Evaluating the impact of spaced practice using computer-assisted language learning (CALL) on vocabulary learning in the classroom

Article

Accepted Version

EVALUATING THE IMPACT OF SPACED PRACTICE USING COMPUTER-ASSISTED LANGUAGE LEARNING (CALL) ON VOCABULARY LEARNING IN THE CLASSROOM

Abstract

This study investigated the influence of practice distribution (i.e. spacing between practice sessions) on successful vocabulary learning by examining two different time distributions, i.e. 1-day spacing (1-DS) versus 7-day spacing (7-DS) using a freely available computer-assisted language learning programme (Quizlet). The study achieved high ecological validity through a classroom-based study with low proficiency L2 English language learners at an Omani college of technology. The sample consisted of 96 participants in Control (*n*=33), 1-DS (*n*=34) and 7-DS (*n*=29) Groups. The Control Group was a test-only group with no explicit practice activities for the target words (34 nouns). Meanwhile, the 1-DS Group (one day spacing between each practice session) and 7-DS Group (seven day spacing between each practice session) received four 20-minute practice sessions using Quizlet to learn the target words. The participants completed baseline tests, including vocabulary level tests (VLTs), and working memory tests (WMTs), alongside performance tests at three time points (pre-, immediate post-, delayed post-test). The results revealed that the two experimental Groups (1-DS, 7-DS) both scored significantly and equally higher than the Control Group at post-test, indicating that the spacing of practice sessions did not mediate learning success through this computer-based vocabulary practice. The technique feature analysis (TFA) model by Nation and Webb (2011) was applied to identify the key features of Quizlet activities, which may contribute to its effectiveness for vocabulary learning. The results revealed that the activities met a high percentage of the TFA criteria, which may account for the significant learning gains achieved by both the 1-DS and 7-DS groups.

Introduction

In recent years, second language (L2) English proficiency has become a de facto requirement of university study in many countries around the world. L2 learners are thought to need knowledge of 10,000 words in order to comprehend university texts (Averianova, 2015). However, in a variety of contexts it has been found that students' L2 English vocabulary breadth is often much smaller than 10,000 words (Alqarni, 2019; Laufer, 2000; Nation, 2006). For example, in Oman, despite students receiving around 1,350 hours of English language instruction at school (from age 6 to 18), the English vocabulary size of a new intake of Omani students at university level is often much lower, commonly no more than 2,000 words (Horst et al., 2000). Similar disparities have also been observed in other contexts. For example, the breadth of vocabulary knowledge for Japanese English as a foreign language (EFL) student at

university is typically between 2,000 and 2,300 words, after being exposed to between 800 and 1,200 hours of English instruction (Barrow et al., 1999; Shillaw, 1995). Meanwhile, Indonesian students at this stage of their education, following 900 hours of English tutoring, can recognise approximately 1,220 words only (Nurweni & Read, 1999). Therefore, identifying effective methods for developing learners' L2 English vocabulary is particularly important to prepare students for studying in an English-medium instruction context.

Several studies have indicated that having sufficient vocabulary knowledge underpins language proficiency and performance in all four language skills, i.e. reading, writing, listening and speaking (for example, Milton, 2013; Nunan, 1999; Smith, 2003). Therefore, an ongoing concern of English language instructors is how to efficiently and effectively improve vocabulary knowledge in a real-life setting (the classroom), especially amongst low-beginner level students who are struggling to progress. Further, language teachers need to maximise the benefit of the limited learning time available in the classroom; therefore, the question of how to structure sessions and distribute practice activities within the time available is of particular importance.

Whilst there is extensive research from cognitive psychology demonstrating the benefits of longer spacing between practice sessions for long-term knowledge retention, the growing body of L2 research on this topic has produced mixed results (e.g., Küpper-Tetzel et al., 2014; Rogers & Cheung, 2018; Serrano & Huang, 2018). Further, methodological shortcomings, such as a lack of control group and variation in timing of delayed post-tests, has limited the comparability and generalisability of findings.

This study, therefore, aimed to contribute to research on this topic, through a novel, ecologically valid investigation of the impact of practice distribution (1-day spacing and 7-day spacing) on intentional vocabulary learning by lower proficiency adult learners of L2 English. The study incorporated a Control group and ensured comparable timing of delayed post-tests for all experimental groups, addressing the methodological shortcomings of existing studies. Further, the study utilised a freely-available computer-assisted language learning (CALL) application (Quizlet, 2020) to facilitate explicit, intentional vocabulary practice during the study intervention. CALL applications provide the opportunity to optimally space practice in order to facilitate learning (Nation, 2001) and offer "a currently unexploited opportunity to schedule study sessions to optimize long-term retention" (Rohrer & Pashler, 2007, p.186). The present study, therefore, sought to investigate, in a classroom setting, the effectiveness of systematic, repeated, intentional vocabulary practice delivered via an online learning tool, for

low proficiency L2 English learners, and whether the spacing of practice sessions impacted the learning benefits observed.

Literature Review

Intentional Vocabulary Learning

Intentional vocabulary learning is defined as "any activity geared at committing lexical information to memory" (Hulstijn, 2001, p. 267), that is the intention to learn new words, via activities explicitly designed to introduce and practice unknown lexical items. In contrast, incidental vocabulary learning requires extensive exposure over a long period of time, whereas direct intentional learning can facilitate the rapid development of lexical knowledge (Elgort, 2011); an advantage for the classroom context where the time available and amount of exposure to the target language is often limited. Intentional vocabulary learning is thought to be particularly useful in the early stages of vocabulary learning and for lower proficiency learners (Ma & Kelly, 2006; Nation & Meara, 2010), who, due to their limited existing vocabulary knowledge, are less likely to be able to infer the meaning of unknown words when they are encountered incidentally. Nation (2001), therefore, emphasises the usefulness of introducing and practising high-frequency words via direct, explicit instruction in the early stages of language learning. Explicit practice techniques, such as flashcards, can be particularly useful for efficiently and effectively learning new vocabulary (Tung, 2015) and rapidly expanding learners' vocabulary size, particularly when learners are guided in how to use them efficiently (Nation, 2011). For example, using the first language (L1) or pictures to clearly convey the meaning of target vocabulary, incorporating spaced repetition, and flexibility in which and how many words are practised at any one time, can promote memorization and retention of target vocabulary (Nation, 2011, 2013).

In recent years, attention has turned to the benefits of CALL applications to support direct vocabulary practice (Tung, 2015) and provide the opportunity to optimally space practice in order to facilitate learning and retention (Nation, 2001).

Intentional Vocabulary Learning via a CALL programme: Quizlet

CALL programmes have been recommended for creating an ideal environment for learners at an early stage of their acquisition of EFL (Tam et al., 2010), because they can enhance confidence and motivation (Krish et al., 2011). Dizon (2016) recommends Quizlet as an interactive programme that facilitates intentional vocabulary learning. The Quizlet website provides several vocabulary activities in the form of word lists and flashcards. It is widely used,

with over 350 million study sets, more than 3 billion study sessions and 50 million active users every month across 130 countries (Quizlet, 2020). Broadly speaking, CALL programmes nowadays go far beyond traditional paper and pencil methods, offering numerous types of activities for practising various aspects of the language (Özer & Koçoğlu, 2017), and supporting a learner-centred experience (Al-Khatib, 2011).

Whilst Quizlet is widely used in teaching and learning contexts, few research studies have investigated the efficacy of Quizlet for vocabulary learning (Crandell, 2017). Previous studies (e.g., Anjaniputra & Salsabila, 2018; Chien, 2015; Jackson III, 2015) have primarily examined learners' perceptions of the tool, with few conducting experimental studies in real-life settings to identify how it improves productive (i.e. use in writing/speaking) and receptive (i.e. understanding when listening/reading) vocabulary knowledge. Only a handful of recent studies (e.g., Korlu & Mede, 2018; Özer & Koçoğlu, 2017; Zambrano Acosta, 2018) have implemented a quasi-experimental design to investigate Quizlet's effectiveness for vocabulary learning. The findings reveal greater learning gains at post-test for Quizlet than control groups, as well as positive views expressed by participants. However, to the best of the present authors' knowledge, Özer and Koçoğlu (2017) represents the only quasi-experimental study attempting to measure Quizlet's effectiveness for long-term retention via a delayed post-test in comparison to a paper-based vocabulary Notebook. Eighty-nine randomly selected participants were divided into four classes: two experimental groups utilising either Quizlet or Notebook and two control groups (without the use of Quizlet or a vocabulary notebook). All classes followed the same curriculum; however, the Quizlet group were provided with a flashcard software programme (Quizlet) for learning and recalling new words from three consecutive units of their textbook, whereas the Notebook group maintained paper-based vocabulary notebooks for the same purpose. Özer and Koçoğlu (2017) found that the experimental groups (Quizlet and Notebook), significantly outperformed the control groups in the post- and delayed post-tests, with the Quizlet group making slightly more improvement than the Notebook group. Nevertheless, the delayed post-test took place just two weeks later, therefore the impact of Quizlet on longer-term vocabulary retention is not known. Further research is needed which measures retention at intervals greater than two weeks (Ahmadi, 2014).

Although there is a great deal of research that looks at progress in vocabulary learning, the existing research tends to focus on fairly homogenous populations (i.e. intermediate-high ability university students) (for example, Anjaniputra & Salsabila, 2018; Dizon, 2016; Korlu & Mede, 2018). However, Barr (2016) and Sanosi (2018) are amongst the very few researchers who have investigated low-proficiency English learners at university level, with a view to

enhancing their vocabulary knowledge through Quizlet. Barr (2016) used a sample of 32 first-year Japanese students, who were instructed to use Quizlet to prepare for vocabulary tests by viewing and completing the flashcard and gap-filling tasks for the Quizlet vocabulary sets provided in class. The Quizlet users outperformed non-Quizlet users, and also recorded moderately higher scores on test items which required use of the target vocabulary in unfamiliar sentence contexts.

In a similar vein, Sanosi (2018) investigated Quizlet's effect on vocabulary acquisition of 42 EFL beginners in their first year at a Saudi university. Using a matched experimental (who attended regular classes plus used Quizlet) and control group (who only attended regular classes), the authors found that the experimental group outperformed the control group significantly at post-test. Sanosi concluded that Quizlet seems to be an ideal vocabulary-learning tool, for use both within and outside the classroom. However, Sanosi's study did not include a delayed post-test to assess long-term recall. Further, it is not possible to draw conclusions on the optimal amount and frequency of practice, as level of exposure varied between participants.

Based on the above discussion, more experimental research is needed to establish the learning effectiveness of Quizlet in terms of both immediate and long-term learning gains. Additionally, research is needed with non-typical populations, such as lower proficiency learners, e.g. those who remain at a level equivalent to Common European Framework of Reference for Languages (CEFR) level A1 despite 10+ years of instruction in the target language. Further, technology offers versatility and flexibility, in terms of making learning materials accessible anytime and anywhere. However, as yet, little is known about whether practice distribution (i.e. the frequency and spacing of practice sessions) influences learning outcomes using CALL applications, such as Quizlet in classroom settings.

Technique Feature Analysis (TFA)

A wide range of recent studies have emphasised the advantages of using software to help students learn vocabulary (e.g., Korlu & Mede, 2018; Özer & Koçoğlu, 2017; Zambrano Acosta, 2018). These studies have generally favoured Quizlet, based on students' results and perceptions, as noted above. To examine the potential effectiveness of a given task or learning software, Hu and Nassaji (2016) recommend using the Technique Feature Analysis (TFA) framework as a means of assessing the depth of processing facilitated by a task, and as an indication of its effectiveness for facilitating learning.

