Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

6-6-2022 4:00 PM

Spaced Practice and Second Language Vocabulary Learning

Sukyung Kim, *The University of Western Ontario*

Supervisor: Webb, Stuart, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Education

© Sukyung Kim 2022

Follow this and additional works at: https://ir.lib.uwo.ca/etd

 Part of the Language and Literacy Education Commons

Recommended Citation

Kim, Sukyung, "Spaced Practice and Second Language Vocabulary Learning" (2022). *Electronic Thesis and Dissertation Repository*. 8600.
https://ir.lib.uwo.ca/etd/8600

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

This thesis aims to investigate whether learners can increase second or foreign language (L2) vocabulary learning through spaced practice, in which repeated practice is spaced out in time or through other intervening events. It is well acknowledged that spaced practice promotes learning and enhances retention. Despite robust positive effects of spaced practice in learning and memory, the degree to which spaced practice effects are meaningful for L2 learning is still not clear. For example, the majority of spaced practice studies on L2 vocabulary learning has focused on paired-associate learning (e.g., flashcard learning). There are many different learning activities for vocabulary learning, and more research investigating the effects of spaced practice in different vocabulary learning conditions is warranted. This thesis is made up of three studies in the integrated article format and is organized into five chapters: An introduction to the topic of spaced practice (Chapter 1), the three studies (Chapters 2, 3, and 4), and a concluding chapter (Chapter 5).

Study 1 (Chapter 2) meta-analyzed earlier studies of spaced practice in L2 learning. 98 effect sizes from 48 experiments (N = 3,411) were retrieved. This study compared the effects of three aspects of spacing (spaced vs. massed, longer vs. shorter spacing, and equal vs. expanding spacing) on immediate and delayed posttests to calculate mean effect sizes. This study also examined the extent to which nine empirically motivated variables moderated the effects of spaced practice. Results showed that (a) spacing had a medium-to-large effect on L2 learning; (b) shorter spacing was as effective as longer spacing in immediate posttests but was less effective in delayed posttests than longer spacing; (c) equal and expanding spacing were statistically equivalent; and (d) variability in spacing effect size across studies was explained methodologically by the learning target, number of sessions, type of practice, activity type, feedback timing, and retention interval. This study has already been published in the journal *Language Learning* (Wiley).

Study 2 (Chapter 3) examined the effects of spaced practice on L2 vocabulary learning through fill-in-the-blanks and flashcards activities. 150 Korean learners were divided into five groups: one control (no treatment) and four experimental groups, based on learning condition (fill-in-the-blanks vs. flashcards) and spacing type (massed [no spacing interval] vs.

spaced [1-day interval]). The participants studied forty-eight low frequency English words. Results showed that the effects of spaced practice were greater for fill-in-the-blanks than flashcards on an immediate posttest and that spaced practice was more effective than massed practice for both activities on a 2-week delayed posttest. The results suggest that fill-in-the-blanks may be affected by spacing in the same way as flashcards. This study is currently under review at the journal *The Modern Language Journal* (Wiley).

Study 3 (Chapter 4) examined the effects of spaced practice on the learning and retention of forty-eight low frequency English words through sentence production and flashcards activities. 150 Korean university students were randomly assigned to five groups: one control (no treatment) and four experimental groups, based on learning condition (sentence production versus flashcards) and spacing schedule (massed [no interval] versus spaced [1-day interval]). Results showed that spaced practice was as effective as massed practice in vocabulary learning for sentence production and flashcards activities on an immediate posttest but that spaced practice was more effective than massed practice for both activities on a 2-week delayed posttest. This suggests that both activities may be affected similarly by spacing. This study is currently under review at the journal *TESOL Quarterly* (Wiley).

Taken as a whole, the current thesis showed large effects of spaced practice on L2 vocabulary learning and retention but the effects seemed to depend on how words were learned (e.g., whether the practice is spaced within a session or between multiple sessions; whether retrieval practice is provided or not). The thesis also showed that spaced practice may contribute to vocabulary learning in other ways apart from flashcards. Pedagogically, the findings suggest that it may be useful for teachers and students to use spacing when scheduling activities for practice repetitions inside and outside classroom. The findings of the three studies in the current thesis are important because they show the value of spacing in other L2 vocabulary learning conditions. This thesis then concludes with methodological and pedagogical implications for L2 vocabulary learning as well as suggestions for future research.

Keywords: Spaced practice, Second language vocabulary learning, Meta-analysis, Second language learning, Vocabulary learning activities, Fill-in-the-blanks, Flashcards, Sentence production, Transfer appropriate processing, Feedback timing

Summary for Lay Audience

Learners study second or foreign language (L2) words in language classrooms, but they often forget the words. Encountering words repeatedly (i.e., repeated practice) and testing studied words contribute to vocabulary learning. Furthermore, when repeated practice is spaced out in time or through other intervening events (i.e., spaced practice), the potential for learning and retention improves. This thesis investigates whether learners can increase L2 word learning through spaced practice. It consists of three articles focusing on effects of spaced practice. First, to clarify the overall effects of spaced practice, Study 1 systematically reviewed earlier studies of spaced practice in L2 learning. It is widely acknowledged that spaced practice has a positive effect on flashcard learning. To examine whether other vocabulary learning activities are affected by spaced practice, Study 2 compared fill-in-the-blanks activities to flashcards. Study 3 compared sentence production activities to flashcards. Results showed that spaced practice benefits L2 learning but the effects seemed to depend on how words were learned (e.g., the number of learning sessions, whether studied words were tested or not) (Study 1). Results also showed that fill-in-the-blanks and sentence production activities may be affected by spaced practice in the same way as flashcards (Studies 2 and 3). These findings suggest the value of spaced practice occurs with other L2 vocabulary learning conditions. This thesis concludes with methodological and pedagogical implications for L2 vocabulary learning.

Co-Authorship Statement

Three studies (Chapters 2, 3, and 4) in this thesis are co-authored papers. Study 1 has already been published in the journal *Language Learning* (Wiley). Studies 2 and 3 are currently under review. Study 2 was submitted to *Studies in Second Language Acquisition* (Cambridge University Press) on January 25th, 2022. Study 3 was submitted to *TESOL Quarterly* (Wiley) on February 25th, 2022. I am responsible for designing the computer-based materials, collecting, processing, and analyzing the data, and preparing the manuscript. The contribution made by the other author was in the form of generating research ideas and questions, developing materials (e.g., target words, activities, tests), providing consultation and feedback throughout all stages of this project, and finalizing and editing the manuscript.

Acknowledgements

I would like to express my heartfelt appreciation to many people and organizations. This research could not have been completed without their great support. First, I would like to express my deepest thanks to my supervisor, Stuart Webb for his guidance, encouragement, and faith throughout this long journey. He helped me believe that this research was possible and made the research process a positive one. I would also like to thank Dr. Frank Boers and Dr. Batia Laufer for their comments at an early stage of my PhD research. I would like to extend my gratitude to my other committee members, Dr. Irina Elgort, Dr. Barbara Fenesi, and Dr. Laura Batterink, for their perspectives and thoughtful commentary.

I could not have completed this research without Dr. Junkyu Lee and Dr. Hyunjung Kim, and people working at Hankuk University of Foreign Studies for helping me collect data and supporting me, Dr. Yun Lee and Dr. Hyunsook Yoon for their help and advice from the beginning of my academic career, and Korean students in Canada and South Korea who participated in the pilot study and main experiments. I would also like to thank to Dr. Tatsuya Nakata for his skeptical and helpful comments throughout my PhD, Dr. Paul Tremblay for his advice and suggestions on statistical analyses in the three studies, and the vocabulary group members at Western. A very special thank you to Sunho Choi for his assistance and great support throughout my academic years.

Finally, I would like to dedicate this work to my family. My special thanks go to my parents and my brother, Si Uk, for their love and support during my whole academic years and always. Many thanks also go to HaeRyun, Kim, Junghwan Choi, Jungwoo, Choi, and Eunjung Gil.

Table of Contents

List of Tables

List of Figures

Chapter 1

No Figures

Chapter 2

List of Appendices

Chapter 2

Chapters 3 and 4

List of Abbreviations

BNC	British National Corpus
COCA	Corpus of Contemporary American English
CI	Confidence Interval
EFL	English as a Foreign Language
L1	First Language
L2	Second Language
M	Mean
RI	Retention Interval
SD	Standard Deviation
VLT	Vocabulary Levels Test

Chapter 1: Introduction

The purpose of this thesis is to investigate whether learners can increase second or foreign language (L2) vocabulary learning through spaced practice. When learners are provided with new words in the classroom and they can produce the correct answer to questions addressing the words, we may say that they have learned the words. However, what we learn tends to be easily forgotten (Baddeley, 1999). Certainly, learners of English as a foreign language (EFL) know a relatively small proportion of the vocabulary known by adult native speakers of English (Siyanova-Chanturia & Webb, 2016; Webb & Nation, 2017). This suggests that more effective ways of learning and retaining words are needed.

There are many ways to learn words. Morgan and Rinvolucri (2004) described 118 activities to develop vocabulary knowledge, and Webb and Nation (2017) profiled 23 approaches to learning vocabulary. Many studies of deliberate learning of L2 vocabulary have demonstrated that words can be learned through activities. Learning words from flashcards and word lists leads to gains in knowledge of form-meaning connection (e.g., Elgort, 2011; Mondria & Wiersma, 2004; Nakata, 2008; Webb, 2009). Learning words through fill-in-the-blanks has also shown positive effects in vocabulary learning (e.g., Folse, 2006; Rott, 2012). Writing words in sentences had also been found to contribute to vocabulary learning (e.g., Javanbakht, 2011; Webb, 2005). The degree to which these different approaches are effective or could be modified to increase their effectiveness has received relatively little attention, however.

The ways in which activities are performed provide for certain learning conditions, which can contribute to vocabulary learning (Webb & Nation, 2017). Encountering words repeatedly is essential for learning vocabulary (e.g., Brown, Waring, & Donkaewbua, 2008; Chen & Truscott, 2010; Horst, Cobb, & Meara, 1998; Pellicer-Sánchez & Schmitt, 2010; Pigada & Schmitt, 2006; Saragi, Nation, & Meister, 1978; Teng, 2016; Waring & Takaki, 2003; Webb, 2007; Zahar, Cobb, & Spada, 2001). Also, testing words that are learned (i.e., retrieval practice) is beneficial to vocabulary learning and retention (e.g., Barcroft, 2007, 2015; Nakata, 2017; Royer, 1973; van den Broek, Takashima, Segers, & Verhoeven, 2018).

Many studies have revealed that retrieving stored information can be a more potent learning opportunity than restudying the information (i.e., repeated study) (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006, 2011). Bjork (2011) pointed out that the process of retrieving does not merely test the information stored in memory; it also modifies the representation of the information in memory, which enables the information to become more recallable in the future. Furthermore, the power of repeated retrievals may depend on how the retrieval practice is scheduled. Spacing—providing an interval between learning opportunities (Anderson, 2000, p.235)—the repeated retrieval practice for a given item has become one of the mainstays of learning and memory research (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a review). A range of evidence in both cognitive psychology (e.g., Bahrick, 1979, Experiment 2; Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bahrick & Phelps, 1987; Bloom & Shuell, 1981; Cepeda, Coburn, Rohrer, Wixted, Mozer, & Pashler, 2009; Pyc & Rawson, 2009) and L2 vocabulary acquisition (e.g., Çekiç & Bakla, 2019; Küpper-Tetzel, Erdfelder, & Dickhäuser, 2014; Lotfolahi & Salehi, 2016; Nakata & Suzuki, 2019; Rogers & Cheung, 2020) supports the idea that spaced retrieval practice improves the learning and retention of vocabulary.

The current research aims to examine the effects of spaced practice. The remainder of this chapter will discuss theories relevant to spaced practice effects. It will then present earlier meta-analytic reviews of spaced practice effects and a brief review of literature investigating effects of spaced practice on L2 learning, followed by the rationale for the current research. After this chapter, three studies will be introduced: Chapter 2 (Study 1: a meta-analytic review of spaced practice effects), Chapter 3 (Study 2: an empirical study comparing the effects of spaced practice in L2 vocabulary learning through fill-in-the-blanks and flashcard activities), and Chapter 4 (Study 3: an empirical study comparing the effects of spaced practice in L2 vocabulary learning through sentence production and flashcard activities). Chapter 5 concludes the thesis with a brief summary of the findings from the three studies, implications for L2 vocabulary teaching and learning, and suggestions for future research.

1.1 Theories of Spaced Learning

Three interrelated ideas comprise theoretical explanations of the effects of spaced learning: first, spacing makes learning more difficult, but desirably so; second, forgetting that occurs via spacing strengthens remembering; and third, spacing enhances relearning.

1.1.1 Desirable Difficulties

Spacing learners' study sessions further apart makes learning more difficult and impedes their performance during learning (Bjork & Bjork, 2011). Bjork (1994) proposed spacing as one of the most effective manipulations to introduce desirable difficulties. In the theoretical idea of desirable difficulties, the word desirable is the key. Desirability of difficulties in spaced practice makes encoding of to-be-learned material richer during learning and requires learners engage in more effort for successful retrieval, which can trigger both encoding and retrieval processes that support learning, comprehension, and remembering (Bjork & Kroll, 2015; McDaniel & Butler, 2011). Although the conditions that are desirably difficult reduce the rate of apparent learning (Bjork & Kroll, 2015), they optimize long-term retention and slow forgetting, specifically in L2 vocabulary learning (Bahrick, 1979; Schneider, Healy, & Bourne, 2002).

1.1.2 Forgetting

Spacing may allow time for learners to forget previously learned information. However, Bjork and Bjork (1992) argued that previously learned information remains in memory; it does not decay but does become inaccessible. From this perspective, forgetting—losing access to information in memory—occurs because the retrieval of previously obtained information is inhibited by competition from other recently learned information associated with the same retrieval cue in memory (Bjork, 2011). However, it has also been assumed that forgetting often creates conditions suitable for effective learning (Bjork, 2011; Jacoby, 1978).

Bjork (2011) mentioned that since inaccessible (or competing) information also remains in memory, it can be recognized when encountered again and can subsequently be relearned at an accelerated rate. For example, we may not remember the address of our childhood home. However, we may be able to retrieve the address from our memories when we visit familiar places and see the street name. Such retrieved information can be recalled more easily in the near future because it was accessed in the recent past (Bjork, 2011). Jacoby (1978) noted that if students are allowed to forget previously learned information by inserting an interval of space between interventions (or by the use of an interceding exercise), they actually try to recall forgotten information to solve problems instead of merely remembering previously tendered responses. Therefore, the interruption ultimately leads to better performance in later instances of recall.

1.1.3 Relearning Effect

The relearning condition occurs after spacing. When the information to be remembered is repeated in the relearning condition, learners can strengthen the knowledge in their minds (Bjork, 2011). Bjork and Allen (1970) outlined two accounts of the relearning effect in spaced practice. The first idea is the consolidation of the first presentation (of a given item) and asserts that the first presentation is more effective in spaced practice than it is in repeated practice. In other words, the relearning effect occurs when two succeeding presentations are not close. Landauer (1969) assumed that the consolidation induced by repetition is less cumulative if two presentations occur too closely together. Conversely, the consolidation of relearning is likely to be more effective if a recurring presentation is spaced (Landauer, 1969). The second idea attributes the relearning effect to a second presentation's encoding variability, namely, spacing can make relearning (through the second presentation) more independent from the learning that occurred during the first presentation, i.e., the second presentation can constitute an entirely new encoding if the spacing between sessions is longer. This concept implies that temporal variations between learning sessions can reduce context dependency and can therefore help learners better encode the second presentation (Bjork & Allen, 1970).

To summarize, spacing makes learning difficult but ultimately also leads to better retrieval success. Spaced learning contributes to forgetting, but the process of learning and forgetting enhances remembering. Further, relearning the forgotten information after spacing can represent another learning step, and the acts of forgetting, remembering, and relearning result in greater long-term retention. The following section will present earlier reviews and empirical investigations of spaced practice effects.

1.2 Reviews of Research Investigating Spaced Practice Effects

Earlier reviews have suggested that spaced practice benefited verbal learning and memory (e.g., Cepeda et al., 2006; Donovan & Radosevich, 1999). Donovan and Radosevich (1999) meta-analyzed 63 studies (112 effect sizes) of spaced practice effects on learning and memory and found a medium-to-large effect of spaced practice (mean weighted effect size, d = 0.46, 95% CI [0.42, 0.50]) in comparison to massed practice. Cepeda et al. (2006) meta-analyzed 317 experiments from 184 studies of spaced practice effects on verbal learning and found that spaced practice contributed to better learning and retention than massed practice.

