the domesticated lower-scoring words (dog, sheep, cat).

**Figure 3**

*Percentage of Participants That Remembered Each Item*
Discussion

The present study aimed to investigate the optimal scheduling of vocabulary learning with digital flashcards under conditions applicable to a classroom. Using Quizlet, secondary school students aged 11-18 learned 16 novel FL words at either ISI-1 or ISI-7 (within-subjects) and were tested at either RI-7 or RI-28 (between subjects) on both productive and receptive measures.

Results showed a small but statistically significant difference between ISI conditions, according to which ISI-7 led to better retention at both RI-7 and RI-28. This contrasts with previous research in several important ways. Firstly, previous research on FL vocabulary with secondary-school learners did not find a lag effect using other types of tasks (Küpper-Tetzel et al., 2014; Rogers & Cheung, 2020, 2021; Serrano & Huang, 2018, 2021). Secondly, previous lab studies have only found an advantage to a longer ISI at the longer RI (Bahrick, 1979; Bird, 2010; Li & DeKeyser, 2019), whereas our results showed the lag effect to be consistent at the shorter and longer RIs for productive measures. Finally, a grammar-learning experiment using Quizlet with the same intervals and in the same setting (Serfaty & Serrano, 2022) found no global advantage to either condition. Instead, ISI-7 was only better when learners found the task to be less challenging due to their individual characteristics (e.g., proficiency, task-completion time). To our knowledge, this is the first time that differential lag effects have been shown for grammar and vocabulary learning with the same learners, training tasks, and intervals. To interpret this difference, it would be reasonable to assume that single items of vocabulary are simpler to remember than the complex rules of long sentences. This assumption is supported by the much shorter training times in the present study compared with the grammar study. Therefore, the more difficult ISI-7 may have added desirable difficulty to the comparatively simple vocabulary learning task but not to the more complex grammar task.

The present study also compared productive and receptive vocabulary knowledge. Although the training involved productive recall practice, significantly higher scores were obtained on the receptive test, in line with claims that receptive vocabulary knowledge develops earlier and is easier to attain than productive knowledge (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004). The advantage of ISI-7 for productive knowledge is theoretically interesting, but the difference in scores (18% vs. 27% at RI-28) was small. However, in the receptive test, the difference in scores (28% vs. 47% at RI-28) would be quite meaningful in an educational context. Similarly, Chen and Truscott (2010) found bigger effects on receptive tests than on productive tests for different quantities of input.

A speculative interpretation could be that receptive knowledge was better retained between sessions than productive knowledge, especially from the longer ISI. This would be similar to findings from Barclay and Pellicer-Sánchez (2021), in which form-recognition ability was retained while form-recall ability decayed. Training data supports this interpretation: words were generally not retrievable productively at the start of S3. However, since relearning was faster than learning, some partial knowledge (i.e., receptive knowledge) must have been retained. When participants saw a cue, they had the opportunity to retrieve items productively. The subsequent feedback could then serve to remind participants of retained receptive knowledge, because receptive knowledge is the comprehension or recognition of the target form. If both types of knowledge can be retrieved, and receptive knowledge is better retained between sessions than productive knowledge, then receptive knowledge would also be more effectively reinforced through relearning. This reinforcement would be stronger for ISI-7 words than for ISI-1 words due to greater retrieval effort or better consolidation, culminating in a stronger lag effect for receptive knowledge.

It then follows that the lag effect would be stronger for productive knowledge if it was better retained between training sessions. This could be achieved by adding more sessions (Rawson et al., 2018). It is clear that two sessions of Quizlet for FL vocabulary, as an isolated activity, are not enough. Nakata et al. (2021) showed the importance of cumulatively reviewing vocabulary over a long period of time in order to avoid forgetting. Rawson et al. (2018) have advocated for more studies involving several sessions that result in higher scores, pointing out that such low scores would not be useful to real students needing to pass exams.
Another finding from the training data requires explanation. Current accounts of the lag effect (Koval, 2022; Smith & Scarf, 2017) emphasize that successful retrieval is necessary for a longer lag to have a facilitative effect on retention. However, words from ISI-7 were not typed correctly at the first round of S3, and therefore it could not be claimed that successful retrieval followed the longer lag. Despite this, a lag effect was detected.

One explanation could be that successful retrieval is still valid if it comes after a round of feedback. For instance, if an ISI-7 item is retrieved in Round 2, it did not follow the longer lag, but it was still successfully retrieved for the first time in seven days. If items from both ISI-1 and ISI-7 are retrieved during Round 2, the ISI-7 items would still require more effort. Training data showed that ISI-1 words were more accessible in memory than ISI-7 words, even if neither could be retrieved in full during Round 1. Therefore, the lag effect would still apply.