The TFA framework proposed by Nation and Webb (2011) aims to describe the key cognitive processes involved in vocabulary acquisition along five core dimensions: motivation, noticing, retrieval, generation, retention. It offers a rationale for adopting intentional vocabulary teaching, allowing researchers to assess whether the chosen activities offer opportunities for vocabulary practice, rehearsal and retrieval to deepen short- and long-term memory processing. For instance, having a 'generation' component; either receptive (listening or reading) or productive (using the word in a new context), is important for enhancing 'noticing' (drawing attention to unknown words) in the direction of the learner's knowledge gap (Swain, 2005). Moreover, generation provides learners with opportunities for retrieval and rehearsal, thereby developing their vocabulary knowledge (Keating, 2008; Laufer, 2006). Nation and Webb (2011) argue that productive retrieval, illustrated in an audio-visual presentation (such as software flashcards), is recommended to reinforce the mental links between form and meaning. Meanwhile, Nakata (2008, p. 5) describes rehearsal "as an activity to encode new information into our long-term memory through overt or silent articulation". The framework emphasises the importance of providing learners with repeated retrieval opportunities; however, it remains unclear how often and how frequent that practice should be.

Spaced Practice

The provision of multiple, spaced opportunities for information retrieval is a key component of the TFA Framework. However, an important question, particularly relevant to the language classroom where time available for practice is often limited, is how much time is needed between practice sessions. The study of spacing schedules concerns what is known as the 'lag effect' (Rogers, 2017), i.e. whether there is better knowledge retention and recall when there is a longer interval ('spacing') between practice sessions (Nakata & Suzuki, 2019; Serrano & Huang, 2018). A small number of studies have looked at longer and shorter spacing for L2 grammar learning (e.g. Bird, 2010; Kasprowicz et al., 2019; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2015) and for L2 vocabulary acquisition (e.g., Küpper-Tetzel et al., 2014; Rogers & Cheung, 2018; Serrano & Huang, 2018); however, research thus far has produced mixed results. Some have identified that longer spacing (e.g. 7 days) leads to better retention (Bird, 2010; Rogers, 2015; Serrano & Huang, 2018), whereas others have found that shorter spacing (e.g. 1 day or 3.3 days) has a significant advantage (Rogers & Cheung, 2018; Suzuki, 2017), and still others have revealed no statistically significant differences between shorter and longer spacing schedules (Kasprowicz et al, 2019; Küpper-Tetzel et al., 2014). It should be

mentioned here that to the current authors' knowledge, Kasprowicz et al. (2019) is the only lag effect study to include a control group to account for potential test-retest effects.

A number of lag effect studies (e.g. Rogers, 2015; Suzuki, 2017) have investigated the relationship between the intersession interval (ISI) (i.e. the time period between study sessions), and the retention interval (RI) (i.e. the gap between the last study session and the test), in order to establish the extent to which increasing the time between practice sessions improves long-term knowledge retention. Rohrer and Pashler (2007) concluded that the optimal length of the ISI can be determined by the RI. They suggest that the optimal ISI should be between 10% and 30% of the RI, for example if the RI is 30 days, the optimal ISI would be between 3 and 9 days. However, whether this ratio applies to learning in a classroom context, and for vocabulary learning and retention, is still unknown.

Serrano and Huang (2018) investigated the effects of time distribution on five repetitions of reading a short passage for L2 fluency and incidental vocabulary learning. The study focused on the optimal lag effect, i.e. the ideal interval between reading sessions of the same text, to promote the maximum impact on L2 vocabulary acquisition. 71 Taiwanese EFL students (aged 16 years) were grouped into intensive (1-day ISI) and spaced (7-day ISI) groups. Both groups completed a post-test immediately after the final practice session followed by a delayed post-test, scheduled at an optimal (according to Rohrer & Pashler, 2007) 25% RI for each group. For the Intensive (1-day ISI) group the delayed post-test took place four days after the final practice session, and for the Spaced (7-day ISI) group, the delayed post-test took place 28 days after the final practice session. Although both groups showed gains at post-test, the results revealed that shorter spacing (Intensive, 1–day ISI) led to greater vocabulary learning at immediate post-test. However, longer spacing (Spaced, 7-day ISI) supported greater long-term retention, with the Spaced group performing significantly better at their 28-day RI delayed post-test, than the Intensive group did at their 4-day RI delayed post-test. Whilst these findings suggest that more distributed practice may lead to better long-term retention, it is important to note that the study design did not allow for comparison of the two experimental groups at equivalent delayed post-tests. Therefore, we do not know how the Intensive group might have performed at the 28-day RI delayed post-test, nor the Spaced group at the 4-day RI delayed post-test.

This current study seeks to extend the design of Serrano and Huang's study and address this methodological issue by also examining the performance of each group at each delayed post-test (see Methods section). Further, the present study builds on existing lag effect research to compare the impact of longer and shorter spacing distributions, but for low proficiency L2

English university learners within an instructed context; an under-researched population within existing lag effect studies.

Research Questions

Two research questions (RQs) are addressed in this study:

RQ1: To what extent does intentional, direct vocabulary practice via a CALL programme (Quizlet) promote vocabulary learning and retention amongst low proficiency learners in the classroom, in terms of a) recognition and b) recall of target vocabulary?
RQ2: To what extent does the distribution of practice sessions (1-day spacing versus 7-day spacing) moderate the benefits of using Quizlet to promote vocabulary learning and retention amongst low proficiency learners, at immediate post-test and delayed post-test?

Method

Participants

A quasi-experimental design was adopted, with six intact classes of L2 learners (N = 96) drawn from an Omani college Foundation Program. The learners had received 12 years of English instruction prior to joining the Foundation Program but remained at Level 1 (as measured by the placement test[1] completed when joining the Program), equivalent to Level A1 on the CEFR and therefore can be considered low proficiency L2 English learners.

The participants included 22 females and 74 males with an average age of 18 years. These six classes were divided into three groups (two classes per group): a Control (n=33), a 1-day spacing (1-DS, n=34) and a 7-day spacing (7-DS, n=29) group. The 1-DS and 7-DS groups received four practice sessions (of 20 minutes each) to rehearse the target words (34 nouns) using Quizlet (see activity descriptions below). The Control Group was a test-only group; they continued with their usual classroom teaching during the intervention but did not engage in any

[1] The placement test (lasting 90 minutes) measures the English language level of each new student intake. It tests English grammar and vocabulary knowledge and includes 100 questions (multiple-choice and gap-filling exercises), with a point awarded for each correct answer. Students are assigned to a level based on these test results (0-20=Level 1; 21-50=Level 2; 51-70=Level 3; 70-90=Level 4, and over 90=post-foundation). New students joining the universities and higher education institutions (HEIs) in Oman are required to take an English placement test, so that they can be allocated to the appropriate level (Level 1 to Level 4) on the Foundation Programme. For this purpose, HEIs administer an in-house placement test, designed according to their learning outcomes as HEIs (Al-Mamari, 2012). Very few students tend to pass these placement tests and enrol directly on a college or university programme. Al Mahrooqi (2012, as cited in Kamanpoori, 2014) found that 80% of students (from a sample of 8,000 students) were required to undertake English courses in the Foundation Programme, prior to accessing higher education.

explicit/intentional instruction relating to the target vocabulary, although may have had incidental exposure through their regular English teaching.

Target Words

The 34 target words (Appendix 1) were nouns in the 2,000-word and academic word lists, selected from the Level 2 Vocabulary Log[2], to ensure that they went beyond the participants' current vocabulary knowledge (i.e. Level 1). The pre-test, as well as the baseline vocabulary levels tests (see next section), confirmed students were unfamiliar with the target vocabulary prior to the study commencing. The Level 2 Vocabulary Log consists of 100 words, representing different parts of speech, including 34 nouns, which were chosen for the present study. The researchers used the *Compleat Lexical Tutor* website (https://www.lextutor.ca/vp/comp/) to confirm the frequency of vocabulary (Cobb, n.d.; Heatley et al., 2002).

The equivalent meanings in the students' L1 (i.e. Arabic) were provided within the Quizlet activities, to ensure that the meaning of the words was clear for the participants. In addition, engagement with L1 translation can help to transfer target L2 vocabulary into long-term memory, particularly for lower proficiency learners, for whom processing of the meaning of target L2 vocabulary is thought to be mediated by L1 translation (Jiang, 2000, 2002). Only one part of speech (nouns) was selected for this study, because nouns are among the main components of sentences (Webb, 2005). In addition, some scholars argue that certain parts of speech are harder than others to learn (e.g., Childers & Tomasello, 2006). Therefore, the authors chose the noun, representing a large proportion of the items on the Level 2 list. This ensured that the target words were as similar as possible in their frequency and range, while also presenting a degree of challenge to the learners; thereby avoiding the effect of differing levels of difficulty between the target words.

Study Instruments

The study instruments[3] consisted of two baseline tests, including a 2000-5000 word vocabulary level test (VLT) and working memory tests (WMTs) as well as a vocabulary test measuring learners' active (L1 to L2) and passive (L2 to L1) recognition and recall of the target

[2] A Vocabulary Log is a vocabulary list provided to students of each level. Words in the Vocabulary Log are extracted from the corresponding course books to facilitate students' understanding of the subject matter, when they encounter these vocabulary items during their lessons. Students are asked to find the L1 equivalents of all words provided in the Vocabulary Log, identify the part of speech, and use the word in sentences. The teacher's responsibility is to check that students complete this task and later conduct spelling tests.

3 All materials will be available via the IRIS Database.

vocabulary, utilised to measure performance at pre-, post- and delayed post-test. The VLT and vocabulary tests were pen-and-paper tests administered to the whole class together. The WMT tests were administered one-to-one with each participant.

In terms of the baseline tests, the VLT (Schmitt et al., 2001) is a vocabulary test assessing learners' vocabulary size and knowledge and was administered to check that the participants' vocabulary knowledge fell below the identified vocabulary range (2000-word) for the intervention. The participants were shown 60 words from each frequency band (2000, 3000, 5000) presented in groups of six words alongside three definitions and were asked to identify the word that matched each definition. In addition, WMTs (forward / backward digit span, Appendix 2) were employed. A central cognitive process involved in vocabulary learning under a spaced practice schedule is 'retrieval', i.e. the process of retrieving information (phonological form, orthographic form, meaning) about L2 words from storage in the memory (Nakata, 2015). On each encounter with a word, the learner retrieves their previous knowledge of its form and meaning and assimilates this with the new input, thereby supporting further retrieval and eventual transfer into long-term memory (Baddeley, 1990). Working memory, as "a cognitive device for online information retrieval and processing" (Teng & Zhang, 2021, p. 3) is likely to play a key role in this process. WM has been identified as key predictor of L2 learning (Al-Hammadi, 2012; Larsen-Freeman, 2014; Linck et al., 2014) and has been found to impact lexical learning, particularly under instructional conditions (e.g. intentional vocabulary learning) involving simultaneous attention to form and meaning (Ruiz et al., 2021). Therefore, the present study utilised WMTs to assess the learners' ability to retrieve and manipulate auditory information and to detect any variation in the students' recall ability (Climie & Rostad, 2011; Elsayyad, 2014). In the forward digit span test (designed to test ability to retain simple sequences of auditory information, Climie & Rostad, 2011), participants heard sequences of numbers which increased in length from two to nine digits and had to repeat each sequence verbatim. For the backward digit span test (testing ability to manipulate simple sequences of auditory information, Climie & Rostad, 2011), participants again heard sequences of numbers increasing from two to nine digits but had to repeat the numbers in reverse order. Both tests included a total of 16 items, but ended if the participant repeated two sequences incorrectly. Both the VLTs and WMTs ensured the study sample equivalence and scores from both sets of baseline tests were included as covariates in the analysis to control for any potential confounding effects.