Given abundant evidence of spaced practice benefits in cognitive psychology research, there has been a great deal of research investigating the effects of spaced practice on L2 learning. Several studies have shown that spaced practice promoted better L2 learning and more enhanced long-term retention than massed practice, in which repetitions occur in immediate succession without the allocation of intervening time (e.g., Bloom & Shuell, 1981; Nakata & Suzuki, 2019; Suzuki, Yokosawa, & Aline, 2020). Other studies have revealed that spaced practice was as effective as massed practice on immediate learning of L2 vocabulary (e.g., Lee & Choe, 2014). Many studies have also examined the relative effects of different types of spaced practice (i.e., different length of spacing between learning opportunities, e.g., shorter spacing versus longer spacing) on L2 vocabulary learning. Several studies have indicated that longer spacing was more beneficial than shorter spacing (e.g., Bahrick, 1979, Experiment 2; Bahrick et al., 1993; Bahrick & Phelps, 1987; Pashler, Zarow, & Triplett, 2003; Pyc & Rawson, 2009, Experiment 1; Rogers, 2015). In contrast, other studies have

demonstrated that shorter spacing was more effective than longer spacing (e.g., Cepeda et al., 2009, Experiment 1; Küpper-Tetzel et al., 2014; Rogers & Cheung, 2020). Some studies, however, have indicated that shorter spacing was as effective as longer spacing on retention (e.g., Kasprowicz, Marsden, & Sephton, 2019).

Taken together, the findings of the earlier reviews and empirical studies have showed positive effects of spaced practice on learning and memory. However, Donovan and Radosevich (1999) and Cepeda et al. (2006) included very limited L2 studies (10% of the sample out of 112 effect sizes examined verbal learning with face-name pairs, L1-L1 word pairs, and L2-L1 word pairs, Donovan & Radosevich, 1999; 4% of the studies out of 184 research reports, Cepeda et al., 2006), and earlier studies of L2 learning have shown inconsistent results. It would be useful to clarify the overall effects of spaced practice on L2 learning, which would provide more accurate and meaningful evidence that spaced practice promotes L2 learning and enhances its retention.

1.3 Motivation for the Current Research

Given the limited number of L2 studies included in previous meta-analytic reviews of spaced practice (Cepeda et al., 2006; Donovan & Radosevich, 1999) and the inconsistent results obtained from earlier L2 studies on spaced practice, it is, therefore, important to clarify the overall effects of spaced practice in order to provide pedagogical guidance and useful directions for further research.

Numerous L2 studies of spaced practice have investigated L2 vocabulary learning and demonstrated benefits of spaced practice (e.g., Bahrick, 1979; Bloom & Shuell, 1981; Koval, 2020; Lotfolahi & Salehi, 2016; Nakata & Suzuki, 2019). However, most research on L2 vocabulary has attended to the paired-associate learning condition (e.g., flashcards). Flashcards is a common and efficient activity (Webb et al., 2020), but learning from flashcards is only one of many activities that are undertaken to deliberately learn words (e.g., matching words to their meanings, writing target words in given sentences, choosing the

correct meanings of target words, and producing original sentences using target words). Since there are so many different activities to learn words (Morgan & Rinvolucri, 2004; Webb & Nation, 2017), the effects of spacing cannot yet be generalized to other L2 vocabulary learning conditions. Furthermore, a lack of research beyond flashcards might have constrained the degree to which spaced practice effects are meaningful. It therefore would be pedagogically valuable to determine the extent to which spaced practice may contribute to vocabulary learning in different activities, because it may help teachers and learners to increase vocabulary learning gains.

The current research focuses on the effects of spaced practice. In the first study of three studies in this thesis, earlier L2 studies were systematically reviewed to clarify the overall effects of spaced practice. Systematic research synthesis of spaced practice effects on L2 learning may allow clear and meaningful synthetic conclusions to be drawn from a single category of studies. For example, regarding learning target (vocabulary, grammar, and pronunciation), it may be useful to understand the extent to which spaced practice affects vocabulary learning in relation to studies of L2 grammar and pronunciation learning.

Next, to examine whether other vocabulary learning activities are affected by spacing, the second study in this thesis examined whether spacing has the same effects on different L2 vocabulary learning activities. The effect of spaced practice on vocabulary learning through fill-in-the-blanks was compared to its effect with flashcards. The third study in this thesis examined the effects of spaced practice on vocabulary learning comparing sentence production to flashcards. Comparing the gains in vocabulary learning through other learning activities and flashcards may provide some indication of the degree to which spacing effects found through paired-associate learning conditions (e.g., flashcards) may be generalized to different vocabulary learning activities. The findings from the last two studies may be pedagogically valuable in further developing L2 vocabulary teaching and learning strategies by determining whether spaced practice with different activities promotes learning and encourages retention.

1.4 Thesis Format

The thesis involves three studies in the integrated article thesis format. Study 1 (Chapter 2) is a meta-analysis on the effects of spaced practice on L2 learning. In this article, 37 L2 studies (98 effect sizes from 48 experiments) of spaced practice were reviewed. The article has already been published by the top-tier international peer-reviewed journal *Language Learning* (Wiley) and is also available online through the following link: https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12479 (The Effects of Spaced Practice on Second Language Learning: A Meta-Analysis). Study 2 (Chapter 3) examined whether spacing has a similar effect on L2 vocabulary learning and retention in fill-in-the-blank and flashcard activities. In this article, different learning conditions were designed, based on activities (fill-in-the-blanks versus flashcards) and spacing schedules (massed [no interval] versus spaced [1-day interval]). Learning and testing correspondence effects (whether matching learning condition to test format affects learning) and the effects of feedback timing (whether feedback is provided immediately or with a delay) on vocabulary learning are also addressed. This article is currently under review at the top-tier international peer-reviewed journal *The Modern Language Journal* (Wiley). Study 3 (Chapter 4) examined the effects of spaced practice on the learning and retention of L2 vocabulary through sentence production and flashcards activities. In this article, different spacing schedules (massed versus spaced [1-day interval]) and activities (sentence production versus flashcards) were variables to compare the effects of spacing in different learning conditions. Learning and testing correspondence effects and the effects of feedback timing on vocabulary learning are also addressed in this chapter. This article is currently under review in the top-tier international peer-reviewed journal *TESOL Quarterly* (Wiley). In the final chapter (Chapter 5), the findings in all three studies are discussed and followed by the conclusion.

1.5 References

Anderson, J. R. (2000). *Learning and memory: An integrated approach* (2nd ed.). New York: Wiley.

Baddeley, A. (1999). *Human memory: Theory and practice* (Revised ed.). UK: Psychology Press.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*(3), 296−308. http://doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4*(5), 316−321. http://doi.org/10.1111/j.1467-9280.1993.tb00571.x

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13*(2), 344−349.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*(1), 35−56.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185−205). Cambridge, MA: The MIT Press.

Bjork, R. A. (2011). On the symbiosis of learning, remembering, and forgetting. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 1−22). New York: Psychology Press.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding. *Journal of Verbal Learning and Verbal Behavior, 9*(5), 567−572. http://doi.org/10.1016/S0022-5371(70)80103-7

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35−67). Hillsdale, NJ: Erlbaum.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56−64). New York: Worth Publishers.

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *American Journal of Psychology*, *128*(2), 241−252. http://doi.org/10.5406/amerjpsyc.128.2.0241

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research 74*(4), 245−248.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*(2), 136−163.

Çekiç, A., & Bakla, A. (2019). The effects of spacing patterns on incidental L2 vocabulary learning through reading with electronic glosses. *Instructional Science*, https://doi.org/10.1007/s11251-019-09483-4

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236−246. http://doi.org/10.1027/1618-3169.56.4.236

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

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*(5), 693−713. http://doi.org/10.1093/applin/amq031

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. https://doi.org/10.1037/0021-9010.84.5.795

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

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*, 273−293.

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.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal or Verbal Learning and Verbal Behavior, 17*(6), 649−667. http://doi.org/10.1016/S0022-5371(78)90393-6

Javanbakht, Z. O. (2011). The impact of tasks on male Iranian elementary EFL learners' incidental vocabulary learning. *Language Education in Asia, 2*(1), 28−42. http://doi.org/10.5746/LEiA/11/V2/I1/A03/Javanbakht

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966−968.

Koval, N. G. (2020). *Testing the reminding account of the lag effect in L2 vocabulary acquisition from L2-L1 retrieval practice within a paired-associate learning format* (Published doctoral dissertation). Michigan State University, The United States.

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.

Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review, 76*(1), 82−96. http://doi.org/10.1037/h0026746

Lee, E., & Choe, M. H. (2014). The effect of spaced repetitions on Korean elementary students' L2 English vocabulary learning. *Studies in English Education, 19*(1), 55−75.

Lotfolahi, A. R., & Salehi, H. (2016). Learners' perceptions of the effectiveness of spaced learning schedule in L2 vocabulary learning. *SAGE Open, 6*(2), 1−9. http://doi.org/10.1177/2158244016646148

McDaniel, M. A., & Bulter, A. C. (2011). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 175−198). New York: Psychology Press.

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive and receptive + productive L2 vocabulary learning: What difference does it make? In B. Laufer (Ed.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp.79−100). Amsterdam, the Netherlands: Benjamins.

Morgan, J., & Rinvolucri, M. (2004). *Vocabulary*. Oxford: Oxford University Press.

Nakata, T. (2008). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning, 24*(1), 17−38. http://doi.org/10.1080/09588221.2010.520675

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition, 39*(4), 653−679.

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. http://doi.org/10.1017/S0272263118000219

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1051−1057. http://doi.org/10.1037/0278-7393.29.6.1051

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language, 22*(1), 31−55.

Pigada, M., & Schmitt, N., (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language, 18*(1), 1−28.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437−447.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181−210. http://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 23−47). New York: Psychology Press.

Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*(4), 857−866. http://doi.org/10.1002/tesq.252

Rogers, J., & Cheung, A. (2020). Input spacing and the learning of L2 vocabulary in a classroom context. Language Teaching Research, 24, 616–641. http://doi.org/10.1177/1362168818805251

Rott, S. (2012). The effect of task-induced involvement on L2 vocabulary acquisition: An approximate replication of Hulstijn and Laufer (2001). In G. Porte (Ed.), *Replication Research in Applied Linguistics* (pp. 228–267). New York: Cambridge University Press.

Royer, J. M. (1973). Memory effects for test like events during acquisition of foreign language vocabulary. *Psychological Reports*, *32*, 195−198.

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72−78. http://doi.org/10.1016/0346-251X(78)90027-1

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*(2), 419−440. http://doi.org/10.1006/jmla.2001.2813

Siyanova-Chanturia, A., & Webb, S. (2016). Teaching vocabulary in the EFL context. In H. P. Widodo & W. A. Renandya (Eds.), *English language teaching today: Building a closer link between theory and practice* (pp. 227−239). Switzerland: Springer International Publishing.

Suzuki, Y., Yokosawa, S., & Aline, D. (2020). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research*. http://doi.org/10.1177/1362168820913985

Teng, F. (2016). The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading. *The Language Learning Journal.* http://doi.org/10.1080/09571736/2016.1244217

van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, *68*(2), 546−585.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130−163.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*(1), 33−52.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46−65. http://doi.org/10.1093/applin/aml048

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. RELC Journal, 30(3), 360−376.

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

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal, 104*(4), 715−738. http://doi.org/10.1111/modl.1267

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review, 57*(4), 541−572. http://doi.org/10.3138/cmlr.57.4.541

Chapter 2: The Effects of Spaced Practice on Second Language Learning: A Meta-Analysis

2.1 Introduction

Massed practice involves studying the same items in succession without any intervening time or items, whereas *spaced practice* involves studying items separated by an interval of time or other items. For example, massed practice in second language (L2) learning could involve learning *cat, dog,* and *fish* in the sequence *cat, cat, cat, dog, dog, dog, fish, fish, fish,* whereas spaced practice could involve learning the same items in a sequence such as *cat, dog, fish, cat, dog, fish, cat, dog, fish.* Research reveals that the inclusion of spacing promotes learning (e.g., Bahrick, 1979). The term *spacing effect* refers to enhanced learning, for a given item, during spaced practice as compared with massed practice.

There are, however, different types of spacing. *Absolute spacing* is the total amount of intervals between all learning opportunities for a given item (Karpicke & Bauernschmidt, 2011). For example, if an item is encountered six times with an encounter occurring every 3 minutes, the absolute spacing is 18 minutes. The distribution of learning opportunities relative to one another, including equal and expanding spacing, is captured by *relative spacing* (Karpicke & Bauernschmidt, 2011). *Equal spacing*, also known as fixed or uniform spacing, expresses the condition where the spacing between encounters for a given item is constant. In *expanding spacing*, the interval between encounters gradually increases. *Lag effects* refer to comparisons of the effects of different amounts of spacing (e.g., relatively short vs. relatively long).

Blocking ensures that the amount of practice devoted to a particular skill (or concept) is massed, and *interleaving* guarantees that practice of the particular skill (or concept) is spaced across multiple learning opportunities and separated by intervening tasks. In interleaved practice, for example, under the category of English tense (as a superordinate concept), learners learn different types of tense an equal number of times but in a different

order (e.g., present, past, future, past, present, future, present, past, future). In blocking practice, learners learn one type of tense, followed by another type (e.g., present, present, present, past, past, past, future, future, future). Although interleaving and spacing are separate constructs (i.e., interleaving operates at a superordinate level; Metcalfe, 2011), they are often confounded, and interleaving effects may reflect the contribution of spacing (Taylor & Rohrer, 2010).

Learning new skills or knowledge typically requires practice, and learning is enhanced when practice is spaced rather than massed (Baddeley, 1999; DeKeyser, 2007). Consequently, the development of spaced practice has become one of the most powerful advancements in learning and memory research. Numerous empirical studies (e.g., Carpenter & DeLosh, 2005) and reviews of literature (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Donovan & Radosevich, 1999) have demonstrated the benefits of spaced practice for skill learning (e.g., music performance, airplane control simulation) and for verbal memory learning tasks such as picture naming, fact recall, and paired-associate learning (e.g., first language [L1] word pairs, L2−L1 word pairs). Although previous reviews included L2−L1 word pairs as a verbal memory task, L2 learning studies were limited in number (no more than seven studies, e.g., Cepeda et al., 2006) or not clearly mentioned (Donovan & Radosevich, 1999), and thus the effects of spaced practice on L2 learning are less clear.

There has been a great deal of research investigating the effects of spaced practice on L2 learning, but the effects reported have been inconsistent. Research has revealed that (a) spaced practice benefited learning and retention of L2 vocabulary (e.g., Bloom & Shuell, 1981) and L2 grammar (e.g., Suzuki, Yokosawa, & Aline, 2020); (b) spaced practice was as effective as massed practice on immediate posttests (Lee & Choe, 2014); (c) longer spacing was superior to shorter spacing on delayed posttests measuring L2 vocabulary (e.g., Pashler, Zarow, & Triplett, 2003) and L2 grammar learning (Rogers, 2015); (d) shorter spacing contributed to greater learning than longer spacing on delayed posttests (e.g., Küpper-Tetzel, Erdfelder, & Dickhäuser, 2014); (e) shorter spacing was as effective as longer spacing on delayed posttests (e.g., Kasprowicz, Marsden, & Sephton, 2019); (f) equal spacing was more effective than expanding spacing on delayed posttests (e.g., Çekiç & Bakla, 2019); and (g)

equal spacing was as effective as expanding spacing on delayed posttests (e.g., Kang, Lindsey, Mozer, & Pashler, 2014).