Alternatively, it is probable that ISI-7 words required more retrieval attempts in S3 than ISI-1 words. Some viewpoints hold that unsuccessful retrievals are also beneficial, priming the learner to pay attention to feedback (Kornell & Vaughn, 2016). For Quizlet in particular, if a word is typed correctly, the user does not see it in feedback. Therefore, more incorrect responses elicit more visual feedback and more retrieval attempts, which could reinforce memory (Nakata, 2017; Webb, 2007). Barclay and Pellicer-Sánchez (2021) found that more attempts to reach the criterion in productive flashcard learning resulted in better retention at their RI-28 form-recognition test. From another perspective, Bahrick and Hall (2005) argued that unsuccessful attempts prompt the learner to identify bad mnemonic strategies. If a word is easy to retrieve due to a short lag, a bad strategy might not be detected, but an unsuccessful retrieval attempt may prompt the learner to develop a better strategy.

**Limitations and Future Directions**

The present study targeted FL vocabulary learning through a popular educational app, Quizlet, for secondary-school students in an international school in Cambodia. A replication would be warranted before findings could be generalized to other contexts and populations. It would also be interesting to replicate the experiment using a research-focused tool (e.g., Gorilla) rather than Quizlet in order to track learners' response times on successful retrievals and the number of trials required to reach the criterion per word. These indicators could confirm our speculation that more effort was induced in retrieving words from the longer lag or that more retrieval attempts led to better retention.

A further limitation of the present study is that the receptive test came after the productive test, using the same items, and that participants only engaged in productive practice during training. It would be interesting to compare lag effects for both productive and receptive practice, using enough target words to test both productive and receptive knowledge without repeating items. Additionally, a future study could conduct a productive and receptive test without feedback at the beginning of S3 in order to compare how much of each kind of knowledge was retained after different lags.

Finally, although having only two sessions per structure served the purpose of investigating lag effects, more sessions would be needed for authentic FL learners to achieve a desirable level of knowledge. It is also possible that adding more sessions could moderate the effects found in the present study. Similarly, using an unknown language was necessary to isolate the effects of the treatment from prior knowledge and extra-experimental practice; however, in an authentic FL classroom, the target vocabulary would be meaningful words chosen to match the content of the lessons. This would inevitably reinforce long-term memory and lead to higher overall retention. It was even observed in the present data that more dangerous animals and more enticing food items were better remembered than their everyday counterparts. Future research in this area might include this distinction as an independent variable and check whether memorability interacts with lag or with the dimension of knowledge being tested.
Conclusions and Pedagogical Implications

The present paper has reported an experiment in which secondary school students used Quizlet to study unknown words over two sessions, using productive recall, with either one day or one week between sessions. The longer interval promoted better retention of learned items, especially on receptive measures. To our knowledge, this is the first study that has confirmed the lag effect for FL vocabulary learning in a secondary school context and the first to demonstrate differential lag effects for productive and receptive measures. These findings have direct pedagogical implications. Digital flashcards offer educators a tool to build a baseline of vocabulary knowledge for individual learners. Although scores were quite low in the present experiment’s delayed posttests, it would be expected and recommended that flashcards are reviewed more than twice in order to preserve memory for longer. Moreover, vocabulary sets should be used as a supplemental activity to meaningful language practice, either beforehand to pre-learn key vocabulary or afterwards to prevent forgetting. Most importantly, our results suggest that educators should schedule these sessions at multi-day intervals, rather than repeating the same set on consecutive days. This should help students to remember what they study for longer and reduce the number of sessions required to build reliable and durable FL vocabulary knowledge.

Acknowledgements

We would like to thank our participants for taking part in this study and express our gratitude to the anonymous reviewers and journal editors for their effort in bringing the paper to publication. We would also like to acknowledge the Spanish Ministry of Science and Innovation, Project PID2019-110536GB-I00, for additional support.

Notes

1. During the publication period for the present paper, another study was made available online involving ISIs and Quizlet in a classroom. Muqaibal et al. (2023) used Quizlet as a tool for learning L2 vocabulary in daily versus weekly sessions, finding no differences between ISI conditions. In this study, students read vocabulary at the start of each session and then engaged in different types of activities on the website. This study was therefore similar to the other cited classroom studies in that they controlled the number of class periods dedicated to study and used varied activities. Thus, although Quizlet was the tool, the study did not employ a lab-like methodology with retrieval between sessions.