The 34 target words (Appendix 1) were divided into two sets for the vocabulary test, with one set of 17 items in the Recall test and the remaining in the Recognition test, to test both

productive and receptive knowledge (Laufer & Goldstein, 2004). The Recall items included nine items requiring translation of L1 Arabic words into L2 English ('active' recall) and eight items requiring translation of L2 English words into L1 Arabic ('passive' recall). The Recognition items included nine multiple choice questions comprising nine L1 Arabic words each with four possible L2 English equivalents ('active' Recognition) and eight multiple-choice questions, comprising eight L2 English words each with four possible L1 Arabic equivalents ('passive' Recognition) (Appendix 3). The distractors in the vocabulary recognition test were nouns chosen from the Cambridge English vocabulary list (Key English Test - KET), appropriate for the A2 level on the CEFR, and the General Service List: 2000 Most useful words list (Bauman, 1995). The reliability of the vocabulary tests was checked using Cronbach's alpha (Cronbach, 1951). The pre-test demonstrated a reliable internal consistency of .84, and both the immediate and delayed post-tests indicated very high reliability of .96. The reliability of the Recall and Recognition tasks at pre-, immediate post-, and delayed post-test are presented separately (Table 1), in line with González-Fernández and Schmitt's (2020, p.481) suggestion that "the recognition and recall masteries of any particular word knowledge component must be seen as separate constructs". The Recall reliability statistic was a little low at pre-test (Table 1), which was likely due to the participants relying on guesswork, as reflected in the low vocabulary scores at pre-test.

Table1

Reliability statistics Derived from The three Vocabulary Tests

Test	Section	Cronbach's alpha	No. of Items
Pre-test	Recall	.67	17
	Recognition	.80	17
Immediate post-test	Recall	.92	17
	Recognition	.95	17
Delayed post-test	Recall	.94	17
	Recognition	.93	17

The intervention

The intervention consisted of four 20-minute sessions during which participants engaged in a variety of English vocabulary-learning activities, using the Quizlet flashcard programme. It offers a range of vocabulary tasks, such as Flashcards, Write, Speller, Gravity, Match and Test, and is also compatible with both the Apple iOS and Google Android mobile platforms. Quizlet is a freely accessible website, mainly focused on vocabulary, where teachers can create their own study sets. The audio-visual function on Quizlet facilitates learning of vocabulary form and meaning.

The learners used a Quizlet study set, created by the first author, to learn 34 new nouns (Appendix 1). These target words were presented to the students through several selected activity types provided by Quizlet, which require recall and recognition of the target vocabulary, including: Flash Cards (i.e., an audio-visual presentation of an individual word in English with its equivalent in Arabic), Match (i.e., matching the English word to its picture), Spell (i.e., listening to the word to write the correct English spelling), Write (i.e., presenting the picture to write the correct word in English) and Test (i.e., match, write and multiple choice tasks). Table 2 details the activities completed in each intervention session (Appendix 4 for further details).

Table 2

Design of Sessions (Time & Activity)

	Task/Mode	Learning Method	Type	Time Spent	No. of Repetitions	
Session 1						
1.	Flashcards	Digital flashcards presenting target L2 word on one side, and L1 meaning on reverse side. Quizlet cycles through all target words one-by-one, showing the L2 word before automatically flipping to reveal L1 meaning.	Recognition	3 min	1	Total= 4
2.	Flashcards	Player 1 displays flashcard showing L2 word and asks player 2 to guess the meaning before clicking to view L1 meaning on reverse side. Players swap roles working through the set of target words.	Recall/ Recognition	10 min	2	
3.	Match	Each player has to match up each L2 word with the correct L1 meaning.	Recognition	7 min	1	
				Total= 20 min		
Session 2						
1.	Flashcards	Digital flashcards presenting target L2 word on one side, and L1 meaning on reverse side. Quizlet cycles through all target words one-by-one, showing the L2 word before automatically flipping to reveal L1 meaning.	Recognition	3 min	1	Total= 3
2.	Spell	Player hears L2 word and has to type correct L2 word. If incorrect, Quizlet reveals missing letters and player tries again until they type correct word.	Recall/ Recognition	17 min	2	
				Total= 20 min		
Session 3						
1.	Flashcards	Digital flashcards presenting target L2 word on one side, and L1 meaning on reverse side. Quizlet cycles through all target words one-by-one, showing the L2 word before automatically flipping to reveal L1 meaning.	Recognition	3 min	1	Total= 3
2.	Write	L1 meaning is displayed, player has to write the correct L2 word.	Recall	17 min	2	
				Total= 20 min		
Session 4						

1.	Flashcards	Digital flashcards presenting target L2 word on one side, and L1 meaning on reverse side. Quizlet cycles through all target words one-by-one, showing the L2 word before automatically flipping to reveal L1 meaning.	Recognition	3 min	Total= 20 min	1	Total= 3
2.	Test	Mixture of multiple choice, Write and Match, and True/False questions used to test recognition and recall of target L2 words.	Recall/ Recognition	17 min		2	

As the Quizlet programme provides several different vocabulary tasks, it was essential to ensure that both experimental groups had the same amount and type of exposure to the target vocabulary items in each training session and across all the sessions. Therefore, the activities were carefully balanced within and across each session to control the total number of repetitions and amount of time spent within the four training sessions. A specific set of vocabulary tasks (i.e., Flashcards, Match, Spell, Write & Test) were chosen to control the type of practice (i.e. recognition and recall) delivered during the training sessions (Table 2). These vocabulary tasks were chosen based on the five main components of the TFA framework (Table 3) and to ensure that the learners had opportunities for both recognition and recall practice of the target vocabulary across the sessions. Two tasks (Gravity and Learn) were excluded as they do not allow the teacher/researcher to control the amount and nature of input for all learners: Gravity is a game-based activity whereby students type the correct word against the clock, while Learn includes multiple randomly selected tasks such as spelling, writing and multiple-choice activities.

All sessions started with Flashcards for three minutes to allow the learners to review the target vocabulary before the learners moved on to the next activity for no more than 17 minutes to practice the target words (see Appendix 4 for examples of each task). Note, in session one, after the initial three-minute Flashcard practice, the learners used the Flashcard activity again in pairs to provide recall practice, before completing the Match task (recognition practice).

The intervention sessions were run by the first author and observed by the class teachers. The intervention took place over a period of four days for the 1-DS Group (1-day between practice sessions) and four weeks for the 7-DS Group (7-days between practice sessions). The vocabulary learning activities were undertaken during regular lessons in the English Language Center labs at the college.

In order to evaluate the vocabulary-learning activities utilised in this study, the researchers cross-checked the features of the TFA Framework, as formulated by Nation and Webb (2011), against the key features of the intervention delivered through Quizlet (Table 3).

Hu and Nassaji (2016) consider this Framework to be a successful predictor of the depth of processing facilitated by a vocabulary activity. The teaching method adopted here, involving Quizlet and spacing between sessions, met 14 out of the 18 criteria within the five TFA components (around 82% of the Framework's features). Therefore, TFA provides a rationale for adopting intentional vocabulary learning through Quizlet.

Table 3
Five-component Framework of Assessment Criteria (Adopted from Nation & Webb, 2011, p.7)

Criteria	Quizlet Activities and Scores				
	Flashcards	**Match**	**Spell**	**Write**	**Test**
Motivation					
Is there a clear vocabulary learning goal?	1	1	1	1	1
Does the activity motivate learning?	1	1	1	1	1
Do the learners select the words?	0	0	0	0	0
Noticing					
Does the activity focus attention on the target words?	1	1	1	1	1
Does the activity raise awareness of new vocabulary learning?	1	1	1	1	1
Does the activity involve retrieval negotiation?	1	1	1	1	1
Retrieval					
Does the activity involve retrieval of the word?	1	1	1	1	1
Is it productive retrieval?	1	0	1	1	1
Is it recall?	1	1	1	1	1
Are there multiple retrievals of each word?	1	1	1	1	1
Is there spacing between retrievals?	1	1	1	1	1
Generation					
Does the activity involve generative use?	1	1	1	1	1
Is it productive?	1	1	1	1	1
Is there marked change that involves the use of other words?	0	0	0	0	0
Retention					
Does the activity ensure successful linking of form and meaning?	1	1	1	1	1
Does the activity involve instantiation?	0	0	0	0	0
Does the activity involve imaging?	1	1	1	1	1
Does the activity avoid interference?	1	1	1	1	1
Average Total score /18 – 14.8	**15**	**14**	**15**	**15**	**15**
Average /100% – 82.2%	**83.3%**	**77.8%**	**83.3%**	**83.3%**	**83.3%**

Procedure

For each session, the students reviewed the study set in Quizlet via the Flashcards activity and then completed another activity; i.e. Match, Spell, Write and then Test respectively. Both the 1-DS and the 7-DS Group performed exactly the same activities for each session (see Appendix 4). Only the timing of each practice session differed between the two experimental groups, as explained above and depicted in Figure 1. Therefore, the overall amount, nature, and

duration of exposure for each individual and within both experimental groups was controlled, to ensure that all participants had a similar experience with Quizlet and equivalent level of exposure to the target vocabulary.

Prior to the intervention, all groups completed the baseline (VLT and WMT) tasks and the pre-test. Following the intervention, the experimental (1-DS and 7-DS) and Control Groups completed immediate and delayed post-tests. The immediate post-test took place directly after the last intervention session. Two delayed post-tests were used. The first delayed post-test was administered four days after the final practice session to half of the participants in each Group (1-DS, 7-DS, Control), while the remaining participants in each group completed the second delayed post-test four weeks (28-days) after the final practice session. Thus, each participant only completed one delayed post-test, with half of the participants in each of the experimental Groups completing the delayed post-test in the optimal ISI:RI ratio (25%) for that group, i.e. for the 1-DS Group, 4-day RI=25% and 28-day RI=3.6%, and for the 7-DS Group, 4-day RI=175% and 28-day RI=25%. This unique design made it possible to address the methodological issue identified in Serrano and Huang (2018) and measure the impact of all three conditions (1-DS, 7-DS and Control) on longer-term retention of the target vocabulary items at both 4-day and 28-day RI, whilst avoiding over-testing the individual participants. Figure 1 illustrates the study design and procedure.