Given the limited number of L2 studies included in previous reviews and the inconsistent results obtained from L2 studies on spaced practice, research aimed at clarifying findings is warranted. Compared to skill learning and verbal memory, there are arguably even more individual differences (e.g., language aptitude, Kasprowicz et al., 2019) and contextual variables (e.g., teaching techniques, Rogers & Cheung, 2020a; multiple modes of L2 input, Serrano & Huang, 2018; type of knowledge to be learned, Suzuki & DeKeyser, 2017a; task complexity, Suzuki et al., 2020) involved in L2 learning. Furthermore, there is abundant evidence of various instructional treatment benefits (e.g., form-focused instruction, implicit inductive teaching) in L2 learning (e.g., Norris & Ortega, 2000). It is, therefore, important to clarify the overall effects of spacing and the different types of spacing on L2 learning in order to provide pedagogical guidance, as well as to identify useful directions for future research. In addition, because learner-related variables (e.g., prior L2 knowledge, Nakata & Suzuki, 2019b) and methodological features (e.g., feedback, Nakata, 2015a) were noted as reasons for the inconsistency in findings, it is important to explore whether and to what extent the effect of spaced practice is moderated by different variables across studies. The present study aims to address these questions by conducting a meta-analysis, one of the most effective tools for comprehensive research synthesis (Hunter & Schmidt, 2004).

2.2 Background

2.2.1 Theories of Spaced Practice Effects

Many theories of spaced practice effects have been proposed and examined. First, spacing between learning opportunities makes learning more difficult, but desirably so (desirable difficulty framework, e.g., Bjork, 1994; Suzuki, Nakata, & DeKeyser, 2019). Second, forgetting occurring via spacing creates more effortful retrieval attempts, which strengthens retention (Bjork, 1975). Third, spacing between learning opportunities enhances subsequent

repeated learning (consolidation, e.g., Wickelgren, 1972). Fourth, spacing between learning opportunities results in more attentional processing, but massed learning results in less processing (deficient processing, e.g., Jacoby, 1978; Koval, 2019). Fifth, reducing the accessibility of information in memory after spacing enhances additional learning of that information (accessibility principle, e.g., Bjork & Bjork, 1992). Sixth, spacing makes subsequent repeated learning more distinctive, and the learning in different contexts is better remembered (contextual variability theory, e.g., Melton, 1970). Seventh, spacing manipulated between retrievals (i.e., testing information from memory) produces benefits on long-term retention (study-phase retrieval, e.g., Toppino & Bloom, 2002).

2.2.2 Previous Meta-Analytic Reviews of Spaced Practice

Donovan and Radosevich (1999) examined a total of 63 studies with 112 effect sizes and found that spaced practice was superior to massed practice. They reported that about 10% of the sample examined verbal memory with tasks (e.g., face–name pairs, low associate pairs; in which all the written and oral tasks were presented in the L1) and with L2−L1 word pairs. However, the number of L2 studies was unclear. Cepeda et al. (2006) meta-analyzed the effect of spaced practice in verbal recall tasks for memory (e.g., picture naming, spelling, low associates; in which all the materials were presented in the L1) and for L2 learning (e.g., learning the meanings of L2 words from paired associates), but only about 4% of the studies out of 184 research reports involved L2 learning. They found that spaced conditions were significantly better than massed conditions. They also found that longer spacing was more effective than shorter spacing at longer retention intervals (the interval between the last learning session and the final posttest). However, they found no obvious difference between equal and expanding spacing. Although Cepeda et al. reviewed the effects of spaced practice, there is as yet no clear description of the extent to which spaced practice affects L2 learning. This is because they mainly investigated the relationship between spacing intervals (the interval between learning opportunities) and retention intervals, and there were few L2 studies examined. Uchihara, Webb, and Yanagisawa's (2019) meta-analysis included spacing as a moderator variable and found that frequency effects in L2 incidental vocabulary learning

(whereby the higher the number of encounters with a word, the better the learning) were larger when words were encountered in massed conditions (defined as within one session), r = .38, 95% CI [.31, .45], than when words were encountered in spaced conditions (defined as learning across multiple sessions), r = .23, 95% CI [.12, .34]. However, spacing was not examined as the sole construct, so a clear picture of spacing effects on L2 vocabulary learning was not obtained.

2.2.3 Review of Moderator Variables on Spacing Effects

2.2.3.1 Age

Several L1 studies have examined the effects of spaced practice at different ages but have obtained inconsistent results: Older children showed spacing effects, but not younger children (e.g., Toppino & DiGeorge, 1984); young adults showed larger spacing effects than older adults (Maddox, Balota, Coane, & Duchek, 2011); there was no age difference between the effects of shorter and longer spacing (e.g., Seabrook, Brown, & Solity, 2005) or between the effects of equal and expanding spacing (e.g., Maddox et al., 2011). Furthermore, some findings conflict with Wilson's (1976) hypothesis that the effects of different types of spacing are dependent on working memory capacity (the ability to not only temporarily store information but also manipulate it for learning, Baddeley, Eysenck, & Anderson, 2015), which develops with age (Gathercole, Pickering, Ambridge, & Wearing, 2004). In L2 studies, spaced practice effects have been observed with adult learners (e.g., Li & DeKeyser, 2019) and with young learners (e.g., Lotfolahi & Salehi, 2017). However, given that no studies have examined age as an independent variable, the effects of spaced practice with L2 learners of different ages remain unclear. Furthermore, given that working memory capacity is significantly positively correlated with L2 learning (e.g., Linck, Osthus, Koeth, & Bunting, 2014), the effects of spaced practice may not be the same among L2 learners of different ages.

2.2.3.2 Learning Target

Most L2 spaced practice studies have investigated L2 vocabulary learning (e.g., Koval, 2020). Positive effects have also been demonstrated with L2 grammar or morphology (e.g., Suzuki et al., 2020) and L2 pronunciation (e.g., Carpenter & Mueller, 2013). However, acquisition of vocabulary and grammar may occur through different processes (Pinker, 1998). For example, Ullman (2015) reported that declarative memory may play different roles in lexical and grammatical aspects of learning and processing. Pronunciation learning is a different skill from vocabulary and grammar learning (Li & DeKeyser, 2019). Therefore, the effects of spaced practice may not be the same among different domains (vocabulary, grammar, and pronunciation) of a L2.

2.2.3.3 Number of Sessions

Spaced practice studies involve spacing within a single session or between multiple sessions. Most single-session studies manipulate item spacing (i.e., studying items separated by an interval of other items), and most multiple-session studies manipulate time spacing (i.e., studying items separated by an interval of time). It is also possible for multiple-session studies to manipulate item spacing (i.e., manipulating item spacing within each session). Spaced practice benefits have been observed when manipulated within a single session (e.g., Nakata & Suzuki, 2019b) as well as between multiple sessions (e.g., Li & DeKeyser, 2019). However, it is not clear whether the number of sessions affects outcomes. Therefore, it may be methodologically and pedagogically valuable to see whether it influences learning through spaced practice.

2.2.3.4 Type of Practice

Spaced practice can involve repeated practice in studying materials (study trials), retrieving information from memory (test trials), or a combination of studying and retrieval (test–

restudy or study–test trials; e.g., Roediger & Karpicke, 2006). Several studies have revealed long-term retention benefits of information relearned in spaced practice (e.g., Verkoeijen, Rikers, & Ö zsoy, 2008). Other studies found that repeatedly assessing information across time promotes learning (e.g., Lawrence, 2013). This suggests that both spaced restudy and retrieval practice are effective for learning and retention. However, studies comparing repeated restudy practice (study trials) to repeated retrieval followed by feedback across time (test–restudy trials) found that the best retention occurred in the test–restudy trials (e.g., Butler & Roediger, 2007). L2 studies have found positive effects of retrieval relative to restudy on L2 vocabulary learning and retention (e.g., Barcroft, 2007). None of these studies, however, involved spacing as an independent variable. Furthermore, there has been no empirical research comparing restudy to retrieval on L2 grammar or pronunciation.

2.2.3.5 Activity Type

Research (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013) has found the benefits of spaced practice to be general across a range of materials, such as verbal materials (e.g., word pairs, facts), visual materials (e.g., pictures, videos), and educational materials (e.g., lectures, mathematical formulas). However, not all tasks yield large benefits of spaced practice. Donovan and Radosevich (1999) found that there was a large spacing effect with a low level of task complexity, d = 0.97, 95% CI [0.88, 1.06], but a small effect with a high level of task complexity, d = 0.07, 95% CI [−0.05, 0.18]. Spaced practice for L2 learning has also been studied with a wide range of activities: paired-associate tasks (e.g., Nakata, 2015a), listening and reading activities for form–meaning mapping (e.g., Kasprowicz et al., 2019), judgment tasks (e.g., Li & DeKeyser, 2019), oral description using pictures (e.g., Suzuki et al., 2020), and exercises such as multiple-choice tasks, fill-in-the-blanks tasks (e.g., Bloom & Shuell, 1981), and crossword puzzles (e.g., Rogers & Cheung, 2020b). These activities are used to help L2 learners to comprehend target items (e.g., multiple-choice tasks, reading texts, listening and identifying the correct spoken forms of words) and to produce target items (e.g., picture description, making sentences, pronouncing words). Donovan and Radosevich (1999) coded foreign language tasks (L2–L1 word pairs) as representing an average level of

task complexity and found a small-to-medium effect of spacing, d = 0.42, 95% CI [0.36, 0.48]. However, there might be a difference in the level of difficulty that learners experience in comprehending versus producing target items, and hence this may impact the magnitude of spacing effects.

2.2.3.6 Provision of Feedback

Studies have demonstrated that spacing effects may be influenced by the provision of feedback after retrieval (e.g., Roediger & Karpicke, 2006). Cepeda et al. (2006) reported that feedback may be a variable that explains differences between equal and expanding spacing; when feedback is provided, expanding spacing benefits performance because feedback minimizes the chance of forgetting an item (Pashler, Cepeda, Wixted, & Rohrer, 2005). However, Cepeda et al. (2006) could not examine the effect of feedback because all three studies included in their meta-analysis for equal and expanding spacing provided feedback. It would be useful to examine the effects of feedback because spaced practice studies that have provided feedback have reported contrasting results. For example, Kang et al. (2014) failed to find a positive effect for expanding spacing with feedback relative to equal spacing with feedback, whereas Nakata (2015a) found expanding spacing with feedback to be superior to equal spacing with feedback. However, it should be noted that Nakata found a significant effect of expanding spacing only on a posttest involving receptive recall (from L2 to L1), with very small effect sizes, d = 0.12−0.19, 95% CI [−0.80, 0.53]. Furthermore, given that feedback to correct learners' responses has generally been found to be beneficial to L2 learning (e.g., Li, 2010), it would be interesting to see whether the effect of spaced practice is moderated by feedback.

2.2.3.7 Feedback Timing

The timing of feedback may also moderate learning through spaced practice. Some studies in cognitive psychology found that delayed feedback (e.g., feedback given after all responses)

had a greater effect on learning than immediate feedback (e.g., Butler, Karpicke, & Roediger, 2007), but others found more benefit from immediate feedback (e.g., Brosvic, Epstein, Cook, & Dihoff, 2005). The superiority of delayed feedback can be explained by the fact that delayed feedback results in more laborious learning circumstances, which fits with the desirable difficulty framework (e.g., Bjork, 1994; Suzuki et al., 2019). In contrast, because immediate feedback is generally provided after each response, it is more likely to make learners fully process feedback after both incorrect and correct responses (Butler & Roediger, 2007).

In L2 studies, Nakata (2015b) examined feedback timing (immediate and delayed) in four different repeated retrieval practice conditions (one, three, five, or seven retrievals). Sixteen English–Japanese word pairs were divided into two sets of eight items. One set was assigned to the immediate feedback condition, in which feedback was provided immediately after each response. The other set was assigned to the delayed feedback condition, in which feedback was provided after all eight items were performed. The interval between the last encounter with a given item and the posttest was controlled. Nakata found no main effect of feedback timing for L2 vocabulary learning on either receptive (from L2 to L1) or productive (from L1 to L2) recall posttests. On the 1-week delayed posttest, he found a significant effect of the immediate feedback on only the receptive recall posttest, with a very small effect size, $d = 0.14$, 95% CI [0.03, 0.51]. However, because this study did not manipulate the spaced learning conditions, the effect of feedback timing on spaced practice for L2 vocabulary learning and retention remains unclear. Furthermore, there has been no empirical research on L2 grammar or pronunciation learning that has directly investigated the interaction between spacing and feedback timing. Given that the impact of feedback on learning and memory has been endorsed by the majority of investigations, it is useful to examine whether immediate or delayed feedback is more conducive to L2 learning in more versus less spaced conditions.

2.2.3.8 Frequency of Practice

Spaced practice studies have included different numbers of encounters with target items, ranging from one or two (e.g., Pyc & Rawson, 2009) to 27 or 30 (e.g., Suzuki, 2017). Greater frequency of practice can provide learners with more time to restudy or more attempts to retrieve. Maddox and Balota (2015) found, in a L1 study using low associate word pairs (e.g., apple–evil), significant increases in retrieval practice performance as the number of tests during the training sessions increased from one to five in a shorter spacing condition, whereas in a longer spacing condition retrieval practice performance increased from the one-test to the three-test condition, but did not increase further in the five-test condition. These findings may suggest that providing more practice does not always lead to better performance or better retention. Nakata (2017) looked at the role of retrieval frequency (one, three, five, or seven retrievals) within a single session for L2 vocabulary learning. He found that five or seven retrievals led to better performance than one or three retrievals on both immediate and delayed posttests. To our knowledge, there is no L2 empirical research investigating the relationship between spaced conditions and frequency of practice.

2.2.3.9 Retention Interval

Spaced practice effects may depend on when knowledge is measured (Cepeda et al., 2006; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Rohrer & Pashler, 2007). Cepeda et al. (2006) found a positive relationship between spacing intervals and retention intervals (RIs); the longer the spacing, the greater the retention. Rohrer and Pashler (2007) reported that spacing effects depended jointly on spacing intervals and RI, arguing that the learning outcomes of different types of spaced practice may be better or worse depending on when the final test is taken. Cepeda et al. (2008) found that longer spacing produces better retention than shorter spacing at long RIs, whereas shorter spacing outperformed longer spacing at short RIs. These findings suggest that the length of RI may have a considerable impact on the effects of spaced practice.

2.3 Method

Research Questions

The current meta-analysis was guided by the following research questions:

1. To what extent does spacing affect L2 learning?

2. To what extent do learning gains differ in relation to type of spacing?

3. Which empirically motivated variables (age, learning target, number of sessions, type of practice, activity type, provision of feedback, feedback timing, frequency of practice, and RI) moderate the effects of spaced practice?

2.3.1 Literature Search

First, we comprehensively searched 22 relevant journals of cognitive psychology, applied psychology, applied linguistics, and second language acquisition for different combinations of key words: *spacing effect, massed, interleaving, blocking, lag effect, shorter spacing, longer spacing, absolute spacing, relative spacing, equal spacing, fixed spacing, uniform spacing, expanding spacing, second language learning*, and *foreign language learning.* We then employed the following electronic databases in order to extend the search: Education Resources Information Center, Linguistics and Language Behavior Abstracts, PsycINFO, and Google Scholar. In addition, we searched references in review articles (e.g., Cepeda et al., 2006) and in book chapters (e.g., Carpenter, 2017). We set 1979 as the starting point because Bahrick's study from that year is one of the classic experiments on spaced practice (as observed by Dunlosky et al., 2013), and because there were very few L2 empirical studies prior to 1979 (cf. Crothers & Suppes, 1967, Experiments 8, 9, 10, and 11), and those that existed did not report sufficient statistical information to calculate effect sizes. We set July 2020 as the completion point for our data collection.

In order to minimize the "file-drawer" problem in research synthesis (the fact that some studies remain in researchers' files because of the publication bias toward studies reporting significant findings; Rosenthal, 1979), we considered retrieving "fugitive" literature (e.g., unpublished papers, doctoral theses, conference presentations). However, due to the difficulty involved in retrieving those sources, we decided to include only doctoral theses that are carefully designed and provide detailed statistical information. We used the electronic database ProQuest Global Dissertations and Theses to search for doctoral theses, employing the same key words as for published studies.

2.3.2 Inclusion Criteria

All reports that appeared initially eligible for the meta-analysis were then examined in reference to a set of inclusion criteria. To be included in the meta-analysis, a study report had to meet all of the following criteria:

1. The study had to examine the effect of spaced practice on L2 learning. We took L2 learning to include learning of L2 vocabulary such as single words or collocations (Snoder, 2017), L2 grammatical structures (e.g., past perfect tense; Bird, 2010), L2 morphological features (e.g., Japanese te-form of the verb; Suzuki & DeKeyser, 2017a), L2 pronunciation (e.g., Mandarin monosyllables such as ba with different tones; Li & DeKeyser, 2019), and orthographic and phonological nonsense items (e.g., Nakata & Elgort, 2021).