References


## Appendix A. Target Items

<table>
<thead>
<tr>
<th>Target</th>
<th>English</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kelev</td>
<td>Dog</td>
<td><img src="image" alt="Pug" /></td>
</tr>
<tr>
<td>Lehem</td>
<td>Bread</td>
<td><img src="image" alt="Bread" /></td>
</tr>
<tr>
<td>Hatul</td>
<td>Cat</td>
<td><img src="image" alt="Cat" /></td>
</tr>
<tr>
<td>Halav</td>
<td>Milk</td>
<td><img src="image" alt="Milk" /></td>
</tr>
<tr>
<td>Namer</td>
<td>Tiger</td>
<td><img src="image" alt="Tiger" /></td>
</tr>
<tr>
<td>Mayim</td>
<td>Water</td>
<td><img src="image" alt="Water" /></td>
</tr>
<tr>
<td>Keves</td>
<td>Sheep</td>
<td><img src="image" alt="Sheep" /></td>
</tr>
<tr>
<td>Tapuz</td>
<td>Orange</td>
<td><img src="image" alt="Orange" /></td>
</tr>
<tr>
<td>Arnav</td>
<td>Rabbit</td>
<td><img src="image" alt="Rabbit" /></td>
</tr>
<tr>
<td>Gezer</td>
<td>Carrot</td>
<td><img src="image" alt="Carrot" /></td>
</tr>
<tr>
<td>Tanin</td>
<td>Crocodile</td>
<td><img src="image" alt="Crocodile" /></td>
</tr>
<tr>
<td>Marak</td>
<td>Soup</td>
<td><img src="image" alt="Soup" /></td>
</tr>
<tr>
<td>Nesher</td>
<td>Eagle</td>
<td><img src="image" alt="Eagle" /></td>
</tr>
<tr>
<td>Basar</td>
<td>Meat</td>
<td><img src="image" alt="Meat" /></td>
</tr>
<tr>
<td>Karish</td>
<td>Shark</td>
<td><img src="image" alt="Shark" /></td>
</tr>
<tr>
<td>Glida</td>
<td>Ice Cream</td>
<td><img src="image" alt="Ice Cream" /></td>
</tr>
</tbody>
</table>
Appendix B. Quizlet Screenshots

Cue:

Feedback after an incorrect response:
**Appendix C. Mean Age and Training Time of Experimental Groups**

<table>
<thead>
<tr>
<th>Posttest RI</th>
<th>Animals First</th>
<th>Food First</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>RI-7</td>
<td>n = 24</td>
<td>n = 24</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Age: $M = 13.46$ (2.13)</td>
<td>Age: $M = 13.17$ (1.86)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time: $M = 25.95$ (15.67)</td>
<td>Time: $M = 21.23$ (7.98)</td>
<td></td>
</tr>
<tr>
<td>RI-28</td>
<td>n = 26</td>
<td>n = 22</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Age: $M = 13.23$ (1.58)</td>
<td>Age: $M = 13.95$ (1.96)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time: $M = 22.75$ (11.63)</td>
<td>Time: $M = 25.05$ (10.25)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>46</td>
<td>97</td>
</tr>
</tbody>
</table>

**Comparison between RI-7 & RI-28**

<table>
<thead>
<tr>
<th></th>
<th>$t$</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.650</td>
<td>.518</td>
</tr>
<tr>
<td>Time</td>
<td>0.337</td>
<td>.737</td>
</tr>
</tbody>
</table>
Appendix D. Google Classroom Screenshot

Assignments appeared in the participant’s classroom at the specified time. Each assignment contained a Google Doc with the Quizlet link matching that participant’s experimental condition.

The Google Doc had space to record their times and provided the link to the Quizlet set. They were required to add screenshots of the final page, which showed their progress in each round. Google Classroom also tracks the time that the doc was opened and submitted.

About the Authors

Jonathan Serfaty completed his PhD at the Universitat de Barcelona and is currently a postdoctoral researcher at the Université Côte d’Azur. His research focuses on how the quantity and conditions of learning L2 vocabulary and grammar affect their long-term retention. He has previously published in Language Learning, Applied Psycholinguistics, and System.

E-mail: jonathan.serfaty@univ-cotedazur.fr
ORCiD: https://orcid.org/0000-0003-2861-6790

Raquel Serrano is Associate Professor in the Department of Modern Languages and Literatures and English Studies at the University of Barcelona. Her research focuses on time distribution, study abroad, and L2 vocabulary learning. Some of her recent publications have appeared in Language Teaching, Language Learning, and Language Awareness.

E-mail: raquelserrano@ub.edu
ORCiD: https://orcid.org/0000-0001-9335-4702