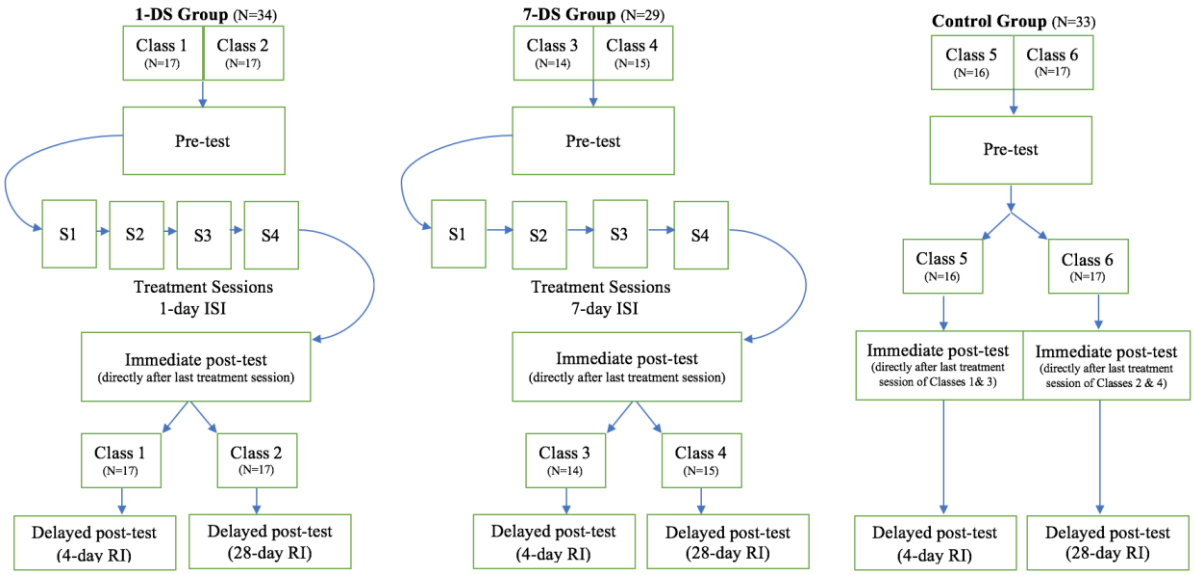


Figure 1: Study Design

Analysis

The data analysis involved two phases. Before the intervention, a Kruskal-Wallis test and a one-way ANOVA were implemented for the baseline tests (VLT and WMT) to compare the 1-DS, 7-DS, and Control Groups. Prior to running the analysis, assumptions of normality and

homogeneity of variance were checked. The data for the WMT met the assumptions and therefore were analysed using one-way ANOVA. The Kruskal-Wallis test was reported for the VLTs, which did not meet the assumptions.

In the second phase, multilevel linear mixed effects modelling, with maximum-likelihood estimation, was employed to examine learners' performance on the pre-, post-, and delayed post-tests for the vocabulary Recognition and Recall tests respectively, taking account of fixed effects for the independent variables (Time, Group, Retention Interval) and covariates (2,000 VLT, 3,000 VLT, 5,000 VLT, forward WMT and backward WMT) as well as random effects including by-participant random intercepts and random slopes for time. Pairwise comparisons with Bonferroni correction were run to identify the cause of any significant interactions between the independent variables (Time*Group, RI*Time, RI*Group, RI*Time*Group).

A model selection approach was used whereby a model including fixed effects for Time, Group and RI and associated interactions (Time*Group, RI*Time, RI*Group, RI*Time*Group was conducted initially (Model 1). The model was then built-up step by step, with the additions of each covariate (Models 2 to 6), before adding by-participant random intercepts (Model 7) and finally random slopes for time (Model 8). For model 7, a variance components covariance structure was used; whereas for model 8, an autoregressive covariance structure was chosen (Field, 2021). At each stage, the model fit was assessed using a chi-square likelihood ratio test, comparing the goodness of fit (*-2LL*) for the current model, against the previous model. Non-significant fixed factors and covariates were retained in subsequent models to enable examination of all independent variables and covariates in the final model. For the vocabulary Recognition task data, Model 8 converged successfully; however, it did not improve the model fit over and above Model 7 ($\chi^2(4) = 7.157$, *ns*). However, Model 7, which included by-participant random intercepts (Var(u_{oj}) = 3.866, *SE* = .767, 95% *CI* = 2.640, 5.721), significantly improved model fit ($\chi^2(1) = 64.794$, $p = .001$) over and above Model 6, which included fixed effects for independent variables and covariates only. For the vocabulary Recall task data, Model 8 failed to converge. However, Model 7, which included by-participant random intercepts (Var(u_{oj}) = 4.377, *SE* = .882, 95% *CI* = 2.949, 6.497), significantly improved model fit ($\chi^2(1) = 59.683$, $p = .001$) over and above Model 6. Therefore, for both tests the results based on Model 7 are reported, which included fixed factors (Time, Group, RI) and interactions (Time*Group, RI*Time, RI*Group, RI*Time*Group), covariates (2,000 VLT, 3,000 VLT, 5,000 VLT, forward WMT, backward WMT) and by-participant random intercepts.

Effect sizes (Cohen's *d*) were calculated to identify the magnitude of change / difference observed between- and within-subjects and interpreted using Plonsky and Oswald's (2014) field-specific benchmarks. To assess the reliability of effect sizes, 95% confidence intervals were calculated, whereupon any confidence intervals that did not cross zero were judged to be reliable indicators of an effect (Plonsky & Oswald, 2014).

Findings

Descriptive statistics for the baseline tests (WMTs and VLTs) are presented in Table 4. There were no significant differences between the WMT scores (forward or backward) of the three groups (Table 4) in either the 4-day or 28-day RI sub-groups. In terms of participants' baseline vocabulary knowledge, analysis via Kruskal-Wallis test demonstrated significant differences between the participants' results for both the 2,000 and 3,000 VLTs within the 4-day RI sub-group. Pairwise comparisons with adjusted *p*-values checked for any differences between the Control vs. 1-DS (p= .150, r= -.34, 95% *CI*= -0.86, 0.51), Control vs. 7-DS (p= .060, r= -.42, 95% *CI*= -0.97, 0.47), and 1-DS vs. 7-DS (p=1.000, r= -.08, 95% *CI*= -0.83, 0.58) scores for the 2,000 VLT and between the Control vs. 1-DS (p= .173, r=- .33, 95% *CI*= -0.78, 0.59),), Control vs. 7-DS (p=.055, r=-.43, 95% *CI*= -0.93, 0.51), and 1-DS vs. 7-DS (p=1.000, r=-.10, 95% *CI*= -0.92, 0.50) for the 3,000 VLT. The results revealed no statistical differences between the groups for the 2,000 and 3,000 VLTs, with small effect sizes. Although small effects (> .40) were observed for the Control vs. 7-DS groups at the 2,000 and 3,000 levels, the confidence intervals crossed zero, suggesting that these effects were unreliable. Further, there were no significant differences in the 5,000 VLT scores for 4-day RI sub-group and no significant differences at any level for the 28-day RI sub-group (Table 4).

Therefore, the three main Groups (1-DS, 7-DS, Control) within each sub-group did not exhibit differences in their existing vocabulary knowledge prior to the study; all participants scored below the 2,000-word frequency level, indicating a low level of existing vocabulary knowledge. In addition, there was no significant difference in the participants' capacity for vocabulary retention or recall (as evidenced in the WMTs). To further account for any potential variation in performance on the Recognition and Recall tasks, which may have been due to these factors, the participants' scores on the VLTs and WMTs were included in the mixed effects models as covariates.

Table 4

Descriptive statistics and between-group comparisons for the VLTs and WMTs (4-day RI and 28-day RI sub-groups)

RI	Group	*n*	VLTs *M*(*SD*)			WM Tests *M*(*SD*)	
			2,000 VLT	**3,000 VLT**	**5,000 VLT**	**Forward**	**Backward**
4-day	**Control**	16	2.00 (*4.21*)	0.88 (*3.24*)	0.19 (*.75*)	7.94 (*1.65*)	5.63 (*1.15*)
	1-DS	17	2.59 (*2.12*)	1.12 (*1.41*)	0.53 (*1.38*)	8.18 (*1.55*)	5.94 (*1.98*)
	7-DS	14	2.86 (*2.18*)	1.43 (*1.60*)	0.64 (*1.50*)	8.57 (*1.22*)	5.86 (*1.79*)
Between-group comparisons ***p* values**			*p*= .043	*p*= .043	*p*= .479	*p*=.514	*p*= .891
28-day	**Control**	17	3.00 (*3.02*)	1.12 (*1.65*)	0.88 (*2.15*)	7.88 (*1.65*)	5.35 (*1.77*)
	1-DS	17	2.71 (*2.87*)	1.00 (*1.77*)	0.24 (*.56*)	7.71 (*1.61*)	5.82 (*1.70*)
	7-DS	15	1.93 (*2.40*)	1.33 (*1.68*)	0.40 (*1.06*)	8.27 (*1.62*)	6.40 (*1.96*)
Between-group comparisons ***p values***			*p*= .433	*p*= .764	*p*= .755	*p*=.618	*p*= .272

Note: Maximum score for each VLT is 30.00 and for WMTs are 18 for each (Forward and Backward).

The descriptive statistics in Table 5 suggest substantial progress in Recall and Recognition of the target vocabulary by both the 1-DS and 7-DS groups (within both the 4-day RI and 28-day RI sub-groups) at immediate post-test and sustained at delayed post-test. These Groups made slightly more improvement in their Recognition than in their Recall. Conversely, the Control Group's mean scores over the three timepoints reflected no change in either their Recognition or Recall.

Table 5

Descriptive statistical data for vocabulary tests (4-day RI and 28-day RI sub-groups)

			Recall			Recognition		
Group	RI	*n*	Pre-test	Immediate	Delayed	Pre-test	Immediate	Delayed
Control	4-day	16	0.56 (*1.55*)	0.88 (*1.71*)	0.63 (*1.31*)	5.31 (*3.79*)	5.50 (*4.05*)	6.75 (*3.82*)
	28-day	17	0.76 (*1.60*)	1.53 (*2.43*)	1.06 (*1.98*)	6.71 (*3.75*)	6.88 (*4.11*)	6.88 (*4.20*)
	All	33	0.67 (*1.55*)	1.21 (*2.10*)	0.85 (*1.68*)	6.03 (*3.78*)	6.21 (*4.08*)	6.82 (*3.96*)
1-DS	4-day	17	1.18 (*1.47*)	10.59 (*4.61*)	9.47 (*4.60*)	5.65 (*2.78*)	15.53 (*2.94*)	15.12 (*3.69*)
	28-day	17	0.59 (*1.00*)	8.94 (*6.15*)	6.41 (*5.19*)	5.35 (*3.61*)	16.06 (*1.35*)	15.12 (*2.55*)
	All	34	0.88 (*1.27*)	9.76 (*5.42*)	7.94 (*5.07*)	5.50 (*3.17*)	15.79 (*2.27*)	15.12 (*3.12*)
7-DS	4-day	14	0.64 (*1.08*)	9.50 (*4.75*)	9.43 (*4.52*)	5.57 (*3.37*)	16.07 (*1.49*)	15.71 (*1.77*)
	28-day	15	0.87 (*1.30*)	9.13 (*5.04*)	7.27 (*5.16*)	6.07 (*4.32*)	15.27 (*3.79*)	14.93 (*3.20*)
	All	29	0.76 (*1.19*)	9.31 (*4.82*)	8.31 (*4.90*)	5.83 (*3.83*)	15.66 (*2.89*)	15.31 (*2.59*)

Note: Maximum test scores of 17.00 for Recall and Recognition respectively

The following sub-sections will report the results of the mixed effects modelling to examine the impact of the intervention, practice distribution (1-DS / 7-DS) and retention interval (4-day / 28-day) on the students' Recall and Recognition of the target vocabulary at pre-, immediate and delayed post-tests.

Vocabulary recognition test

The final model (7) revealed a significant fixed effect of Time: $F(2, 190)= 358.256$, $p< .001$, and Group: $F(2, 95)= 57.608$, $p< .001$, while the Group*Time interaction was also statistically significant: $F(4, 190)= 79.823$, $p< .001$. Notably, there was no significant effect of Retention Interval (RI): $F(1, 95)= .833$, $p= .364$ and no significant interaction between RI*Time: $F(2, 190)= .942$, $p= .391$, or between RI*Group: $F(2, 95)= .561$, $p= .572$, and no significant three-way interaction between RI*Time*Group: $F(4, 190)= .782$, $p= .538$. The 2000 VLT: $F(1, 95)= 13.641$, $p< .001$, and forward WMT: $F(1, 95)= 5.618$, $p< .020$ were identified as significant covariates for the vocabulary recognition task.