2. The study had to feature a comparison of one type of practice with another type of practice in order to examine the effects of spaced practice (i.e., comparing spaced with massed practice, longer with shorter spacing, or equal with expanding spacing). For example, Uchihara et al. (2019) meta-analysis included massed and spaced conditions as a moderator to examine frequency effects in L2 incidental vocabulary learning. However, the studies included in their meta-analysis were not included in the current meta-analysis because none of them qualified as a comparative study examining the effects of spaced practice.

3. Studies comparing blocking to interleaving were included (Carpenter & Mueller, 2013; Nakata & Suzuki, 2019b; Pan, Tajran, Lovelett, Osuna, & Rickard, 2019; Suzuki et al., 2020). Blocking corresponds to massed practice or shorter spacing (not pure massed practice), whereas interleaving is equivalent to spaced practice or longer spacing (see Appendix S2 in the Supporting Information online for the category criteria).

4. The study had to provide clear spacing intervals. For example, we excluded the study by Lightbown and Spada (1994) that compared 18 hours per week to 2 hours per week because it was not clear whether the time distribution was either shorter or longer, or equal or expanding. Additionally, we excluded studies involving spaced practice with different criterion levels via a dropout method (where items that were correctly retrieved during a trial were removed from the to-be-practiced list in the subsequent trial), because the number of test–restudy trials per item was variable between participants (e.g., five-drop group, Pyc & Rawson, 2007).

5. The study had to control for participants' preexisting knowledge of target items (vocabulary, grammatical features, and pronunciation rules). Conducting a pretest to show no statistically significant difference between groups on the pretest (e.g., Suzuki et al., 2020), using nonsense items (e.g., Nakata & Elgort, 2021) or a miniature artificial language (e.g., Suzuki, 2017), and pilot testing of target items with another population (e.g., Nakata & Elgort, 2021) are common ways of controlling for prior knowledge.

6. Studies adopting both intentional and incidental learning conditions for the target L2 items were included. In the former, target items are explicitly taught or studied (e.g., Bird, 2010). In the latter, the target items are not explicitly taught or studied, and participants are not told about subsequent posttests (e.g., Serrano & Huang, 2018).

7. The study had to provide enough statistical information for effect size calculation. Several studies (e.g., Bahrick, 1979) did not provide enough information to calculate effect sizes. We contacted authors and were grateful to receive additional information that allowed us to complete the current meta-analysis (our thanks to Emilie Gerbier, Jeffrey Karpicke, Sean Kang, Steve Pan, and Thomas Toppino).

8. When the study included more than one experiment with different participants, each experiment contributed an effect size in the meta-analysis (e.g., Pan et al., 2019).

9. Replicated or extended studies had to involve different data samples. For example, Suzuki (2017) reported the same data as Suzuki (2018, 2019); Suzuki and DeKeyser (2017a) reported the same data as Suzuki and DeKeyser (2017b). In this meta-analysis, we included the study by Suzuki (2017), which was replicated, and the study by Suzuki and DeKeyser (2017a), which examined the effects of spaced practice as the main focus, whereas we excluded the authors' subsequent studies (Suzuki, 2018, 2019; Suzuki & DeKeyser, 2017b), which reanalyzed the same data using cognitive aptitude (e.g., working memory) from the perspective of aptitude– treatment interaction.

10. The study was written in English.

11. Studies adopting both within-participants and between-participants designs were included. In a within-participants design, the independent variable (spacing) is manipulated within participants. For example, half of the items might be studied in a massed condition whereas the other half are studied in a spaced condition. In a between-participants design, spacing is manipulated between participants. For instance, half of the participants study the items with a massed condition and the other half study them with a spaced condition.

The PRISMA flow diagram presented in Figure 1 depicts the study inclusion criteria (Moher, Liberati, Tetzlaff, & Altman, 2009) and provides the number of included and excluded references. More detailed information is reported in Appendix S1 in the online Supporting Information for this article. Forty-eight experiments reported in 37 studies (N = 3,411) were selected for this meta-analysis. The 48 experiments were then divided into three different categories of spaced schedules: (a) spaced versus massed, (b) longer versus shorter, and (3) equal versus expanding comparisons (see Appendix S2 in the Supporting Information online for the category criteria). Each category was meta-analyzed separately.

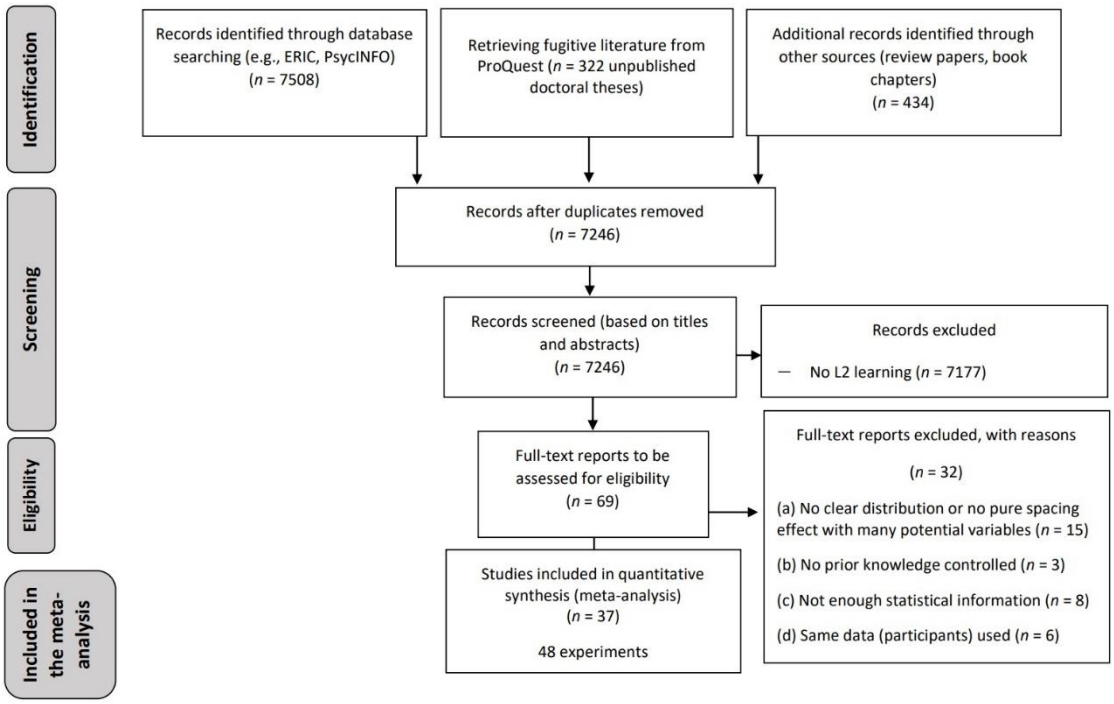


Figure 1 PRISMA flow diagram.

2.3.3 Coding: Dependent and Moderator Variables

The dependent variables were the effect sizes derived from the included L2 studies. The effect sizes were classified as either immediate effects (from immediate posttest scores) or delayed effects (from delayed posttest scores). Some previous studies on spaced practice involved filler tasks (e.g., 5-minute U.S. state naming, Carpenter & Mueller, 2013; 73-second 10 two-digit additions, Nakata, 2015a), followed by immediate posttests, and other studies involved delayed posttests 1 day after treatments (e.g., Pashler et al., 2003). In the current meta-analysis, a test was defined as an immediate posttest if it was taken on the same day as the treatment (i.e., immediately after the last training session or after a filler task administered at the end of the last training session); a test was defined as a delayed posttest if it was taken after a delay of 1 day or greater following the treatment.

Following Suzuki (2017), when a study administered two or more delayed posttests, only the last posttest's score was included and coded as the dependent variable. For example, when a study administered two delayed posttests (e.g., 7-day and 35-day delayed posttests), the first delayed posttest was regarded as a learning session, and the second (last) posttest's score was included and coded as the dependent variable (for delayed effect). When the posttest was administered at three different RIs (e.g., 1-day, 4-week, and 8-week delayed posttests; Schuetze, 2014), the first and second delayed posttests were regarded as learning sessions, and the last delayed posttest's score was coded as the dependent variable. Note that this was the case only if the RI was manipulated within participants.

When there were multiple types of posttests, a shifting unit of analysis was adopted (Patall, Cooper, & Robinson, 2008). For example, if a study involved two different types of immediate and/or delayed posttests (e.g., matching and grammaticality judgment tests, Kasprowicz et al., 2019; error correction and translation tests, Miles, 2014), two separate effect sizes were calculated. For estimating the overall effect of choice, we averaged these two effect sizes so that the sample contributed only one effect size. However, we did not include reaction time data, because such data were provided in only a few of the studies included in the meta-analysis (k = 4: Li & DeKeyser, 2019; Nakata & Elgort, 2021; Suzuki, 2017; Suzuki & DeKeyser, 2017a), and they are based on different metrics (e.g., speed rate or word processing). For example, Suzuki's (2017) study measured accuracy (from vocabulary and grammar tests) and speed (from reaction time), and we included only data from the accuracy measure. Another reason to not include reaction time data was that Avery and Marsden (2019) found that effect sizes from reaction time data are quite a lot lower than the field averages, and they speculated that this could be because the standard deviations are normally wider than for other metrics.

In the current meta-analysis, therefore, each of the three categories (spaced vs. massed, longer vs. shorter, and equal vs. expanding) includes two different timings of the outcome measures: immediate and delayed effects in the spaced versus massed category, immediate and delayed effects in the longer versus shorter category, and immediate and delayed effects in the equal versus expanding category.

We included a total of nine moderator variables: one learner-related variable (age) and eight methodology-related variables (learning target, number of sessions, type of practice, activity type, provision of feedback, feedback timing, frequency of practice, and RI) (see Appendix S3 in the Supporting Information online for the coding scheme). The coding sheet with the data (Kim & Webb, 2021) is publicly available at http://www.iris-database.org.

2.3.3.1 Age

Because a limited number of studies reported the age of participants (21 of 37 studies, 57%), age was initially categorized according to grade levels (e.g., Grade 3, Rogers & Cheung, 2020a). However, because some studies involved participants with a wide range of grade levels (e.g., Grades 9−12, Bloom & Shuell, 1981; Grades 3−8, Lotfolahi & Salehi, 2016) or involved adults ranging from 20 to 63 years (Kang et al., 2014), which makes it difficult to determine the differential effects of spaced practice on learners of different grade levels, this variable was coded as young learners (Grades 1−12) and adult learners (university students or older).

2.3.3.2 Learning Target

This variable consists of three types of L2 items: vocabulary (both single words and multiword items), grammar (including morphological structure), and pronunciation (a monosyllabic item with different tones or pronunciation rules).

2.3.3.3 Number of Sessions

This variable was coded as single session and multiple sessions. Note that the number of sessions includes only training sessions and does not include testing (immediate or delayed posttest) sessions. For example, if a study used time spacing (e.g., a 10- minute interval

between trials) within one training session, followed by testing sessions (e.g., one immediate and two delayed posttests), the study is coded as single session.

2.3.3.4 Type of Practice

Practice includes two types of conditions: study trial and test trial. A study trial refers to an opportunity to restudy the target items that participants learned or studied. A test trial refers to an opportunity to recall or retrieve the target items that participants learned or studied. Note that feedback provided after a test trial can also be an opportunity to restudy the target items that participants learned in the initial learning session. Type of practice was coded as being one of five types: test–restudy (all) trial (testing, followed by restudying all target items); test–restudy (not recalled) trial (testing, followed by restudying only the items that were not recalled); study trial; test trial; and study–test trial (for details, see Tables S4.2 and S4.3, Appendix S4, in the Supporting Information online).

2.3.3.5 Activity Type

This variable was coded as one of: paired associate; comprehension activities; production activities; and combined activities that involved both comprehension and production activities. Paired-associate learning included learning from word lists or word cards. As in the descriptions of activities reported in the meta-analysis by Shintani, Li, and Ellis (2013), L2 activities other than paired-associate learning were coded as comprehension or production activities. Additionally, activities that involved both comprehension (e.g., multiple-choice tasks) and production (e.g., fill-in-the-blanks tasks) were coded as combined activities. Note that although a paired-associate task can involve either receptive retrieval (comprehending the L1 meaning of a L2 word) or productive retrieval (producing the L2 word corresponding to a L1 word given), we consider paired-associate tasks as a separate type of activity, distinct from comprehension, production, and combined activities (for details, see Tables S4.4 and S4.5, Appendix S4, in the Supporting Information online).

2.3.3.6 Provision of Feedback and Feedback Timing

Provision of feedback was coded according to the absence or presence of feedback. The presence of feedback was further categorized into two subgroups according to feedback timing (whether feedback was provided immediately after each response or with a delay).

2.3.3.7 Frequency of Practice

Frequency of practice was reported as the amount of repeated practice (excluding the initial presentation to learn target items). Thus, this is different from the total number of sessions, which includes the presentation, practice, and posttest sessions used in the treatment. For example, Nakata and Suzuki (2019a) included two sessions: The first session consisted of the pretest, learning session (presentation followed by three test trials), and immediate posttest, whereas the second session involved a delayed posttest. Frequency of practice in this study is 3 and the total number of sessions is 2. Following Suzuki (2017), when a study administered two posttests (immediate and delayed) and RI was manipulated within participants, the immediate posttest can be regarded as a learning session. When a study administered three posttests (immediate and two delayed), the immediate posttest and the first delayed posttest are regarded as learning sessions. Thus, in Nakata and Suzuki's (2019a) study, whereas frequency of practice was 3 at the time point for immediate posttest, frequency of practice was 4 at the time point for delayed posttest.[1]

2.3.3.8 Retention Interval

RI was coded as the number of days between the last learning session and the final posttest. In the current meta-analysis, six studies administered multiple delayed posttests (Bird, 2010; Li & DeKeyser, 2019; Lotfolahi & Salehi, 2016; Schuetze, 2014; Suzuki, 2017; Suzuki & DeKeyser, 2017a). Suzuki (2017) pointed out that the first delayed posttest could influence

the retention of knowledge measured by the second delayed posttest. Hence, the first delayed posttest was considered another retrieval practice in Suzuki's (2017) study. Following Suzuki (2017), if a study involved 7-day and 35-day delayed posttests, the calculated RI is 28 days (RI of the last delayed posttest − RI of the delayed posttest administered before the last delayed posttest; 35 days − 7 days = 28 days).[2] It should be noted that this was the case only if the RI was manipulated within participants.[3]

2.3.4 Reliability of the Coding

To assess the reliability of our coding procedures, 12 studies (approximately 32% of 37 studies) were randomly selected and independently coded by a second rater. The second rater is an expert in the area of spaced practice research with a doctoral thesis examining the effects of spaced practice on L2 vocabulary learning. The number of discrepancies between the two raters was calculated by performing Cohen's kappa test (a statistic for either interrater or intrarater reliability testing). The overall agreement was rated at 99% (almost perfect agreement; Cohen, 1960). After all disagreements were resolved through discussion, the first author coded the remaining studies (see Appendix S5 in the Supporting Information online for coding reliability, including Cohen's kappa [k] for each variable that was coded).

2.3.5 Data Analysis

We used Comprehensive Meta-Analysis (version 3.3) software (Borenstein, Hedges, Higgins, & Rothstein, 2013) to calculate the overall effect sizes and conduct analyses for nine moderator variables. In order to address the first research question, we aggregated effect sizes from the studies included in the spaced versus massed comparison to produce a weighted mean effect size. For the second research question, we aggregated effect sizes from the studies included in the longer versus shorter and equal versus expanding categories. To aggregate effect sizes, we used a random-effects model (using the unrestricted maximum likelihood method) so that variation in intervention effects across studies was accommodated

(Borenstein, Hedges, Higgins, & Rothstein, 2009). A significant between-group Q value indicates a heterogeneous distribution with a common effect size among identified samples and thus facilitates subsequent moderator analyses. However, a nonsignificant Q value is not always taken as assurance that the effects are consistent, because the Q statistic and its p value only address the variability of the null hypothesis (Borenstein et al., 2009). In the current meta-analysis, therefore, we also report I 2 statistics (the proportion of variation in effect sizes across studies), tau (the standard deviation of true effects), and prediction interval (how widely the effect sizes vary across studies), which are intended to quantify heterogeneity (the distribution of effects; Borenstein, Higgins, Hedges, & Rothstein, 2017). For the last research question, we conducted moderator analyses in all of the three categories (spaced vs. massed; longer vs. shorter; and equal vs. expanding). A random-effects meta-regression (using the unrestricted maximum likelihood method) was performed for continuous variables (frequency of practice and RI). The statistical significance is assessed if the p value of the data analysis is less than the prespecified alpha of 0.05.