Pairwise comparisons indicated that the significant Time*Group interaction reflected statistically significant differences in the learning trajectories of the experimental groups compared to the Control group over the three timepoints (Figure 2). This is reflected in large effect sizes for both experimental groups between the pre-test and immediate post-test and the pre-test and delayed post-test (see Table 6). There were no significant changes in the experimental Groups' scores on the Recognition task between the immediate and the delayed post-test, suggesting that both Groups maintained their higher scores at delayed post-test. In contrast, pairwise comparisons revealed no significant changes in the Control Group's scores on the Recognition test across the three timepoints, as reflected in the very small effect sizes and confidence intervals crossing zero.

Table 6

Pairwise comparisons over the three time points by Group (Recognition test)

		4-day RI			28-day RI		
		1-DS	7-DS	Control	1-DS	7-DS	Control
Pre-Post	*p*	0.001	0.001	1.000	0.001	0.001	1.000
	d (CIs)	3.45 (1.95, 4.95)	4.03 (2.21, 5.85)	.05 (-.93, 1.03)	3.93 (2.30, 5.56)	2.26 (.97, 3.56)	.04 (-.91, .99)
Pre-Delayed	*p*	0.001	0.001	0.126	0.001	0.001	1.000
	d (CIs)	2.90 (1.54, 4.26)	3.77 (2.02, 5.51)	.38 (-.61, 1.37)	3.13 (1.71, 4.54)	2.33 (1.02, 3.64)	.04 (-.91, .99
Post-Delayed	*p*	1.000	1.000	0.230	0.506	1.000	1.000
	d (CIs)	-.12 (-1.08, .83)	-.22 (-1.27, .83)	.32 (-.67, 1.30)	-.46 (-1.42, .50)	-.10 (-1.11, .92)	.00 (-.95, .95)

The lack of fixed effect or interactions for RI indicated that the timing of the delayed post-test (4-day/28-day RI) had no significant impact on participants' vocabulary Recognition scores at delayed post-test. Further, within each of the experimental groups (Control, 1-day, 7-day), there were no significant differences in the learning trajectories for the 4-day versus 28-day RI sub-groups across the three timepoints on the Recognition test (Figure 2).

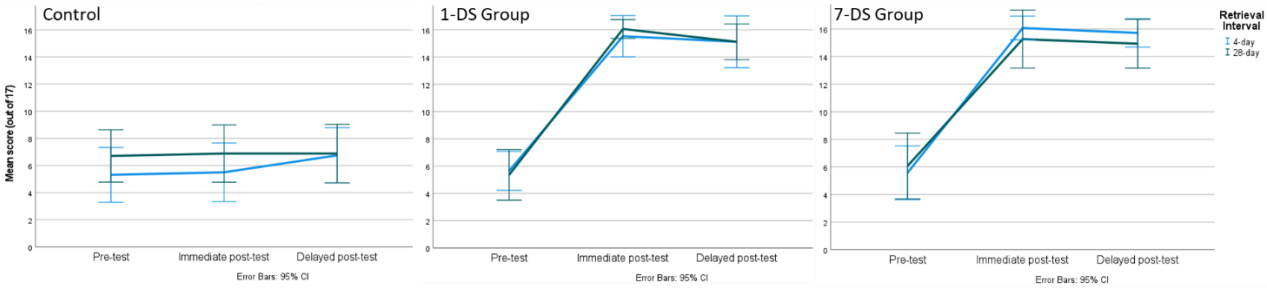


Figure 2: Mean Scores on Recognition Task by Group and Sub-group

Vocabulary recall

The final model (7) revealed a significant fixed effect of Time: $F(2, 190)= 202.82$, $p< .001$, and Group: $F(2, 95)= 41.764$, $p< .001$, and a significant Group*Time interaction: $F(4, 190)= 44.094$, $p< .001$ for the Recall test. No significant effect of RI: $F(1, 95)= 1.412$, $p= .238$ was found and no significant interaction between RI*Group: $F(2, 95)= 2.256$, $p= .11$ or between RI*Time*Group: $F(4, 190)= 1.021$, $p= .397$. However, a significant interaction between RI*Time was yielded: $F(2, 190)= 3.154$, $p= .045$. The 2,000 VLT: $F(1, 95)= 4.157$, $p= .044$, the 3,000 VLT: $F(1, 95)= 9.765$, $p= .002$ and 5,000 VLT: $F(1, 95)= 14.775$, $p< .001$ were identified as significant covariates.

In order to examine the significant interactions (Time*Group and RI*Time) in more detail, pairwise comparisons were run, revealing significant gains for both experimental groups between pre- and post-test and between pre- and delayed post-test; whereas there was no significant change in the Control group's performance across the three time points (Table 7).

Table 7

Pairwise comparisons over the three time points by Group (Recall test)

		4-day RI			28-day RI		
		1-DS	7-DS	Control	1-DS	7-DS	Control
Pre-Post	*p*	0.001	0.001	1.000	0.001	0.001	0.885
	d (CIs)	2.75 (1.42, 4.08)	2.53 (1.16 3.99)	.20 (-.79, 1.18)	1.90 (.75, 3.04)	2.24 (.95, 3.54)	.37 (-.59, 1.33)
Pre-Delayed	*p*	0.001	0.001	1.000	0.001	0.001	1.000
	d (CIs)	2.43 (1.18, 3.68)	2.68 (1.23, 4.12)	.05 (-.93, 1.03)	1.56 (.47, 2.64)	1.70 (.52, 2.88)	.17 (-.79, 1.12)
Post-Delayed	*p*	0.414	1.000	1.000	0.003	0.062	1.000
	d (CIs)	-.24 (-.91, .44)	-.02 (-.76, .73)	-.16 (-1.15, .82)	-.44 (-1.11, .25)	-.36 (-1.08, .37)	-.21 (-1.17, .74)

With regard to the significant interaction between RI*Time, pairwise comparisons indicated that this was due to a small but significant decrease in scores between post- and delayed post-test for the 28-day RI sub-group within the 1-DS group, although the effect size was small and the confidence interval crossed zero suggesting an unreliable effect (Table 7 and Figure 3). Further, although there was no significant difference between the 4-day and 28-day RI sub-groups within the 1-DS group at post-test (p= .092, d= -.30, CI= -.97, .38), there was a small, albeit unreliable, significant difference between these sub-groups at delayed post-test (p= .003, d= -.62, CI= -1.30, .08). There was no significant change in scores for the 28-day RI sub-group within the 7-DS group between post- and delayed post-test, nor for the 4-day RI sub-groups (Table 7). Additionally, there was no significant difference between the 4-day and 28-day sub-groups within the 7-DS group at post- (p= .832, d= -.08, CI= -.80, .66) or delayed post-test (p= .075, d= -.44, CI= -1.17, .31).

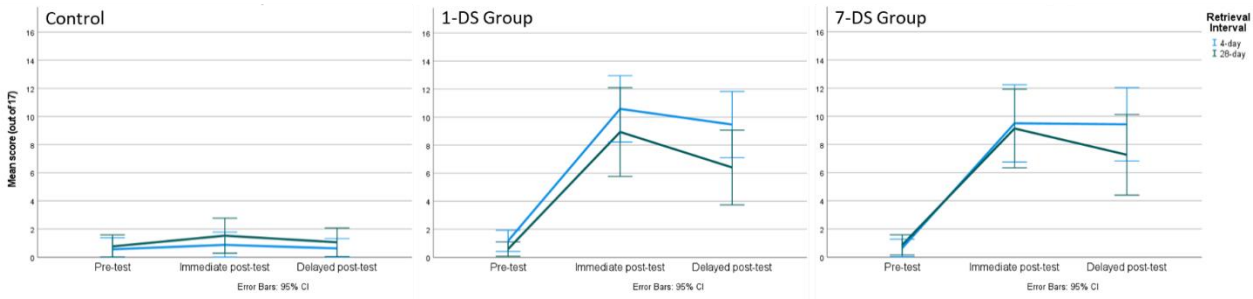


Figure 3: Mean scores for Recall task by Group and Sub-group

Discussion

In response to the first research question (RQ1), the test results demonstrated that intentional vocabulary learning via a CALL application (Quizlet) was effective for facilitating vocabulary learning amongst low proficiency English language learners in terms of Recall and Recognition of vocabulary. This was evidenced by the significant increase in the experimental groups' scores at immediate post-test, sustained at delayed post-test, and outperforming the Control group. Further, there was no change in the Control group's scores over the three test points, indicating limited interference from other possible factors during the intervention, such as the curriculum, test effect, and wider teaching methods deployed during English classes, which could have led to the experimental groups' progress by means other than Quizlet. Therefore, we can be confident that the improvement in the experimental groups' scores was due to the Quizlet intervention they received.

Of note for the Recognition and Recall tests was that both experimental Groups (1-DS and 7-DS) had improved significantly by the immediate post-test, sustaining this improvement

at delayed post-test (for both 4-day RI and 28-day RI). This finding is evident in the significant interaction for Group*Time within the mixed effects models for both tests and is further confirmed by the very large effect sizes (with confidence intervals that do not cross zero), reflecting substantial gains in the experimental Groups' performance on both tasks at post-test, which were maintained at delayed post-test. These findings indicate that both the short spacing (1-day ISI) and long spacing (7-day ISI) between practice sessions were effective for developing learners' ability to correctly recognise the meaning and produce the correct forms of the target words. It should be noted that overall gains on the Recall task were slightly smaller than on the Recognition task, and further, there was a significant decline in the Recall test results for the 1-DS 28-day RI sub-group at delayed post-test. This could be due to the learners being more successful at retaining receptive rather than productive knowledge (Schmitt, 2010). However, as shown in Figure 2, the effect sizes for pre-post gains were still very large. Further, the delayed post-test scores decreased by only a small amount and remained substantially higher than at baseline.

The TFA framework provides one interpretation of why Quizlet is effective. Analysis of the Quizlet activities using the TFA framework (Table 3) highlighted that the explicit, intentional vocabulary practice provided through Quizlet facilitates noticing, retrieval, retention and motivation. Hu and Nassaji (2016, p.31) claim that "No empirical studies, however, have yet examined the predictive power of TFA" in a real practice setting. Consequently, this study (conducted in a real practice setting with students of low English language proficiency) addresses this gap, supporting the TFA Framework as a powerful indicator for the effectiveness of vocabulary learning activities.

The experimental Groups' significant progress supports Tam et al. (2010) study, where CALL programmes were considered capable of creating an ideal environment for learners with limited English language proficiency. This reflects the positive impact of both short- and long-term use of Quizlet (Özer & Koçoğlu, 2017). However, only a few experimental studies (e.g., Korlu & Mede, 2018; Sanosi, 2018) have evaluated Quizlet's effectiveness, generally implementing only pre- and post-test. Therefore, the current study contributes to existing knowledge by demonstrating that explicit, intentional vocabulary practice via Quizlet successfully promoted longer-term retention (up to 28-days).

It is important to note that no comparison group was included in the present study. Therefore, it is not possible to make any claims regarding the relative effectiveness of Quizlet compared to other CALL applications. However, the results of this study have demonstrated that the explicit, intentional practice provided via Quizlet led to significant learning gains. This

finding is particularly notable given the characteristics of the participants involved in this study. As has been observed in numerous countries (see Introduction), students often arrive at university with substantially lower L2 English vocabulary knowledge than is needed for university-level study in English-medium contexts (Laufer, 2000). The present study has demonstrated that focussed bursts of explicit vocabulary practice can successfully facilitate low proficiency learners' vocabulary development in a relatively short space of time.

Turning now to RQ2, the 1-DS and 7-DS Groups (both 4-day and 28-day RI sub-groups) were found to make equal progress between the pre- and immediate post-tests, and to have maintained their progress by the delayed post-test, with no statistically significant differences observed between the experimental groups. This indicated that both the 1-day and 7-day spacing successfully promoted vocabulary learning and retention. Both spacing schedules facilitated the integration of new vocabulary items into short-term memory and their subsequent consolidation into long-term memory regardless of whether the practice was spaced by one or seven days (Cepeda et al., 2006; Rohrer & Pashler, 2007). These findings therefore do not support the prediction of a *lag effect* in relation to the learning of target vocabulary via explicit, intentional practice.

Within the study, the timing of the delayed post-test was counter-balanced across the experimental groups by assigning participants to a sub-group who either completed the delayed post-test at a 4-day or 28-day RI after the intervention. Existing research (Rohrer & Pashler, 2007) would predict that the 1-DS group would have an advantage at the 4-day delayed post-test, and the 7-DS group at the 28-day delayed post-test; reflecting the optimal ISI-RI ratios. Indeed, Serrano and Huang (2018) observed that shorter spacing supported immediate progress in vocabulary learning, while longer spacing sustained long-term retention. Furthermore, it was expected that a longer delay between practice sessions (i.e. 7-day spacing) would make the retrieval of previously learnt information more effortful, thereby reinforcing long-term memory (Toppino & Gerbier, 2014). Thus, the 7-DS Group was predicted to retain more knowledge by the delayed post-tests due to the longer delay between their practice sessions. However, in the present study, no significant fixed effect for Retention Interval (or interaction with Time and/or Group) was observed on the Recognition task, indicating that both Groups performed equally well in the 4-day and 28-day delayed post-tests. This finding contradicts Rohrer and Pashler's prediction that the interval between the final practice session and test needs to fall within an optimal time period. For the Recall task, a significant interaction between Time*RI was observed, which pairwise comparisons revealed was due to a small significant decrease at delayed post-test for the 28-day RI sub-group within the 1-DS group, whereas no corresponding

decrease was observed for either sub-group within the 7-DS group. This finding may in part support the prediction that shorter spacing between practice sessions will facilitate immediate learning but will be less favourable for longer-term knowledge retention. Nevertheless, this finding should be interpreted with caution, given that the effect size was small and unreliable due to the confidence intervals crossing zero. Further, as illustrated in Figure 3 the performance of both sub-groups within the 1-DS and 7-DS groups at delayed post-test remained well above pre-test, suggesting that the learners in all sub-groups had retained a substantial level of knowledge in terms of Recall, as well as Recognition, of the target vocabulary.

Notably, our findings are different to those yielded by Serrano and Huang (2018). However, it is worth acknowledging the key difference in the research design between the current study and that of Serrano and Huang. In Serrano and Huang's study, all participants in each experimental group completed one delayed post-test that matched their group's optimal spacing (4-day RI for the 1-day ISI group; 28-day RI for the 7-day ISI group). In contrast, the present study divided the Experimental Groups into sub-groups, with half of the participants from each group completing one of the delayed post-tests (either 4-day or 28-day RI), enabling direct comparison of each spacing schedule at each retention interval. This research design has contributed to resolving the limitation of Serrano and Huang's study and demonstrated that both more and less frequent practice schedules (1-day and 7-day spacing) can promote longer term vocabulary retention at 4 days and up to 28 days after the final practice session. Due to the design of Serrano and Huang's study, it is not possible to determine whether equivalent levels of retention would also have been observed for their groups at the 'non-optimal' RI (i.e. 28-day RI for the 1-day ISI group and 4-day RI for the 7-day ISI group).

The present study also addressed the limitation regarding lack of Control group, which has been common in distribution of practice studies to date (e.g. Küpper-Tetzel et al., 2014; Rogers & Cheung, 2018; Serrano & Huang, 2018), and found that the experimental groups' progress at post- and delayed post-test is unlikely to be due to a test effect or extraneous factors such as curriculum, teaching method, or teacher effects present outside of the study. Notably, in the present study, a small descriptive increase (see Figure 2) was observed in the Control group's scores at 4-day RI on the Recognition test. It is likely that this is due to a test effect, resulting from the short gap (4 days) following the immediate post-test. There is also the possibility that the Control group may have had incidental exposure to the target vocabulary during their regular English lessons, as real rather than pseudowords were used in the study. However, it is important to note that the change was non-significant, and the effect size remained very small with confidence intervals crossing zero, suggesting that it was an

unreliable effect. Further, their scores remained well below those of the experimental groups suggesting that this small increase is unlikely to reflect a significant learning effect. In contrast, a significant improvement (reflected in large effect sizes) was observed for both experimental groups at the 4-day RI delayed post-test, indicating that the learning gains brought about as a result of the intervention superseded any slight test effect that may have emerged at delayed post-test.

Overall, then our findings would seem to indicate that both shorter and longer spacing led to successful learning and retention of the target vocabulary for these learners. This finding aligns with that of Küpper-Tetzel et al. (2014), who also found that both shorter and longer spacing schedules led to long-term retention. Whilst existing research has clearly established the benefit of spacing practice sessions as opposed to massing practice into a single study session (e.g. Kornell, 2009; Nakata & Webb, 2016), research to date has produced contradictory findings regarding the optimal spacing for vocabulary learning and retention with some finding an advantage for longer-spacing (e.g. Serrano & Huang, 2018), some for shorter spacing (e.g. Rogers & Cheung, 2018) and others no difference between shorter or longer spacing (e.g. this study; Küpper-Tetzel et al., 2014). Therefore, existing research, including the present study, whilst supporting the need for spacing between practice sessions, is inconclusive regarding the 'optimal length' of that spacing. Indeed, in the present study, it would seem that both a 1-day and 7-day interval between practice sessions facilitated sufficiently effortful retrieval of the target vocabulary, but avoided complete forgetting, leading to strong long-term memory traces (Toppino & Gerbier, 2014) and therefore gains maintained at delayed post-test, regardless of whether it occurred after 4 or 28 days. It is important to note that there is substantial variation between existing studies, in terms of the age and context of learning, the target language structures, and the tasks used. It may be, then, that it is not possible to identify one optimal interval for all learners in all contexts and that this may vary based on individual learner as well as context-specific differences (Larsen-Freeman, 2014). Consequently, further research in real classroom settings is needed to explore the factors which impact learning under different spacing schedules.

Limitations and Recommendations for Future Research

The classroom-based context of this study necessitated the use of non-random sampling based on intact classes. Although this could be considered a limitation of the experimental design, steps were taken to ensure the comparability of the classes, including the initial analysis of the baseline measures (VLTs and WMTs) which revealed no significant differences. These

measures were also included in the mixed effects models as covariates alongside by-participant random intercepts, which enabled examination of the fixed effects, once any variance explained by these factors had been taken into account. Notably, the VLT scores (for both tests) and forward WM scores (for the Recognition test) were identified as significant control variables within the models. Therefore, future research is needed to further explore the role of working memory and vocabulary size in relation to explicit vocabulary learning and distribution of practice.

An additional consideration relates to the participants' English proficiency level. The sample was recruited from Level 1 classes (pre-elementary) on the Foundation Programme at one college in Oman. Consequently, the results may not be generalisable to other students with higher levels of English language proficiency, or to other institutions elsewhere. Finally, as noted above, despite the experimental groups making significant learning gains, we cannot conclude that Quizlet is the *most effective* software tool for explicit vocabulary learning, due to the lack of comparison group. Therefore, further intervention studies comparing vocabulary-learning software programmes are needed.

Whilst acknowledging the limitations above, the study findings clearly indicate the pedagogical benefits of explicit vocabulary learning through a variety of activities provided via Quizlet for low-ability learners. This corroborates Nation and Meara (2010), who support the use of explicit vocabulary teaching methods for learners in the initial stages of language learning. Furthermore, the study highlights the importance of providing spacing between practice sessions (either 1-day ISI or 7-day ISI) to allow for the transfer of information into the long-term memory. Finally, the study has provided evidence of the effectiveness of using the TFA Framework (Nation & Webb, 2011) to assess any vocabulary learning activities that might be used in the classroom.

Conclusion

There have been very few empirical studies (see Küpper-Tetzel et al., 2014; Rogers & Cheung, 2018; Serrano & Huang, 2018) investigating the lag effect in terms of vocabulary acquisition that have been conducted in an authentic learning setting. The results of this present study are generally consistent with those obtained by Küpper-Tetzel et al. (2014); i.e. that both short and long spacing of practice sessions lead to beneficial learning outcomes over the shorter and longer term. These results contradict the findings of Rogers and Cheung (2018), who support that shorter spacing between sessions leads to long-term retention. However, these differences could be due to different settings, the teaching method and contexts, as the present

study was conducted with adults in their first year of college, while the participants in previous studies were younger learners at primary or secondary school. Moreover, the type of task and number of training sessions differed. Therefore, more research is needed to get a clearer picture of the influence of the lag effect across different classroom settings, with different groups of learners, and the extent to which any benefits observed are impacted by the nature and number of practice sessions.

To summarise, the present study evaluated the effectiveness of a CALL application (Quizlet) for vocabulary learning and retention under two spaced practice schedules, with low-proficiency learners in a college classroom context. The findings indicated a lack of lag effects on vocabulary learning; i.e. no statistically significant differences between short (1-day) and long (7-day) intervals in terms of immediate knowledge development and longer-term knowledge retention (on either a 4-day or 28-day delayed post-test). The findings highlighted that the vocabulary activities were effective and efficient for vocabulary learning, an observation supported by the inclusion of a high proportion of features identified within the TFA Framework. Therefore, these findings suggest that the lag effect (i.e. benefit of longer spacing between practice sessions) may be less relevant when the practice itself is high quality.

In a wider pedagogical context, the findings of the current research clearly demonstrate the pedagogical advantages of intentional vocabulary learning, as supported by Nation and Meara (2010), for low proficiency language learners. Additionally, the findings illustrate the usefulness of one CALL application for providing explicit, intentional vocabulary practice via digital learning activities to promote vocabulary learning amongst low-proficiency learners. Further research in real learning contexts with diverse participant groups and a wider range of CALL applications is needed to further ameliorate the effectiveness of CALL tools for facilitating vocabulary teaching and learning. Finally, in line with Hu and Nassaji's (2016) recommendation to include more features of TFA in vocabulary learning activities, the researchers suggest using the TFA Framework (Nation & Webb, 2011) to assess activities planned for use in the classroom to ensure that activities offer sufficient opportunities for vocabulary practice, rehearsal and retrieval to deepen processing and promote long-term vocabulary retention.

Acknowledgements

We would like to thank the English Language Centre's teachers and students who volunteered to participate and co-operate freely in this study.