2.3.5.1 Effect Size Calculation

To calculate the effect size of each study, the standardized mean difference from a study that used two independent groups was estimated and converted to Hedges's *g* by multiplying a correction factor: $J = 1 - (3/[4 \times df - 1])$. The overall effect size was calculated by weighing the average effect size for each study according to sample size and then pooling the effect sizes across studies.

Because the current study examines the effectiveness of spaced practice (spaced vs. massed, longer vs. shorter, and equal vs. expanding), comparative effect sizes were computed. A comparative effect size represents the effect of treated groups in comparison with baseline groups (Shintani et al., 2013). In the spaced versus massed comparison, for example, a significant effect size ($g = 0.50$) in the positive direction implies that spaced practice (the treated condition) is more effective than massed practice (the baseline condition) by 0.5 standard deviation units. In contrast, a significant effect size in the negative direction

($g = -0.50$) suggests that massed practice (the baseline condition) is more effective than spaced practice (the treated condition) by 0.5 standard deviation units. In the longer versus shorter comparison, longer and shorter spacing data were coded as treated and baseline data, respectively. In the equal versus expanding comparison, equal and expanding spacing data were coded as treated and baseline data, respectively.

From 48 experiments, we identified 26 effect sizes in the spaced versus massed comparison, including 11 with immediate posttests and 15 with delayed posttests. In the longer versus shorter spacing comparison, we identified 49 effect sizes, including 17 with immediate posttests and 32 with delayed posttests. Finally, in the equal versus expanding comparison, we identified 23 effect sizes, including 7 with immediate posttests and 16 with delayed posttests.

The detection of outliers was performed to ensure the robustness of the results, because the presence of studies with extreme effect sizes may have an impact on the results. Following previous meta-analyses (e.g., that by Shintani et al., 2013), the effect sizes contributed by the included studies were transformed into z scores, and any value (regardless of whether it was positive or negative) larger than 2.0 was removed from the analysis. Outlier detection was performed repeatedly until there were no further outliers. One outlier was identified from the z-score examination (Lotfolahi & Salehi, 2017: z = 2.152).

Finally, we assessed publication bias in the current data sets. Because most studies in this meta-analysis were published (35, alongside one contribution to conference proceedings, Khoii & Abed, 2017, and one doctoral thesis, Koval, 2020), our meta-analysis is more likely to include statistically significant findings than statistically nonsignificant findings (Lipsey & Wilson, 2001); therefore, a bias might influence the results of our meta-analysis. Results demonstrated that publication bias is considered to be a potential threat to conclusions drawn about the effects of spaced practice. The true magnitudes of effects of spaced practice on L2 learning might be smaller than those reported in this meta-analysis, though it is not known how much smaller and whether it would affect all three categories (spaced vs. massed, longer vs. shorter, and equal vs. expanding) of comparisons and all the moderator variables in the

same way (see Appendix S6 in the Supporting Information online for publication bias analyses).

2.4 Results

2.4.1 To What Extent Does Spacing Affect L2 Learning?

Results showed that spaced practice led to significant improvement in L2 learning and retention compared to massed practice (see Figures 2 and 3). Spaced practice was significantly more effective than massed practice on the immediate posttests, g = 0.58, 95% CI [0.16, 1.00], a medium effect according to Cohen's benchmarks (1988; g = 0.2 for small, 0.5 for medium, and 0.8 for large), and small-to-medium with reference to the benchmarks found by Plonsky and Oswald (2014; between-group contrast, g = 0.4 for small, 0.7 for medium, and 1.0 for large). For the domain of psychology (g = 0.32, median effect, Schäfer & Schwarz, 2019), however, the spacing effect of 0.58 from our meta-analysis could be considered large. A significant Q value (Q = 54.72, p < .001) indicates that the true effect size is not identical in all the studies. Of the variance in observed effects, 81.72% reflects variance in true effects rather than sampling error (I^2 = 81.72), and the standard deviation of true effects (tau) was 0.631. We predict that the true effects would fall in the range of −0.93 to 2.09, and it would make sense to apply moderator analyses or meta-regression to explain the variance (Borenstein et al., 2009).

A spacing effect was also found on the delayed posttests, g = 0.80, 95% CI [0.44, 1.17], and the confidence interval values (which do not pass zero) suggested that the size of the spacing effect in the long term could be considered medium to large (Plonsky & Oswald, 2014), and large with reference to Cohen (1988) and to Schäfer and Schwarz (2019). A significant Q test (Q = 79.83, p < .001) and high value of I^2 (82.46%) indicated that the observed variance would remain among identified samples. Tau was 0.639, and the prediction interval tells us that most effects would fall in the range of −0.64 to 2.24. This justified subsequent moderator analyses or meta-regression.

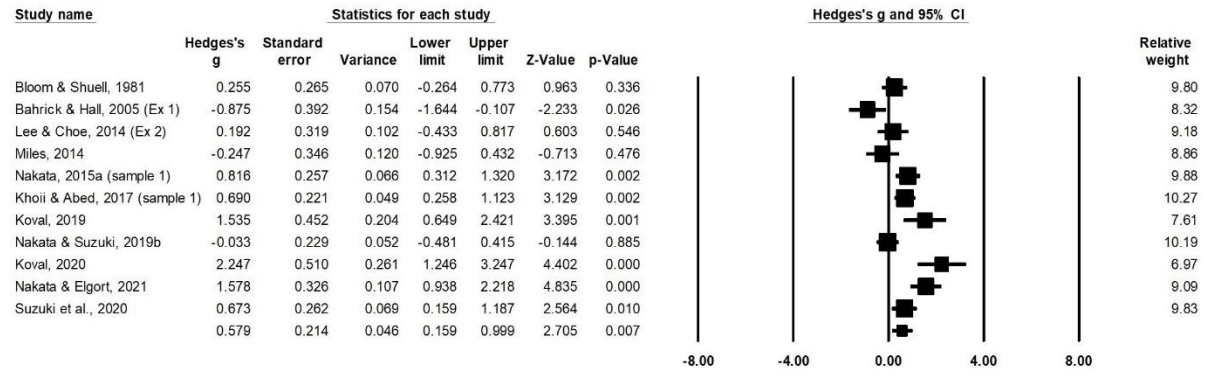


Figure 2 Overall average effect size (indicated by a diamond) of spaced practice when compared to massed practice, and effect sizes with 95% confidence intervals for each study (dependent variable = immediate posttest scores, $k = 11$). Effect sizes are calculated as Hedges's g

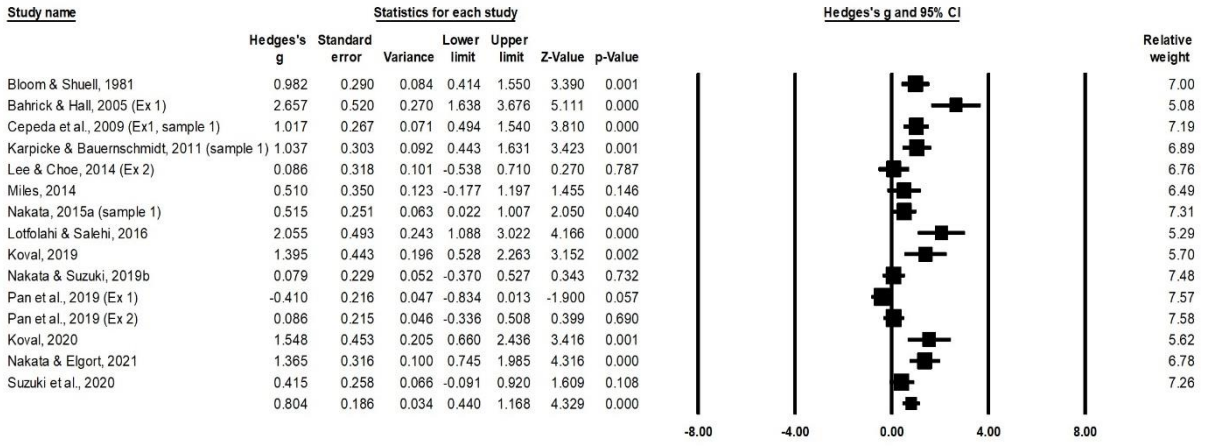


Figure 3 Overall average effect size (indicated by a diamond) of spaced practice when compared to massed practice, and effect sizes with 95% confidence intervals for each study (dependent variable = delayed posttest scores, $k = 15$). Effect sizes are calculated as Hedges's g.

2.4.2 To What Extent do Learning Gains Differ in Relation to Type of Spacing?

Results demonstrated that the effects of shorter and longer spacing were similar on the immediate posttests, $g = -0.15$, 95% CI [−0.37, 0.06]; the confidence intervals crossed zero (see Figure 4), and tau was 0.332. The prediction interval was −0.90 to 0.60, and we predict that the true effects would fall in this wide range. A significant Q value ($Q = 37.07$, $p < .001$) and an I^2 value of 56.84% justified subsequent moderator analyses or meta-regression. However, longer spacing showed a greater effect than shorter spacing on the delayed posttests, $g = 0.40$, 95% CI [0.16, 0.64] (see Figure 5). The confidence interval values, with the lower bound only just above zero, suggested that the size of longer spacing effects in the long term could be considered small (Plonsky & Oswald, 2014), or small to medium with reference to Cohen (1988), but in the medium range within the domain of psychology (Schäfer & Schwarz, 2019). Tau was 0.607, and the prediction interval was −0.87 to 1.67 for the delayed effects. We would predict that the true effect sizes would fall in this wide range. A significant Q value ($Q = 163.63$, $p < .001$) and I 2 value of 81.05% justified subsequent moderator analyses or meta-regression.

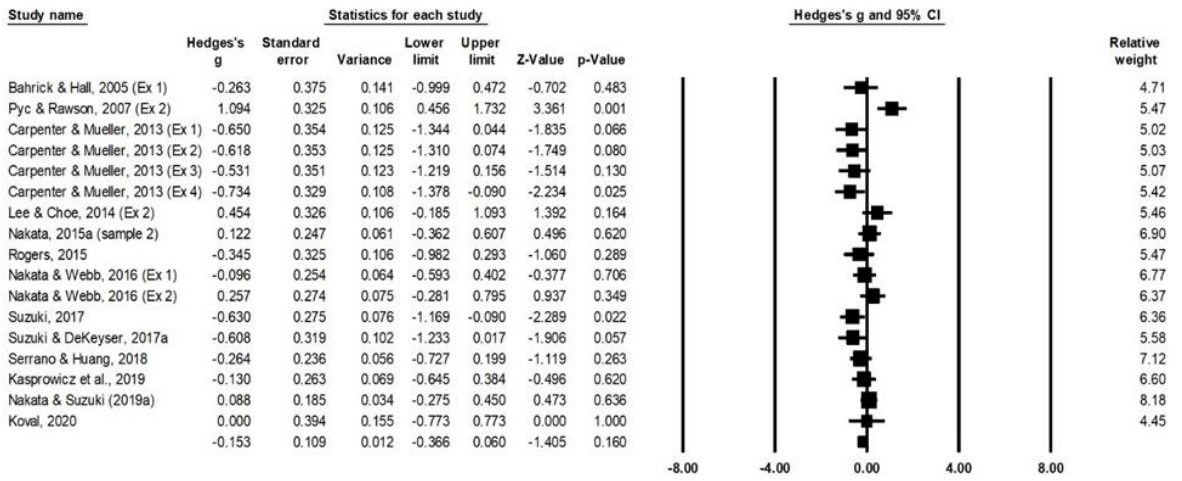


Figure 4 Overall average effect size of longer spaced practice (treated) when compared to shorter spaced practice (baseline), and effect sizes with 95% confidence intervals for each study (dependent variable = immediate posttest scores, $k = 17$). Effect sizes are calculated as Hedges's g.

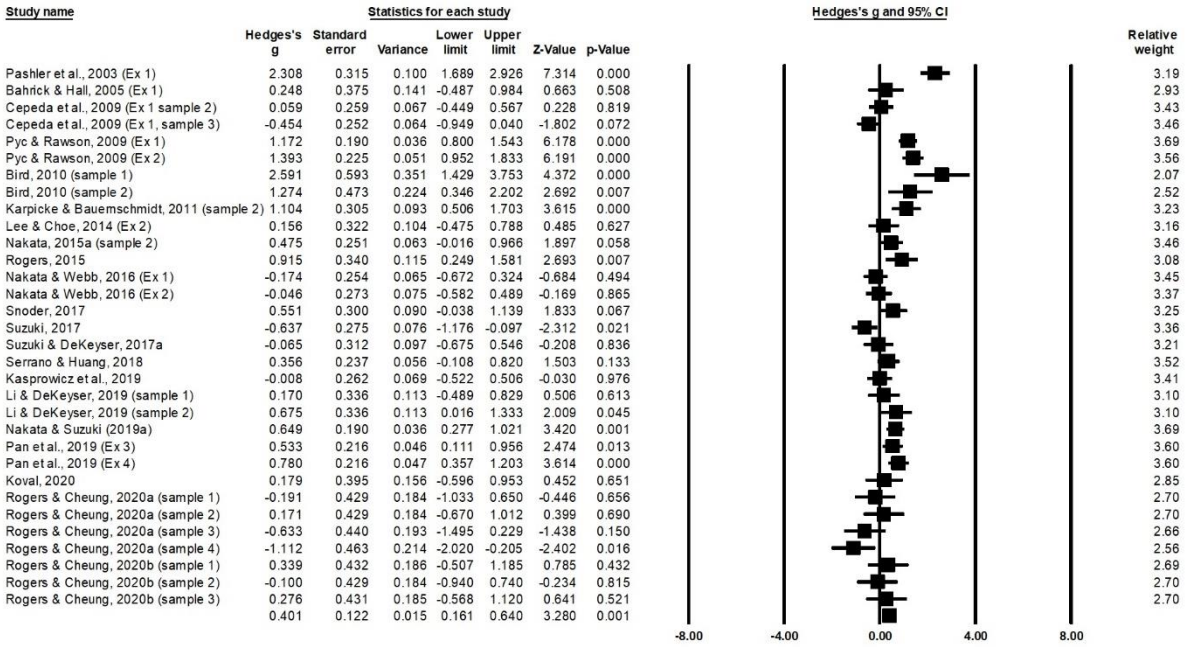


Figure 5 Overall average effect size of longer spaced practice (treated) when compared to shorter spaced practice (baseline), and effect sizes with 95% confidence intervals for each study (dependent variable = delayed posttest scores, *k* = 32). Effect sizes are calculated as Hedges's *g*.

Results showed that equal spacing was as effective as expanding spacing on both immediate posttests, $g = 0.15$, 95% CI [−0.07, 0.37], and delayed posttests, $g = -0.15$, 95% CI [−0.33, 0.03]; the confidence intervals crossed zero (see Figures 6 and 7). I 2 values in this comparison were zero on the immediate posttests and 27.19% on the delayed posttests; a value near zero suggested that almost no observed variance remained, thus no subsequent moderator analysis for the immediate effects is reported; and the value on the delayed posttests indicated that there was a small part ($I^2 = 27.19\%$) of an observed dispersion. Tau was 0.188, and the prediction interval was −0.60 to 0.30. Subsequent moderator analysis and meta-regression for the delayed effects in the equal versus expanding comparison was somewhat justified but should be cautiously interpreted when the results of the analyses suggest that moderator variables may explain the variance.

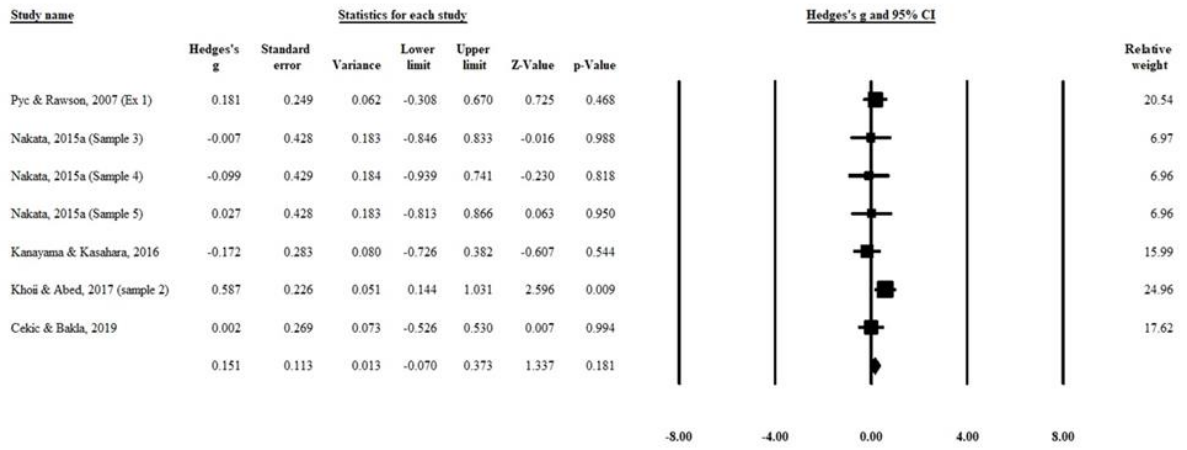


Figure 6 Overall average effect size of equal spaced practice (treated) when compared to expanding spaced practice (baseline), and effect sizes with 95% confidence intervals for each study (dependent variable = immediate posttest scores, $k = 7$). Effect sizes are calculated as Hedges's g.

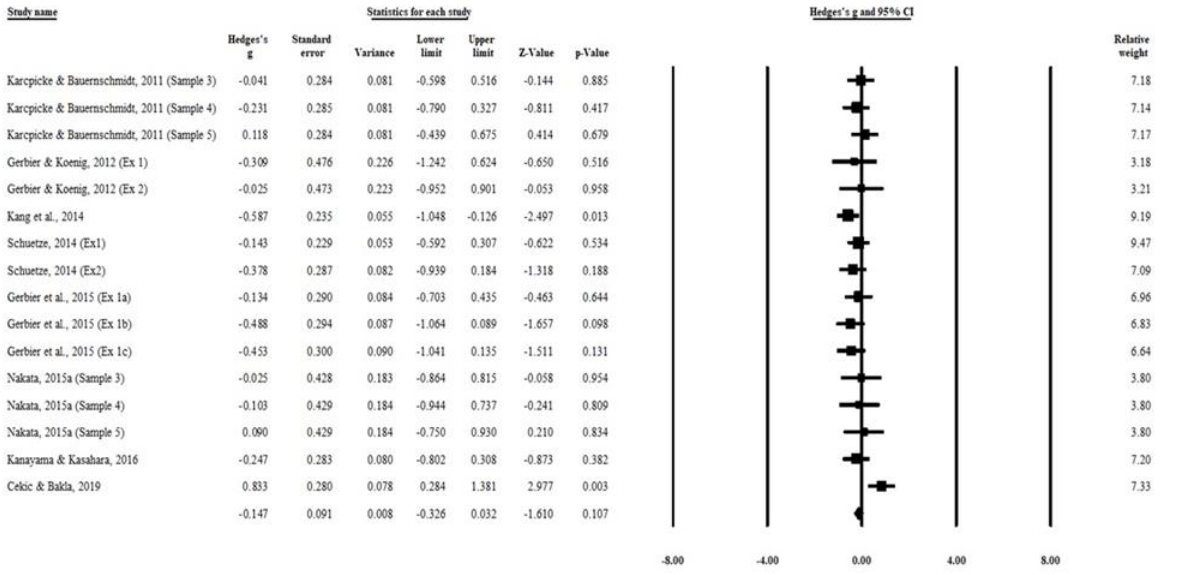


Figure 7 Overall average effect size of equal spaced practice (treated) when compared to expanding spaced practice (baseline), and effect sizes with 95% confidence intervals for each study (dependent variable = delayed posttest scores, $k = 16$). Effect sizes are calculated as Hedges's g.

It should be noted that publication bias analyses indicated that apparent bias exists in the subset of effects from delayed posttests from the spaced versus massed comparison. However, the results of p-uniform (see Appendix S6 in the Supporting Information online) showed that the bias is negligible. In the subset of effects from immediate posttests from the equal versus expanding comparison, I^2 and tau were zero, indicating that estimates of *p*-uniform should be examined. *P*-uniform enables testing of the extent of heterogeneity and considers the statistical significance of effect sizes (van Aert, Wicherts, & van Assen, 2016).

However, the results of both *p*-uniform and the random-effects model were similar (very small effects with confidence intervals that crossed zero), which led to the conclusion that random-effects meta-analysis results may be interpreted as the standard meta-analytic estimates. Because most studies included in the current meta-analysis were published studies (published studies = 35, contribution to conference proceedings = 1, and PhD thesis = 1), a symmetrical distribution may not rule out publication bias. Therefore, the overall effects of spaced practice on L2 learning from the current meta-analysis should be interpreted with caution.

2.4.3 Which Empirically Motivated Variables Moderate the Effects of Spacing?

The *Q* test indicates whether a variable is a significant predictor; that is, whether the effect sizes of baseline and treated conditions effect sizes for that variable are significantly different. However, because of small samples in the current meta-analysis, we interpret the results by focusing more on effect sizes and their confidence interval values. Recall that the moderator analyses are based on the comparative effect sizes; a positive effect size indicates a better effect for the treated group and a negative effect size shows a superior effect for the baseline group. No moderator analysis for the immediate effects in the equal versus expanding comparison ($I^2 = 0$, tau = 0) was reported. Separate meta-regression analyses (using the unrestricted maximum likelihood method) for two continuous variables (frequency of practice and RI) were performed to determine whether these variables were significant predictors of the effects of spaced practice on L2 learning. Moderator analyses for learner and

methodological variables are shown in Tables 1 and 2 (see Appendix S8 in the Supporting Information online for details).

2.4.3.1 Age

Spacing promoted better learning when it involved adult learners, g = 0.66, 95% CI [0.13, 1.20], than when it involved young learners, g = 0.39, 95% CI [−0.44, 1.22]. However, in the long term, the effects were larger with young learners, g = 0.97, 95% CI [0.11, 1.82], than with adult learners, g = 0.77, 95% CI [0.36, 1.18]. Longer spacing significantly led to better retention than shorter spacing when it involved adult learners, g = 0.54, 95% CI [0.27, 0.81]. However, longer spacing was as effective as shorter spacing when it involved young learners. Because the sample sizes for young learners were small (k = 3 in the spaced vs. massed comparison and k = 8 in the longer vs. shorter comparison), readers should be cautious in interpreting the benefits of spaced practice with young learners.

2.4.3.2 Learning Target

Spacing led to better learning and retention when it involved L2 vocabulary, g = 0.76−1.15, 95% CI [0.26, 1.49], than when it involved L2 grammar, g = 0.11−0.14, 95% CI [−0.64, 0.92]. However, the confidence intervals for L2 grammar learning crossed zero, suggesting that the spacing effects could be statistically unstable when learning involves L2 grammar. Shorter spacing was significantly more effective for the immediate learning of L2 pronunciation, g = −0.64, 95% CI [−1.06, −0.21] (not passing through zero), and of grammar, g = −0.41, 95% CI [−0.70, −0.13] (not passing through zero), but longer spacing significantly enhanced retention for L2 grammar and vocabulary; the effect was larger with grammar, g = 0.56, 95% CI [0.06, 1.06], than with vocabulary, g = 0.34, 95% CI [0.04, 0.64]. However, the benefit from longer spacing in the long term remains unclear when it targets pronunciation, because the sample size was small (k = 2 for delayed effects).

Table 1 Moderator analyses for categorical variables (immediate posttests)

	k	*g*	Variance	95% CI		*p*	*Q* tests	
				Lower	Upper		*Q*	*p*
Age								
Spaced vs. Massed							0.30	.58
Young	3	0.39	0.03	-0.44	1.22	.36		
Adult	8	0.66	0.10	0.13	1.20	.01*		
Longer vs. Shorter							0.45	.50
Young	3	-0.03	0.04	-0.42	0.37	.89		
Adult	14	-0.19	0.02	-0.44	0.06	.14		
Equal vs. Expanding							2.18	.14
Young	2	0.35	0.09	0.01	0.69	.05*		
Adult	5	0.01	0.02	-0.29	0.30	.96		
Learning target								
Spaced vs. Massed							1.71	.19
Vocabulary	8	0.76	0.08	0.26	1.25	.00*		
Grammar	3	0.14	0.08	-0.64	0.92	.72		
Longer vs. Shorter							15.59	.00*
Vocabulary	9	0.14	0.02	-0.11	0.38	.28		
Grammar	4	-0.41	0.02	-0.70	-0.13	.01*		
Pronunciation	4	-0.64	0.03	-0.98	-0.30	.00*		
Number of sessions								
Spaced vs. Massed							5.86	.02*
Single session	6	1.04	0.01	0.49	1.59	.00*		
Multiple sessions	5	0.04	0.06	-0.55	0.63	.88		
Longer vs. Shorter							0.78	.38

Single session	10	-0.08	0.03	-0.40	0.23	.60		
Multiple sessions	7	-0.27	0.02	-0.52	-0.01	.04*		
Equal vs. Expanding							0.25	.62
Single session	4	0.07	0.03	-0.29	0.44	.70		
Multiple sessions	3	0.19	0.06	-0.12	0.51	.23		
Type of practice								
Spaced vs. Massed							1.34	.72
Test-restudy (all)	6	0.69	0.13	0.05	1.34	.04*		
Test-restudy (no recalled)	2	0.48	0.05	-0.06	1.55	.39		
Study trial	2	0.81	0.45	-0.34	1.97	.17		
Longer vs. Shorter							11.74	.01*
Test-restudy (all)	6	0.22	0.02	-0.08	0.51	.16		
Test-restudy (no recalled)	3	-0.54	0.03	-0.89	-0.18	.00*		
Study trial	5	-0.41	0.05	-0.86	0.04	.07		
Study-test trial	3	-0.24	0.02	-0.54	0.07	.13		
Activity type								
Spaced vs. Massed							1.91	.59
Paired associate	3	0.67	0.60	-0.29	1.63	.17		
Comprehension activities	3	0.97	0.37	0.04	1.91	.04*		
Production activities	2	0.68	0.03	-0.42	1.78	.22		
Combined activities	3	0.07	0.03	-0.85	1.00	.88		
Longer vs. Shorter							13.75	.00*
Paired associate	7	0.17	0.02	-0.11	0.45	.24		
Comprehension activities	4	-0.38	0.03	-0.69	-0.07	.02*		
Production activities	3	-0.64	0.03	-0.99	-0.28	.00*		
Combined activities	3	-0.15	0.08	-0.71	0.41	.60		
Provision of feedback								

Spaced vs. Massed								1.32	.25
	Absence	2	0.02	0.06	-0.99	1.02	.98		
	Presence	7	0.69	0.09	0.15	1.23	.01*		

Note. *Statistically significant at $p < .05$.

Table 2 Moderator analyses for categorical variables (delayed posttests)

				95% CI		*p*	*Q* tests	
	k	*g*	Variance	Lower	Upper		*Q*	*p*
Age								
Spaced vs. Massed							0.16	.69
Young	3	0.97	0.25	0.11	1.82	.03*		
Adult	12	0.77	0.04	0.36	1.18	.00*		
Longer vs. Shorter							4.35	.04*
Young	8	-0.04	0.03	-0.52	0.44	.86		
Adult	24	0.54	0.02	0.27	0.81	.00*		
Learning target								
Spaced vs. Massed							13.78	.00*
Vocabulary	10	1.15	0.04	0.81	1.49	.00*		
Grammar	5	0.11	0.03	-0.32	0.54	.61		
Longer vs. Shorter							0.54	.76
Vocabulary	22	0.34	0.02	0.04	0.64	.03*		
Grammar	8	0.56	0.07	0.06	1.06	.03*		

Pronunciation	2	0.42	0.06	-0.57	1.42	.41		
Number of sessions								
Spaced vs. Massed							1.91	.17
Single session	9	0.61	0.05	0.16	1.05	.01*		
Multiple sessions	6	1.12	0.10	0.55	1.69	.00*		
Longer vs. Shorter							6.83	.01*
Single session	11	0.76	0.04	0.42	1.11	.00*		
Multiple sessions	21	0.18	0.02	-0.10	0.45	.21		
Equal vs. Expanding							0.68	.41
Single session	6	-0.04	0.02	-0.35	0.28	.81		
Multiple sessions	10	-0.20	0.02	-0.42	0.02	.08		
Type of practice								
Spaced vs. Massed							3.35	.34
Test-restudy (all)	10	0.70	0.05	0.25	1.14	.00*		
Test-restudy (no recalled)	2	1.73	0.65	-0.36	1.73	.20		
Study trial	2	0.69	0.43	-0.93	1.95	.49		
Longer vs. Shorter							15.86	.00*
Test-restudy (all)	16	0.38	0.02	0.10	0.67	.01*		
Test-restudy (no recalled)	6	1.06	0.09	0.61	1.50	.00*		
Study trial	6	-0.12	0.06	-0.62	0.38	.64		
Study-test trial	3	0.40	0.06	-0.23	1.03	.22		
Equal vs. Expanding							15.33	.00*
Test-restudy (all)	8	-0.32	0.01	-0.54	-0.10	.00*		
Test-restudy (no recalled)	3	-0.05	0.03	-0.37	0.27	.76		
Study trial	2	-0.17	0.11	-0.82	0.49	.62		
Test trial	2	-0.23	0.03	-0.59	0.12	.19		
Activity type								

Spaced vs. Massed							6.26	.10
Paired associate	6	1.36	0.08	0.80	1.92	.00*		
Comprehension activities	5	0.43	0.10	-0.15	1.00	.14		
Combined activities	3	0.53	0.07	-0.23	1.29	.52		
Longer vs. Shorter							10.72	.01*
Paired associate	12	0.58	0.05	0.23	0.93	.00*		
Comprehension activities	9	0.73	0.03	0.31	1.15	.00*		
Production activities	8	-0.24	0.03	-0.72	0.24	.32		
Combined activities	3	0.16	0.03	-0.55	0.86	.66		
Equal vs. Expanding							13.42	.00*
Paired associate	13	-0.23	0.01	-0.41	-0.06	.01*		
Production activities	2	-0.23	0.03	-0.59	0.12	.19		
<u>Provision of feedback</u>								
Spaced vs. Massed							0.00	.95
Absence	3	0.85	0.03	0.02	1.68	.05*		
Presence	10	0.82	0.06	0.36	1.27	.00*		
Longer vs. Shorter							0.71	.40
Absence	4	0.24	0.12	-0.41	0.89	.47		
Presence	23	0.55	0.02	0.27	0.82	.00*		
Equal vs. Expanding							0.01	.93
Absence	6	-0.16	0.01	-0.45	0.14	.31		
Presence	8	-0.14	0.03	-0.42	0.15	.36		
<u>Feedback timing</u>								
Spaced vs. Massed							10.40	.00*
Immediate feedback	8	0.52	0.05	0.10	0.94	.02*		
Delayed feedback	2	2.35	0.13	1.36	3.34	.00*		
Longer vs. Shorter							2.83	.09

Immediate feedback	15	0.39	0.03	0.08	0.71	.01*		
Delayed feedback	8	0.87	0.06	0.41	1.34	.00*		
Equal vs. Expanding							1.06	.30
Immediate feedback	5	0.04	0.09	-0.44	0.52	.88		
Delayed feedback	3	-0.36	0.03	-0.94	0.22	.23		

Note. * Statistically significant at $p < .05$.

2.4.3.3 Number of Sessions

We found a significantly large benefit of spacing on improving immediate L2 performance when it involved a single session, g = 1.04, 95% CI [0.49, 1.59]. However, better retention occurred when it involved multiple sessions, g = 1.12, 95% CI [0.55, 1.69], than when it involved a single session, g = 0.61, 95% CI [0.16, 1.05]. Longer spacing significantly promoted greater retention than shorter spacing when it involved a single session, g = 0.76, 95% CI [0.42, 1.11]. However, when it involved multiple sessions, longer spacing was as effective as shorter spacing. Small effects of expanding spacing for retention were found when it involved a single session, g = −0.04, 95% CI [−0.35, 0.28], and multiple sessions, g = −0.20, 95% CI [−0.42, 0.02], but the effects were not statistically reliable.