References

Ahmadi, M. (2014). Semantic and structural elaboration in L2 vocabulary learning and retention. *Procedia-Social and Behavioral Sciences, 98*, 109-115. https://doi.org/10.1016/j.sbspro.2014.03.395

AL-Hammadi, F. S. (2012). The role of recognition memory in L2 development. *Journal of King Saud University-Languages and Translation, 24*(2), 83-93. https://doi.org/10.1016/j.jksult.2012.05.003

Al-Khatib, H. (2011). Technology enhanced learning: Virtual realities; Concrete results case study on the impact of TEL on learning. *European Journal of Open, Distance and E-Learning*. *14*(1). 1–12. https://eric.ed.gov/?id=EJ936385

Al-Mamari, A. S. (2012, February 20-21). *General foundation program in higher education institutions in Oman national standards: Implementation and challenges* [Paper presentation]. Oman Quality Network Regional Conference, Muscat, Oman. https://omjournal.org/images/229_M_Deatials_Pdf_.pdf

Alqarni, I. R. (2019). Receptive vocabulary size of male and female Saudi English major graduates. *International Journal of English Linguistics*, *9*(1), 111-119. https://doi.org/10.5539/ijel.v9n1p111

Anjaniputra, A.G. & Salsabila, V. A. (2018). The merits of Quizlet for vocabulary learning at tertiary level. *Indonesian EFL Journal, 4*(2), 1-11. https://doi.org/10.25134/ieflj.v4i2.1370

Averianova, I. (2015). Vocabulary acquisition in L2: Does CALL really help? *EUROCALL Conference* (pp. 30-35). Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000306

Baddeley, A. D. (1990). Working memory. *System*, 255-559. http://psych.colorado.edu/~kimlab/baddeley.1992.pdf

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (vol. 8, pp. 47-89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Barr, B.W.B. (2016). Checking the effectiveness of Quizlet as a tool for vocabulary learning. *The Center of EFL Journal, 1*(2), 36-48. https://doi.org/10.15045/ELF_0020104.

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System, 27*(2), 223-247. https://doi.org/10.1016/S0346-251X(99)00018-4

Bauman, J., & Culligan, B. (1995). About the general service list. http://jbauman.com/gsl.html

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics, 31*(4), 635–650. https://doi.org/10.1017/S0142716410000470

Chien, C.W. (2015). Analysis of the effectiveness of three online vocabulary flashcard websites on L2 learners' level of lexical knowledge. *English Language Teaching, 8*(5), 111-121. https://doi.org/10.5539/elt.v8n5p111

Childers, J.B. & Tomasello, M. (2006). Are nouns easier to learn than verbs? Three experimental studies. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs* (pp. 311-355). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195170009.003.0013

Climie, E. A., & Rostad, K. (2011). Test review: Wechsler adult intelligence scale. *Journal of Psychoeducational Assessment, 29*(6), 581-586. https://doi.org/10.1177/0734282911408707

Cobb, T. (n.d.) *Compleat Web VP v.2.6* [computer program]. https://www.lextutor.ca/vp/comp/

Crandell, E.R. (2017). *Quizlet flashcard for the first 500 words of the Academic Vocabulary List* (Unpublished MA dissertation). Brigham Young University. https://scholarsarchive.byu.edu/etd/6335

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://link.springer.com/article/10.1007/BF02310555

Dizon, G. (2016). Quizlet in the EFL classroom: Enhancing academic vocabulary acquisition of Japanese university students. *Teaching English with Technology, 16*(2), 40-56. https://eric.ed.gov/?id=EJ1135913

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning, 61*(2), 367-413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Elsayyad, H. M. (2014). *The relationship between working memory and reading comprehension in L1 Arabic and L2 English for Arabic speaking children* (Unpublished doctoral dissertation). Bath Spa University. http://researchspace.bathspa.ac.uk/5175/1/Elsayyad%20thesis%20final.pdf

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481-505. https://doi.org/10.1093/applin/amy057

Heatley, A., Nation, I.S.P. & Coxhead, A. (2002). Range and Frequency Programs. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207-223. https://www.lextutor.ca/cv/beyond_a_clockwork_orange.html

Hu, H. M. & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement load hypothesis versus technique feature analysis. *System, 56*(1), 28-39. https://doi.org/10.1016/j.system.2015.11.001

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Eds.). *Cognition and Second Language Instruction* (pp 258-287). Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139524780.011

Jackson III, D.B. (2015). A targeted role for L1 in L2 vocabulary acquisition with mobile learning technology. *TESOL Arabia Perspectives*, *23*(1), 6-11. http://issuu.com/tesolarabia-perspectives/docs/feb2015-perspectives

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics, 21*(1), 47–77. https://doi.org/10.1093/applin/21.1.47

Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition, 24*(4), 617-637. http://doi.org/10.1017/S0272263102004047

Kamanpoori, H. (2014, May 5). Making 12 years of effective English learning at schools is important. *Oman Observer*. http://omanobserver.om/making-12-years-of-effective-english-learning-at-schools-is-important/

Kasprowicz, R.E., Marsden, E. & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. *The Modern Language Journal*, *103*(3), 580-606. https://doi.org/10.1111/modl.12586

Keating, G.D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365-386. https://doi.org/10.1177/1362168808089922

Korlu, H. & Mede, E. (2018). Autonomy in vocabulary learning of Turkish EFL learners. *The EuroCALL Review*, *26*(2), 58-70. https://doi.org/10.4995/eurocall.2018.10425

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297-1317. https://doi.org/10.1002/acp.1537

Krish, P., Hussin, S. & Sivapuniam, N. (2011). Language learning and language acquisition in online forums. *3L: Language, Linguistics, Literature*, *17*(2), 91-100. http://journalarticle.ukm.my/3222/1/10-Pramela_Krish_et_al.pdf

Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science, 42*(3), 373-388. http://www.jstor.org/stable/43575234

Larsen-Freeman, D. (2014, March 29). *Complexity theory: Renewing our understanding of language, ;earning, and teaching* [Keynote presentation]. International TESOL Convention. Portland, OR, United States. http://www.tesol.org/attend-and-learn/international-

Laufer, B. (2000). Task effect on instructed vocabulary learning: The hypothesis of 'involvement'. *Selected Papers from AILA'99 Tokyo* (pp. 47-62). Waseda University Press. https://www.academia.edu/10234193/Task_effect_on_instructed_vocabulary_learning_the_hypothesis_of_involvement

Laufer, B. (2006). Comparing focus on form and focus on forms in second-language vocabulary learning. *Canadian Modern Language Review*, *63*(1), 149-166. https://doi.org/10.1353/cml.2006.0047

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399-436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Linck, J., Osthus, P., Koeth, J., & Bunting, M. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic bulletin and Review*, *21*, 861-883. https://doi.org/10.3758/s13423-013-0565-2

Ma, Q., & Kelly, P. (2006). Computer assisted vocabulary learning: Design and evaluation. *Computer Assisted Language Learning*, *19*(1), 15-45. https://doi.org/10.1080/09588220600803998

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 Vocabulary Acquisition, Knowl- edge and Use: New Perspectives on Assessment and Corpus Analysis* (2nd ed., pp. 57-78). European association of second Language Acquisition Monogtaphs. https://www.researchgate.net/publication/256477266_L2_vocabulary_acquisition_knowledge_and_use_New_perspectives_on_assessment_and_corpus_analysis

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL*, *20*(1), 3-20. https://doi.org/10.1017/S0958344008000219

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning?. *Studies in Second Language Acquisition*, *37*(4), 677-711. https://doi.org/10.1017/S0272263114000825

Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, *41*(2), 287-311. https://doi.org/10.1017/S0272263118000219

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning?. *Studies in Second Language Acquisition*, *38*(3), 523-552. https://doi.org/10.1017/S0272263115000236

Nation, I. S. P. (2001) *Learning vocabulary in another language* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review, 63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59
Nation, I. S. P., & Meara, P. (2010) Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (2nd ed., pp. 34-52). Edward Arnold. https://www.academia.edu/37498389/_Norbert_Schmitt_ed_An_Introduction_to_Applied_b_ok_xyz_1_
Nation, I.S.P. (2011). Research into practice: Vocabulary. *Language Teaching, 44*(4), 529-539. https://eric.ed.gov/?id=EJ956199
Nation, I.S.P. (2013). *Learning Vocabulary in Another Language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656
Nation, I.S.P. & Webb, S. (2011). *Researching and Analyzing Vocabulary.* Heinle. https://doi.org/10.1002/tesq.58
Nunan, D. (1999). *Second Language Teaching & Learning*. Heinle & Heinle Publishers. https://eric.ed.gov/?id=ED441344
Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes, 18*(2), 161-175. https://eric.ed.gov/?id=EJ577545
Özer, Y.E. & Koçoğlu, Z. (2017). The use of Quizlet flashcard software and its effects on vocabulary learning. *Language Journal, 168*(1), 61-81. https://doi.org/10.1501/DILDER_0000000238
Plonsky, L. & Oswald, F.L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. https://doi.org/10.1111/lang.12079
Quizlet. (2020). *Our mission* | Quizlet. Retrieved August 17, 2020 from https://quizlet.com/mission
Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly, 49*(4), 857-866. https://doi.org/10.1002/tesq.252
Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics, 38*(6), 906-911. https://doi.org/10.1093/applin/amw052
Rogers, J. & Cheung, A. (2018). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research, 24*(5), 616-641. https://doi.org/10.1177/1362168818805251
Rohrer, D. & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, *16*(4), 183-186. https://doi.org/10.1111/j.1467-8721.2007.00500.x
Ruiz, S., Rebuschat, P. & Meurers, D. (2021). The effects of working memory and declarative memory on instructed second language vocabulary learning: Insights from intelligent CALL. *Language Teaching Research, 25*(4), 510-539. https://doi.org/10.1177/1362168819872859
Sanosi, A.B. (2018). The Effect of Quizlet on vocabulary acquisition. *Asian Journal of Education and e-Learning*, *6*(4) 71-77. https://doi.org/10.24203/ajeel.v6i4.5446
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. https://doi.org/10.1177/026553220101800103
Serrano, R. & Huang, H.Y. (2018). Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text?. *TESOL Quarterly*. *52*(4), 971-994. https://doi.org/10.1002/tesq.445
Shillaw, J. (1995). Using a word list as a focus for vocabulary learning. *The Language Teacher, 19*(2), 58-59. https://www.sid.ir/paper/540794/en
Smith, C.B. (2003). Vocabulary development: Elaboration for writing. *ERIC Topical Bibliography and Commentary.* https://eric.ed.gov/?id=ED480888

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning, 67*(3), 512–545. https://doi.org/10.1111/lang.12236

Suzuki, Y. & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning, 65*(4), 860-895. https://doi.org/10.1111/lang.12138

Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research, 21*(2), 166-188. https://doi.org/10.1177/1362168815617334

Swain, M. (2005). The output hypothesis: Theory and research. In *Handbook of Research in Second Language Teaching and Learning* (pp. 495-508). Routledge. https://doi.org/10.4324/9781410612700-38

Tam, S.S., Kan, N.H. & Ng, L.L. (2010). Low proficiency learners in synchronous computer-assisted and face-to-face interactions. *Turkish Online Journal of Educational Technology-TOJET, 9*(3), 61-75. https://eric.ed.gov/?id=EJ898015

Teng & Zhang (2021). The associations between working memory and the effects of multimedia input on L2 vocabulary learning [Special issue]. *International Review of Applied Linguistics in Language Teaching, AOP*, 1-29. https://doi.org/10.1515/iral-2021-0130

Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. *Psychology of Learning and Motivation, 60*, 113–189. https://doi.org/10.1016/B978-0-12-800090-8.00004-4

Tung, H-T. (2015). Intentional vocabulary learning using digital flashcards. *English Language Teaching, 8*(10), 107-112. https://doi.org/10.5539/elt.v8n10p107

Webb, S. (2005). Receptive and productive vocabulary learning: The effect of reading and writing on word knowledge, *Studies in Second Language Acquisition, 27*(1), 33-52. https://doi.org/10.1017/S0272263105050023

Zambrano Acosta, E.J. (2018). *The Quizlet App and English Language Vocabulary Learning* (Unpublished Master's thesis). Universidad Técnica de Ambato. http://repositorio.uta.edu.ec/jspui/handle/123456789/29097

Appendix 1 Target Word list

Level 2 (Elementary Level) Vocabulary List (Nouns)

No.	Target words	L1 translation
Academic Word List (AWL)		
1.	Access	دخول
2.	Aid	مساعدة
3.	Area	منطقة
4.	Benefit	فائدة
5.	Community	مجتمع
6.	Consumer	مستهلك
7.	Device	جهاز
8.	Economy	اقتصاد
9.	Environment	بيئة
10.	Export	تصدير
11.	Globe	الكرة الارضية
12.	Image	صورة
13.	Instance	مثال
14.	Principal	مدير مدرسة
15.	Project	مشروع
16.	Resident	مقيم
17.	Style	أسلوب
18.	Team	فريق
2000 Level of Frequency		
19.	Balance	توازن
20.	Century	قرن
21.	Education	التعليم
22.	Flood	فيضان
23.	Hunter	صياد
24.	Information	معلومات
25.	Journey	رحلة
26.	Manager	مدير
27.	Ocean	محيط
28.	Skill	مهارة
29.	Solution	حل
30.	Storm	عاصفة
31.	Temperature	درجة الحرارة
32.	Tools	أدوات
33.	Tourists	سياح
34.	Weather	الطقس

Appendix 2 Working Memory Test and Research Protocol

Research Protocol
Digit span tests

Subtest 1: Number Memory Forward
(Rote learning & memory/Attention/Encoding/Auditory Processing)
This subtest is designed to show how well the student can retain simple sequences of auditory information.