2.4.3.4 Type of Practice

Spaced practice promoted better learning and retention when it involved a test– restudy trial (g = 0.48−1.73, 95% CI [−0.06, 2.79], than when it involved a study-only trial, g = 0.69−0.81, 95% CI [−0.36, 1.97]. However, because the sample size for study-only trial was small (k = 2), the smaller effect with a study-only trial should be interpreted with caution. Longer spacing significantly led to greater retention than shorter spacing when it involved a test– restudy trial, g = 0.38−1.06, 95% CI [0.10, 1.50], but longer spacing was as effective as shorter spacing when it involved a study trial or study–test trial. Expanding spacing led to greater retention when it involved a test–restudy trial than when it involved a study trial or test trial. Although the confidence intervals for the test–restudy trial showed statistically reliable effects of expanding spacing, the findings from the equal versus expanding comparison should be interpreted with caution because of small samples (k = 2 for study trial, k = 2 for test trial).

2.4.3.5 Activity Type

Spacing promoted better learning on immediate posttests when it involved comprehension activities, g = 0.97, 95% CI [0.04, 1.91], than when it involved other activities, g = 0.07−0.68, 95% CI [−0.85, 1.78]. However, better retention occurred when it involved a paired-associate task, g = 1.36, 95% CI [0.80, 1.92], than when it involved other activities, g = 0.43−0.53, 95% CI [−0.23, 1.29]. Shorter spacing benefited immediate L2 performance when it involved production activities, g = −0.64, 95% CI [−0.99, −0.28], but longer spacing led to greater retention when it involved comprehension activities and paired associates than when it involved production or combined activities; the positive effect of longer compared to shorter spacing was larger with comprehension activities, g = 0.73, 95% CI [0.31, 1.15], than with paired associates, g = 0.58, 95% CI [0.23, 0.93]. Expanding spacing led to significantly better retention than equal spacing when it involved paired associates, g = −0.23, 95% CI [−0.41, −0.06]. Because the sample size for production activities was small (k = 2), the benefit of expanding spacing with production activities remains unclear.

2.4.3.6 Provision of Feedback

Spaced practice relative to massed practice improved immediate L2 performance more when feedback was provided, g = 0.69, 95% CI [0.15, 1.23], than when feedback was not provided, g = 0.02, 95% CI [−0.99, 1.02]. However, spacing enhanced retention regardless of whether feedback was provided or not. The effect when there was an absence of feedback should be interpreted with caution due to small samples (k = 2 at immediate posttests and k = 3 at delayed posttests in the spaced vs. massed comparison). Longer spacing produced better retention at delayed posttests when feedback was provided, g = 0.55, 95% CI [0.27, 0.82], than when feedback was not provided, g =0.24, 95% CI [−0.41, 0.89]. The confidence intervals (95% CI

[0.27, 0.82]) for the presence of feedback did not include zero, suggesting that larger spacing between feedback and the subsequent trial promotes better retention. Feedback did not have an impact on the comparative effectiveness of equal and expanding spacing.

2.4.3.7 Feedback Timing

Spacing led to greater retention when feedback was provided with a delay, g = 2.35, 95% CI [1.36, 3.34], than when feedback was immediately provided, g = 0.52, 95% CI [0.10, 0.94]. However, the extreme effect size should be interpreted with caution due to the small samples (k = 2 for delayed feedback). Longer spacing led to significantly better retention when delayed feedback was provided, g = 0.87, 95% CI [0.41, 1.34], than when immediate feedback was provided, g = 0.39, 95% CI [0.08, 0.71]. An extremely small to negligible effect in favor of equal spacing was found when immediate feedback was provided, g = 0.04, 95% CI [−0.44, 0.52], and a small effect was found in favor of expanding spacing when delayed feedback was provided, g = −0.36, 95% CI [−0.94, 0.22]. However, for both these effects the confidence intervals crossed zero indicating that these differential effects between equal and expanding spacing regarding feedback timing are unlikely to be statistically reliable.

2.4.3.8 Frequency of Practice

The random-effects meta-regression analyses showed a positive relationship between frequency of practice and the immediate effects (i.e., the greater the frequency of practice, the larger the spacing effects relative to massed practice on immediate learning), but a negative relationship with the delayed effects (i.e., the greater the frequency of practice, the smaller the spacing effects relative to massed practice in the long term). A negative relationship between frequency of practice and effect sizes was found in the longer versus shorter comparison (i.e., the greater the

frequency of practice, the larger the effects for shorter spacing). A negative relationship was also found in the equal versus expanding comparison (i.e., the greater the frequency of practice, the larger the expanding spacing effects). However, the effects of frequency of practice in the three comparisons (spaced vs. massed, longer vs. shorter, and equal vs. expanding) were small to negligible (not statistically significant).

2.4.3.9 Retention Interval

The random-effects meta-regression analyses showed a positive, albeit small and negligible (not statistically significant), relationship between RI and effect sizes in the spaced versus massed comparison (i.e., the longer the RI, the greater the spacing effects relative to massed practice). In the longer versus shorter comparison, the analyses indicated that the longer the RI, the greater the shorter spacing effects, however, the relationship was negligible (not statistically significant). In the equal versus expanding comparison, the results showed a significant negative relationship indicating that the longer the RI, the larger the effects of expanding spacing schedules.

2.5 Discussion

The analyses of comparative effects indicated that spaced practice was significantly more effective for L2 learning (g = 0.58) and retention (g = 0.80) than massed practice. It is notable that spaced practice can lead to better immediate gains than massed practice. The benefits of massed learning have been demonstrated at extremely short RIs (2 or 4 seconds, e.g., Peterson, Saltzman, Hillner, & Land, 1962). Our finding contrasts with results obtained by Peterson et al. (1962) and suggests that spaced practice is a more effective strategy than massed practice to enhance learners' L2 performance immediately. Our finding is consistent with previous meta-analyses (Cepeda et al., 2006; Donovan & Radosevich, 1999). Donovan and Radosevich (1999)

found a mean weighted effect size of 0.45, 95% CI [0.41, 0.50], for immediate learning and 0.51, 95% CI [0.39, 0.64], for retention, indicating that spaced practice was significantly more beneficial than massed practice for both immediate learning and retention. Cepeda et al. (2006) found positive effects of spaced practice at short RIs ranging from 1 second to less than 1 day (averaged percentage correct on the final test: 38.5% for massed practice, 47.6% for spaced practice) and at longer RIs ranging from 1 day to more than 31 days (28.5% for massed practice, 47.4% for spaced practice). It is also important to note that the effects of spacing are considered smaller than those of certain types of L2 instruction (e.g., form-focused or implicit instruction). Norris and Ortega (2000) meta-analyzed the effectiveness of L2 instruction (i.e., focus on form explicit and implicit treatments, and focus on forms explicit and implicit treatments) and found a large effect of all instructional treatments, d = 0.96, 95% CI [0.78, 1.14]. Although the benefits of spaced practice on L2 learning found in the current meta-analysis were smaller (g = 0.58 to 0.80) than the effects of other types of L2 instruction (e.g., focus on form explicit, focus on forms explicit) found by Norris and Ortega, spacing can still be considered to be useful for L2 learning.

The analyses indicated that both shorter and longer spacing have initial benefits, whereas longer spacing has a greater effect on durable learning. Cepeda et al. (2006) also found a pattern with the greatest increases in retention at longer spacing. Consistent with the desirable difficulty framework (e.g., Bjork, 1994), better retention occurred under difficult conditions, such as after longer spacing as opposed to shorter spacing. The overall magnitude of the longer spacing effect (g = 0.40) from our findings is small to medium, in spite of a number of previous memory studies (e.g., Cepeda et al., 2005) that have demonstrated the benefits of longer spacing in the long term. This might be because some inconsistency was shown regarding the effects of shorter and longer spacing on L2 learning, suggesting that other variables affecting the benefits of one type of practice over another could be observable in instructed L2 learning.

The analyses also revealed that there were no significant differences between equal and expanding spacing in either the immediate or the delayed posttests. It should be noted that only a small number of studies included immediate posttests (k = 7), and so we should be cautious about the differential effects of these two spacing types on short-term learning. It is important to note, however, that expanding spacing was as effective as equal spacing in the delayed posttests. This finding suggests that how soon learners retrieve items in the first (initial) retrieval practice or how soon subsequent practice occurs, may not have much impact on long-term retention.

We focused on variables that may moderate the effects of spacing on L2 learning. First, spaced practice promoted better learning and retention of L2 vocabulary and grammar for both young and adult learners. Specifically, adult learners showed greater retention with longer spacing than young learners. This supports Wilson's (1976) hypothesis that the effect of different types of spacing is dependent on working memory capacity; increasing the spacing between items may be more beneficial to older learners than younger learners. Because the sample sizes for young learners were small (k = 3 in the spaced vs. massed comparison and k = 8 in the longer vs. shorter comparison), there would be value in further exploring the effects of spaced practice with young learners.

Second, the effects of different types of spacing were evident in the learning of L2 grammar and pronunciation. Shorter spacing led to greater immediate learning of L2 grammar (g = −0.41) and pronunciation (g = −0.64) than longer spacing. This may be due to the complexity of the task or skill to be learned in grammar and pronunciation learning. It may be more difficult for learners to retrieve grammatical rules in oral production tasks than in comprehension and written tasks (Suzuki, 2017). Brief auditory input in pronunciation learning may be difficult for learners to access after spacing, especially when the spacing is longer (Baddeley, Thomson, & Buchanan, 1975). The benefits of blocking and interleaving may be more relevant for pronunciation and grammar learning than for vocabulary learning. Blocking can help learners identify the commonalities within each concept, whereas interleaving can help learners distinguish among different concepts (Taylor & Rohrer, 2010).

However, when target features (e.g., pronunciation rules) are easily distinguished from each other (e.g., *eau, s, ch*; Carpenter & Mueller, 2013), the benefits of interleaving can be reduced (less pronounced). Thus, shorter spacing (or blocking, with immediate repetition of items sharing the same pronunciation rules) may be particularly beneficial for helping learners to notice and understand the pronunciation rule patterns (Carpenter & Mueller, 2013). Saito and Plonsky (2019) found a medium effect of L2 pronunciation teaching on L2 pronunciation development, $d = 0.68$, 95% CI [0.49, 0.86], for between-group contrasts. Similarly, we found a medium effect of longer spacing for L2 pronunciation learning relative to shorter spacing ($g = -0.64$). However, given that our study sample size was small ($k = 4$), there would be value in further exploring the effects of spacing on L2 pronunciation learning.

Longer spacing promoted better retention for L2 grammar than shorter spacing. One explanation is that learners' comprehension can be impaired by shorter spacing between presentations of different (but related) types of grammatical rules, leading to undesirable difficulties (Metcalfe, 2011). However, learners may devote more attention or processing effort to longer spaced conditions (Jacoby, 1978). Interleaving can benefit the retention of grammatical features (e.g., Nakata & Suzuki, 2019b). Interleaved practice requires that learners repeatedly switch between different kinds of intervening tasks for the target features, which improves discriminability (Taylor & Rohrer, 2010). However, given that the number of blocked and interleaved practice studies on grammar learning was small (Nakata & Suzuki, 2019b; Pan et al., 2019; Suzuki et al., 2020), researchers should be cautious in interpreting the effects of blocking and interleaving for L2 grammar learning. Shintani et al. (2013) found large effects of comprehension-based instruction (e.g., error identification) on receptive knowledge of L2 grammar, $d = 1.09$, 95% CI [0.64, 1.55], and small effects of production-based instruction (e.g., translation) on productive knowledge, $d = -0.21$, 95% CI [−0.39, −0.02]. Shintani's (2015) meta-analysis revealed very large effects of processing instruction (e.g., structured input activities) on receptive knowledge, $d = 2.60$, 95% CI [2.19, 3.00], and productive knowledge, $d = 2.03$, 95% CI [1.65, 2.41], of L2 grammar. We found a small-to-medium effect of spaced practice for L2 grammar learning ($g = 0.56$ for overall effect; $g = 0.88$ for receptive knowledge, $g = 0.42$ for productive knowledge), which is smaller than that found by Shintani (2015) for comprehension-based and processing

instruction but larger than the effect Shintani found for production-based instruction (for details, see Table S7.2, Appendix S7, in the Supporting Information online).

Third, spacing manipulated within one session promoted better immediate L2 performance than spacing manipulated between sessions, but spacing manipulated between sessions led to better retention than spacing manipulated within one session. Because within-session spacing inevitably involves shorter spacing than between-session spacing, spaced practice within a single session may support higher levels of retrieval success at immediate posttests than spaced practice between multiple sessions. The effects of between-session spacing on long-term retention support the distributed practice effect (e.g., Bahrick, Bahrick, Bahrick, & Bahrick, 1993), suggesting that longer spacing (time intervals between multiple sessions are relatively longer than intervals within a session) yields better retention. However, we found a greater effect of longer spacing for the retention of L2 vocabulary when the spacing was manipulated within a single session, g = 0.79, 95% CI [0.32, 1.25], than when it was manipulated between multiple sessions, g = 0.02, 95% CI [−0.23, 0.26]. It should be noted that all within-session studies included in the longer versus shorter comparison (k = 11) involved a retrieval condition as practice. Consistent with study-phase retrieval account (proposing that the benefits of spacing arise from the effects of retrieving information from the first presentation during the second presentation, e.g., Toppino & Bloom, 2002) and the desirable difficulties framework (proposing the desirability of making study more difficult by increasing spacing, e.g., Bjork, 1994), increasing spacing within a single session might be expected to produce superior retention when it involves retrieval conditions.

Fourth, when longer spacing was involved, greater retention occurred in test–restudy trials than in study-only trials. Specifically, the effect of longer spacing was greater in L2 vocabulary learning, g = 1.27, 95% CI [0.75, 1.78], than in L2 grammar learning, g = 0.84, 95% CI [0.39, 1.29]. Consistent with study-phase retrieval theory and the desirable difficulties framework, increasing spacing between test–restudy trials represents a condition that requires more effort, leading to greater learning than study-only trials. It is also notable that we found no clear effects for equal versus expanding spacing in either retrieval or restudy practice. This might be explained by study time and time available to take a posttest. Gerbier and Koenig (2012), in their Experiment 1, allowed unlimited time for studying and

performing the posttest and found the superiority of expanding spacing. However, Gerbier and Koenig in their Experiment 2 and Schuetze (2014) controlled studying time and time on posttest, and they found no benefits for expanding versus equal spacing. Although learning is desirably difficult in the case of spaced practice, learners may compensate for this difficulty by spending more time on tasks (Gerbier & Koenig, 2012).

Fifth, spacing with comprehension activities enhanced learning and retention of L2 vocabulary, g = 1.38−1.56, 95% CI [0.87, 2.08]. However, it is notable that no clear spacing effect was found with paired associates on the immediate posttests. This might be because a paired-associate task has a fast presentation rate (shorter study time), and learners may not encode what they need for deep and useful encoding during practice (Metcalfe, 2011). As the desirable difficulty perspective recommends, massing may be advantageous when initial encoding has not been completed during the first presentation. This suggests that spacing may work at slower presentation rates; during spaced conditions, more study time is needed to encode.

Sixth, there were greater effects of spacing relative to massed practice on L2 vocabulary learning, g = 1.42, 95% CI [0.86, 1.99], when feedback was provided than when feedback was not provided. As the desirable difficulty perspective recommends, spacing after processing feedback can provide a learner with a desirably difficult learning condition on the subsequent trial, improving subsequent retention. However, we found that feedback did not have much impact on the differences between equal and expanding spacing conditions. Cepeda et al. (2006) mentioned that the variability in effects between equal and expanding spacing could be explained by the presence or absence of feedback, which was often a potential confound in the studies comparing these two conditions. However, our findings suggest that the differences in effects across these spacing conditions might be impacted by other variables, rather than feedback.

Seventh, delayed feedback influenced the effects of spaced practice for retention. There were larger effects of spaced practice on L2 vocabulary learning when delayed feedback was provided (g = 2.34, 95% CI [1.64, 3.04], in the spaced vs. massed comparison; g = 0.64, 95% CI [0.15, 1.14], in the longer vs. shorter comparison) than when immediate

feedback was provided (g = 1.04, 95% CI [0.59, 1.49], in the spaced vs. massed comparison; g = 0.37, 95% CI [−0.16, 0.90], in the longer vs. shorter comparison). In the current meta-analysis, most vocabulary studies that provided delayed feedback manipulated spacing between multiple sessions (between-session spacing, k = 6) rather than within one session (within-session spacing, k = 2), whereas vocabulary studies that provided immediate feedback more often involved within-session spacing (k = 9) than between-session spacing (k = 4). One explanation is that delayed feedback that is also between multiple sessions provides (even) longer spacing intervals between opportunities of feedback for a given item than does immediate feedback within one session. This supports distributed practice effects (Bahrick et al., 1993), suggesting that longer spacing promotes better retention. Delayed feedback can also decrease the competition between a learner's incorrect response and the correct response, because an incorrect response tends to be forgotten over time (Butler et al., 2007).