Requires

- Each test component has 8 items with trials of same length.
- Make sure the student feel comfortable and secure, and there are not any interruptions or noise around.
- Read each trial verbatim at the rate of one digit per second

Materials

1- Administration and scoring manual
2- Record form
3- Voice recorder
4- Attendance sheet

Note: there are no time limits, but examiner must be mindful of the rate of passing seconds

Procedures

- On arrival ask the student for their name and group number to assign them in their correct group list (massed or spaced). Check that they have already signed the consent and ask the student to sign the test attendance sheet.
- Start portable voice recorder. Say the student's name, the group name (massed or spaced) and the date.
- The researcher gives instructions for the subtest:

"I'm going to say sets of numbers; when I'm finished with each set, you repeat them back to me in the same order as you heard them."
"Don't worry if you can't remember everything, but try to say as much as you can and to speak clearly."
"First we'll practice. Listen carefully; I can't repeat them once we start. Ready?"

Practice the task:
Make sure the student can hear you well and know what is needed. First, the student need to be trained on the subtest for ' one attempt'.
Example A.
Say: **" Three** (pause) **Eight** (pause). **What numbers did you hear?"**
When the student finished, say **"Ok, now we'll do the test".**

The subtest 1 (Forward) and scoring

الاختبار وحساب درجات الاختبار

Start with the example above. Discontinue when student has made two consecutive 0-point responds.

numbers الأرقام									The correct answer الإجابة الصحيحة	Score الدرجات
6	4									
2	5									
3	1	6								
7	4	9								
6	9	5	7							
3	6	2	9							
8	3	9	4	6						
5	1	7	2	9						
4	2	5	1	8	7					
5	8	4	9	3	6					
1	5	2	8	4	9	7				
8	2	4	7	3	6	1				
9	3	7	5	1	6	4	8			
2	6	4	8	3	7	1	5			
3	8	1	9	5	2	7	4	6		
6	9	5	3	8	1	4	7	2		

Score

1. Score 1 point if the student gives a correct response.
2. Score 0 points if the student gives an incorrect response, say that they do not know the answer, or does not respond within approximately 30 seconds.

Test Protocol
Digit span tests

Subtest 2: Number Memory Reversed (Backward)
(working memory/Transformation of information/Mental manipulation/Visuo-spatial imaging)
This subtest is designed to show how well the student can retain and manipulate simple sequences of auditory information.

Requires

- Each test component has 8 items with trials of same length.
- Make sure the student feel comfortable and secure, and there are not any interruptions or noise around.
- Put a break between the two subtests (forward & backward).
- Read each trial verbatim at the rate of one digit per second

Materials

5- Administration and scoring manual
6- Record form
7- Voice recorder
8- Attendance sheet

Note: there are no time limits, but examiner must be mindful of the rate of passing seconds

Procedures

- On arrival ask the student for their name and group number to assign them in their correct group list (massed or spaced). Ask the student to sign the test attendance sheet.
- Start portable voice recorder. Say the student's name, the group name (massed or spaced) and the date.
- The researcher gives instructions for the subtest:

"I'm going to say sets of numbers; when I'm finished with each set, you repeat them back to me in reversed order."
"So if I say '4 – 1', you say '1-4'."
"First we'll practice. Listen carefully; I can't repeat them once we start. Ready?"

Practice the task:
Make sure the student can hear you well and know what is needed. First, the student need to be trained on the subtest for ' two attempts'.
Example A. 9 - 2
Say: **" Nine** (pause) **Two** (pause). **Now tell me the numbers in reversed order."** (2 – 9)
When the student finished, say **"Ok, now we'll do another set",** and go to Example B.
Example B. 1 - 6
Say: **" One** (pause) **Six** (pause). **"Now tell me the numbers in reversed order."** (6 – 1)
When the student finished, say **"Ok, now we'll go on the test items".**

The subtest 2 (Backward) and scoring

الاختبار وحساب درجات الاختبار

Start with the example above. Discontinue when student has made two consecutive 0-point responds.

numbers الأرقام									The correct answer الإجابة الصحيحة	Score الدرجات
7	3								3 7	
4	9								9 4	
5	2	8							8 2 5	
6	9	2							2 9 6	
4	9	5	3						3 5 9 4	
7	1	6	8						8 6 1 7	
3	7	5	8	1					1 8 5 7 3	
2	9	4	6	3					3 6 4 9 2	
8	2	5	1	9	4				4 9 1 5 2 8	
1	7	4	8	5	9				9 5 8 4 7 1	
4	9	1	7	3	5	8			8 5 3 7 1 9 4	
6	2	9	1	4	7	3			3 7 4 1 9 2 6	
8	1	6	4	9	7	2	5		5 2 7 9 4 6 1 8	
9	3	7	5	1	4	2	8		8 2 4 1 5 7 3 9	
3	6	8	4	2	7	1	5	9	9 5 1 7 2 4 8 6 3	
4	8	2	1	9	5	7	3	6	6 3 7 5 9 1 2 8 4	

Score

1. Score 1 point if the student gives a correct response.
2. Score 0 points if the student gives an incorrect response, say that they do not know the answer, or does not respond within approximately 30 seconds.

Appendix 3 Vocabulary Test Sample

Example from 'Active' Recall test

Test 1: Translate the following words into English.

1. _____ مقيم
2. _____ مدير مدرسة

Example from 'Passive' Recall test

Test 2: Translate the following words into Arabic.

1. access _____
2. area _____

Example from 'Active' Recognition test

Test 3: Select the correct English equivalent for each of the following words and circle it.

1. مدير
   a. manager b. writer c. connector d. agreement
2. مهارة
   a. skill b. sleep c. management d. article

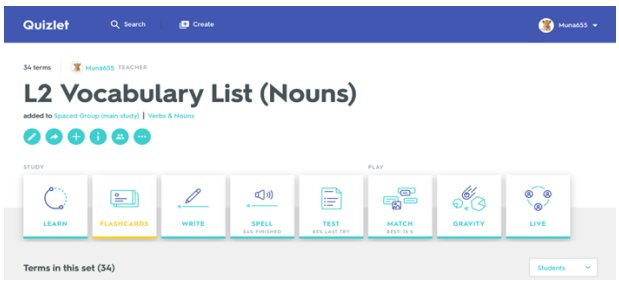
Example from 'Passive' Recognition test

Test 4: Select the correct L1 translation for each of the following words and circle it.

1. device
   a. مطبخ b. اختيار c. مجلد d. جهاز
2. information
   a. معلومات b. رسائل c. اخبار d. تقارير

Appendix 4 Design of Sessions (Time & Activity) with Screenshots from the Intervention Materials

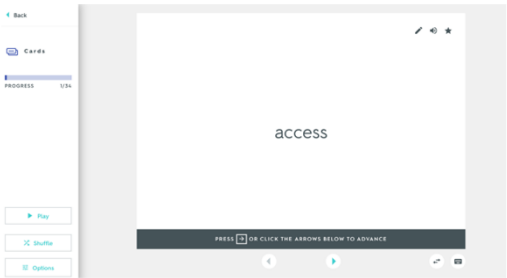
Quizlet's Flashcard Set Landing Page



Session One

	Activity	Method	Type	Time	Repetition
1.	Flashcards	Digital flashcards presenting target L2 word on one side, and L1 meaning on reverse side. Quizlet automatically cycles through all target words one-by-one, showing the L2 word before automatically flipping to reveal L1 meaning.	Recognition	3 mins	1
2.	Flashcards	Player 1 displays flashcard showing L2 word and asks player 2 to guess the meaning before clicking to view L1 meaning on reverse side. Players swap roles working through the set of target words.	Recall/ Recognition	10 mins	2
3.	Match	Each player has to match up each L2 word with the correct L1 meaning.	Recognition	7 mins	1
	Total			20 mins	4

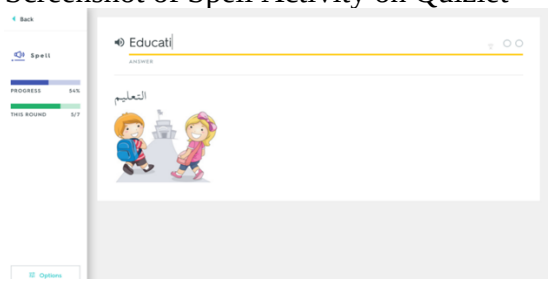
Screenshot of Study Flashcards on Quizlet



Session Two

	Activity	Method	Type	Time	Repetition
1.	Flashcards	Digital flashcards presenting target L2 word on one side, and L1 meaning on reverse side. Quizlet cycles through all target words one-by-one, showing the L2 word before automatically flipping to reveal L1 meaning.	Recognition	3 mins	1
2.	Spell	Player hears L2 word and has to type correct L2 word form. If incorrect, Quizlet reveals missing letters and player tries again until they type correct word, before moving onto the next item.	Recall/ Recognition	17 mins	2
	Total			20 mins	3

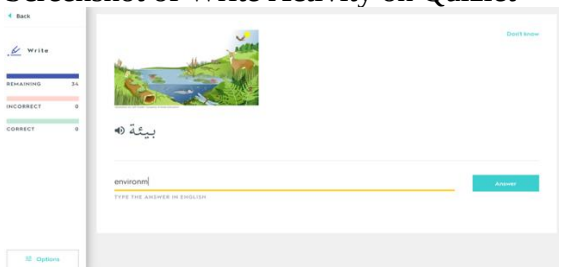
Screenshot of Spell Activity on Quizlet



Session Three

	Activity	Method	Type	Time	Repetition
1.	Flashcards	Digital flashcards presenting target L2 word on one side, and L1 meaning on reverse side. Quizlet cycles through all target words one-by-one, showing the L2 word before automatically flipping to reveal L1 meaning.	Recognition	3 mins	1
2.	Write	L1 meaning is displayed, player has to write the correct L2 word	Recall	17 mins	2
	Total			20 mins	3

Screenshot of Write Activity on Quizlet



Session Four

	Activity	Method	Type	Time	Repetition
1.	Flashcards	Digital flashcards presenting target L2 word on one side, and L1 meaning on reverse side. Quizlet cycles through all target words one-by-one, showing the L2 word before automatically flipping to reveal L1 meaning.	Recognition	3 mins	1
2.	Test	Mixture of multiple choice, Write and Match, and True/False questions used to test recognition and recall of target L2 words.	Recall/ Recognition	17 mins	2
	Total			20 mins	3

Screenshot of Test on Quizlet