In the current meta-analysis, almost all the studies that provided delayed feedback after a test trial targeted L2 vocabulary: Only one L2 grammar study included delayed feedback (Bird, 2010), whereas seven grammar studies included immediate feedback, and one L2 pronunciation study included immediate feedback (Li & DeKeyser, 2019). We should be careful in interpreting the effects of feedback timing on L2 grammar and pronunciation learning, and further research in this area would be valuable.

It is noteworthy that delayed feedback provided in classroom-based studies with paper-and-pencil tasks (Bird, 2010) and computer-based studies (e.g., Gerbier, Toppino, & Koenig, 2015) may lead to different recall rates, because it may be possible for learners to look over all of their responses on the papers in the classroom-based studies, whereas this might not be the case with computer-based delayed feedback. Therefore, the operationalization of feedback timing should be carefully considered when a study is carried out with paper-and-pencil tasks.

Eighth, frequency of practice did not have a significant influence on the effects of spaced practice on L2 learning. However, a closer inspection of the data revealed that the results may be accounted for by other potential confounding variables. It was found that grammar studies included much greater frequency of practice than vocabulary studies (2−30

repetitions in grammar studies compared with 2−9 repetitions in vocabulary studies). Grammar studies that engaged greater values (e.g., 10−30 repetitions) showed differential effects of spaced practice in relation to number of sessions (i.e., whether the practice was manipulated within a session or between multiple sessions). The study by Suzuki et al. (2020) was a within-session study (10−11 repetitions) and showed a diminished spacing effect on the delayed posttest (g = 0.67 on the immediate posttest and g = 0.41 on the delayed posttest). However, the study by Suzuki (2017) was a between-session study (27−30 repetitions), and the effect did not attenuate on the delayed posttest (g = −0.63 on the immediate posttest and g = −0.64 on the delayed posttest [note that a negative value here indicates the superiority of a baseline condition relative to a treated condition]). This may suggest that spaced practice promotes better learning and retention of L2 grammar when the practice is manipulated between sessions rather than within a session (see Tables S9.5 and S9.6, Appendix S9, in the Supporting Information online).

Finally, the effects of expanding spacing were greater than those of equal spacing when the RI was longer. The authors of some previous studies have argued that the advantage produced by expanding spacing is strongly related to the timing of the first retrieval attempt during practice (e.g., Carpenter & DeLosh, 2005). However, Logan and Balota (2008) found that fewer items (low associate word pairs, e.g., cloth-sheep) were recalled in the expanding condition compared to the equal condition on a 24-hour-delayed posttest. Furthermore, in our data, L2 studies that controlled the timing of the first retrieval attempt (e.g., Gerbier & Koenig, 2012; Gerbier et al., 2015) found expanding spacing to be superior to equal spacing on 2-day-delayed posttests. Consistent with contextual variability theory (e.g., Melton, 1970) and the accessibility principle (e.g., Jacoby, 1978), the gradual expansion of spacing between learning opportunities can lead to greater contextual variation and serve to decrease the accessibility of a target item but increase reprocessing of the item in spaced repetitions. Overall, our findings suggest that the timing of the final posttest and gradual expansion of the spacing interval between learning opportunities (rather than the timing of the initial retrieval attempt) may have a profound effect on spaced practice. However, as only two studies controlled for the initial retrieval attempt, more research is warranted to test this interpretation.

It is pertinent to mention that some of the results of the moderator analyses (age, learning target, activity type, feedback timing) as interpreted above were not statistically significant due to small study sample sizes. However, tentative explanations were offered because the findings could be noteworthy, and we hope that these explanations will provide some direction for future research initiatives.

We turn now to the pedagogical implications of our findings. There are many such implications for both young and adult L2 learners. First, teachers may need to revisit target words over spaced time intervals. However, the analyses indicated that it might be useful to space the learning of pronunciation rules with shorter rather than longer spacing, specifically when the rules are not easily distinguished from each other. This may allow students the time needed to recognize the patterns and fully comprehend the rules. Second, teachers may need to revisit target words across a single session. For better retention, teachers could use longer spacing within a single session and/or, for likely even larger benefits, (also) space items over multiple days. Third, teachers may need to intersperse spaced retrieval (i.e., tests) with some kind of restudying practice. For example, teachers could revisit target words that had not been correctly recalled by students when tested or could provide feedback with a delay (e.g., feedback given after testing all items). Furthermore, there could be some value in spaced learning with comprehension activities (e.g., reading sentences or listening to words, followed by comprehension questions), but teachers may need to make sure that the activities are desirably challenging for students and that there is sufficient study (or presentation) time to help students fully comprehend target items or features (e.g., Hausman & Kornell, 2014).

2.6 Limitations and Future Directions

This meta-analysis identified several limitations that would be useful to address in further research on spaced practice. First, there have been comparatively few studies of relative spacing (i.e., equal or expanding spacing). Second, we found a need for additional research investigating the effects of spacing on L2 learning that (a) involves young learners, (b) targets L2 grammar and pronunciation learning, (c) includes production activities, (d) includes delayed feedback, and (e) measures productive knowledge. Moreover, there is a need for

clearer reporting of participants' L2 proficiency (as also observed in the synthesis by Park, Solon, Dehghan-Chaleshtori, & Ghanbar, 2021), which could help teachers to understand how learner differences may interact with the effects of spaced practice. Although learners may be learning through the same activities across and within courses, their L2 proficiency (and aptitude) will vary. Differential effects of spacing might be expected for learners of one proficiency level as compared to learners of a different proficiency level in the same learning condition (see Serrano, 2011). Finally, we were not able to rule out publication bias in the current meta-analysis. Therefore, the overall effects of spaced practice on L2 learning from the current synthesis should be interpreted with caution.

2.7 Conclusion

This meta-analysis revealed that although the spacing effect was robust, the size was in the range of small to medium (g = 0.58) for immediate effects (i.e., immediately after the last training session) and medium to large (g = 0.80) for delayed effects (i.e., a delay of one day or greater following the treatment). It also revealed that longer spacing was more effective than shorter spacing for long-term retention (small-to-medium effect, g = 0.40), but that learning gains were not significantly different between the equal and expanding spacing conditions. Some of the differences between the effects of different spacing conditions were explained by particular variables (e.g., learning target, number of sessions).

2.8 Notes

1 An anonymous reviewer pointed out that there were some studies (k = 12) that involved different types of posttests (e.g., receptive and productive) administered as immediate posttests, and that in such cases each different type of posttest could be considered as a separate learning session when coding the frequency of practice. To examine whether this affected the results, we did further analyses. We coded multiple types of posttests as one learning session and also, separately, we coded multiple

types of posttests as separate learning sessions. We did the analyses in both ways, and the results showed no difference (see Appendix S9 in the Supporting Information online for details).

2 An anonymous reviewer pointed out that retention interval is a key variable in examining spaced practice effects and suggested that the first delayed posttest should be used as a dependent variable (for examining delayed effects) when a study involved two or three delayed posttests. We recoded and further analyzed whether this choice affected the results. We found no statistically significant difference between our earlier coding (where the interval between the first or second delayed posttest [if there were three delayed posttests] and the final delayed posttest was used to examine delayed effects) and this coding suggested by the reviewer (where the first delayed posttest was used to examine delayed effects) in both the spaced versus massed comparison and the longer versus shorter spacing comparison (see Appendix S9 in the Supporting Information online for details).

3 Serrano and Huang's (2018) study was excluded because their RI was manipulated between participants, not within participants.

2.9 References

Note. The full reference list of the studies included in the meta-analysis is available in Appendix S10 in the Supporting Information online.

Avery, N., & Marsden, E. J. (2019). A meta-analysis of sensitivity to grammatical information during self-paced reading: Towards a framework of reference for reading time effect sizes. *Studies in Second Language Acquisition*, *41*(5), 1055-1087. https://doi.org/10.1017/S0272263119000196

Baddeley, A. (1999). *Human memory: Theory and practice* (revised ed.). East Sussex: Psychology Press. https://doi.org/10.1016/s0145-2134(00)00166-6

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2015). *Memory* (2nd ed). New York: Psychology Press. https://doi.org/10.4324/9781315749860

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575–589. https://doi.org/10.1016/S0022-5371(75)80045-4

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*(3), 296–308. https://doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*(5), 316–321. https://doi.org/10.1111/j.1467-9280.1993.tb00571.x

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *31*, 635−650. http://doi.org/10.1017/S0142716410000172

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). NJ: Lawrence Erlbaum Associates. https://doi.org/10.2307/1421430

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/4561.003.0011

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research 74*(4), 245−248. http://doi.org/10.1080/00220671.1981.10885317

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley and Sons, Ltd. https://doi.org/10.1002/9780470743386

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2013). Comprehensive Meta-Analysis Version 3. Englewood, NJ: Biostat, Inc.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(5), 5–18. https://doi.org/10.1002/jrsm.1230

Brosvic, G. M., Epstein, M. L., Cook, M. J., & Dihoff, R. E. (2005). Efficacy of error for the correction of initially incorrect assumptions and of feedback for the affirmation of correct responding: Learning in the classroom. *Psychological Record*, *55*(3), 401–418. https://doi.org/10.1007/BF03395518

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

Carpenter, S. K. (2017). Spacing effects on learning and memory. In J. T. Wixted (Ed.), *Cognitive psychology of memory*, *Vol. 2 of Learning and memory: A comprehensive reference* (2nd ed.) (pp. 465−485). Amsterdam, The Netherlands: Academic Press. https://doi.org/10.1016/b978-0-12-809324-5.21054-7

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*(5), 619–636. https://doi.org/10.1002/acp.1101

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*, 671−682. http://doi.org/10.3758/s13421-012-0291-4

Çekiç, A., & Bakla, A. (2019). The effects of spacing patterns on incidental L2 vocabulary learning through reading with electronic glosses. *Instructional Science, 47*(3), 353−371. https://doi.org/10.1007/s11251-019-09483-4

Cepeda, N. J., Mozer, M. C., Coburn, N., Rohrer, H., Wixted, J. T., & Pashler, H. (2005). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236–246. https://doi.org/10.1027/1618-3169.56.4.236

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.

Crothers, E., & Suppes, P. (1967). *Experiments in second-language learning*. New York: Academic Press. https://doi.org/10.1016/b978-0-12-395568-5.50010-7

DeKeyser, R. M. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511667275

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*(5), 795–805. https://doi.org/10.1037/0021-9010.84.5.795

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*(2), 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Gerbier, E., & Koenig, O. (2012). Influence of multiple-day temporal distribution of repetitions on memory: A comparison of uniform, expanding, and contracting schedules. *The Quarterly Journal of Experimental Psychology*, *65*(3), 514−525. http://doi.org/10.1080/17470218.2011.600806

Gerbier, E., Toppino, T. C., & Koenig, O. (2015). Optimising retention through multiple study opportunities over days: The benefit of an expanding schedule of repetition. *Memory*, *23*(6), 943−954. http://doi.org/10.1080/09658211.2014.944916

Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in Memory and Cognition*, *3*(3), 153–160. https://doi.org/10.1016/j.jarmac.2014.03.003

Hunter, J., & Schmidt, F. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. London: SAGE Publications.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 649–667. https://doi.org/10.1016/S0022-5371(78)90393-6

Kang, S., Lindsey, R., Mozer, M., & Pashler, H. (2014). Retrieval practice over the long time: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, *21*(6), 1544−1550. http://doi.org/10.3758/s13423-014-0636-z

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250−1257. http://doi.org/10.1037/a0023436

Kasprowicz, R., 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. http://doi.org/10.1111/modl.12586

Khoii, R., & Abed, K. F. (2017). Effects of equal spacing, expanding spacing, and

massed condition on EFL learners' receptive and productive vocabulary retrieval. In Pixel, *Proceedings of ICT for language learning* (19th ed.) (pp. 500−504). Florence, Italy: ICT for Language Learning. Retrieved from https://conference.pixel-online.net/ICT4LL/files/ict4ll/ed0010/FP/0960-SLA2580-FP-ICT4LL10.pdf

Kim, S. K., & Webb, S. (2021). *Coding scheme. Materials from "The effects of spaced practice on second language learning: A meta-analysis"* [Coding scheme]. IRIS Database, University of York, UK. https://doi.org/10.48316/rn3w-1b17

Koval, N. G. (2019). Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics*, *40*(5), 1−37. http://doi.org/10.1017/S0142716419000158

Koval, N. G. (2020). *Testing the reminding account of the lag effect in L2 vocabulary acquisition from L2-L1 retrieval practice within a paired-associate learning format* (Published doctoral dissertation). Michigan State University. http://doi.org/10.25335/pg4k-p594

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. https://doi.org/10.1007/s11251-013-9285-2

Lawrence, N. K. (2013). Cumulative exams in the introductory psychology course. *Teaching of Psychology*, *40*(1), 15–19. https://doi.org/10.1177/0098628312465858

Lee, E., & Choe, M-H. (2014). The effect of spaced repetitions on Korean elementary students' L2 English vocabulary learning. *Studies in English Education*, *19*(1), 55−75.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Li, M., & DeKeyser, R. (2019). Distribution of practice effects in the acquisition and retention of L2 Mandarin tonal word production. *The Modern Language Journal*, *103*(3), 607−628. http://doi.org/10.1111/modl.12580

Lightbown, P. M., & Spada, N. (1994). An innovative program for primary ESL students in Quebec. *TESOL Quarterly*, *28*(3), 563–579. https://doi.org/10.2307/3587308

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin and Review*, *21*(4), 861–883. https://doi.org/10.3758/s13423-013-0565-2

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. London: SAGE Publications. https://doi.org/10.1177/01632780122034902

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *15*(3), 257–280. https://doi.org/10.1080/13825580701322171

Lotfolahi, A. R., & Salehi, H. (2016). Learners' perceptions of the effectiveness of spaced learning schedule in L2 vocabulary learning. *SAGE Open, 6*(2), 1−9. http://doi.org/10.1177/2158244016646148

Lotfolahi, A. R., & Salehi, H. (2017). Spacing effects in vocabulary learning: Young EFL learners in focus. *Cogent Education*, *4*(1), 1–10. https://doi.org/10.1080/2331186X.2017.1287391

Maddox, G. B., & Balota, D. A. (2015). Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty. *Memory and Cognition*, *43*(5), 760–774. https://doi.org/10.3758/s13421-014-0499-6

Maddox, G. B., Balota, D. A., Coane, J. H., & Duchek, J. M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology and Aging*, *26*(3), 661–670. https://doi.org/10.1037/a0022942

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, 9(5), 596–606. https://doi.org/10.1016/S0022-5371(70)80107-4

Metcalfe, J. (2011). Desirable difficulties and studying in the region of proximal learning. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 259–276). London / New York: Psychology Press. https://doi.org/10.4324/9780203842539-18

Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System, 42*, 412−428. http://doi.org/10.1016/j.system.2014.01.014

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Nakata, T. (2015a). 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. http://doi.org/10.1017/S0272263114000825

Nakata, T. (2015b). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research*, *19*(4), 416–434. https://doi.org/10.1177/1362168814541721

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, *39*(4), 653–679. https://doi.org/10.1017/S0272263116000280

Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, *37*(2), 233–260. http://doi.org/10.1177/0267658320927764

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

Nakata, T., & Suzuki, Y. (2019b). Mixing grammar exercises facilitates long-term retention: Effects of blocking, interleaving, and increasing practice. *The Modern Language Journal*, *103*(3), 629–647. http://doi.org/10.1111/modl.12581

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., & Rickard, T. C. (2019). Does interleaved

practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills. *Journal of Educational Psychology, 111,* 1172−1188. http://doi.org/10.1037/edu0000336

Park, H. I., Solon, M., Dehghan-Chaleshtori, M., & Ghanbar, H. (2021). Proficiency reporting practices in research on second language acquisition: Have we made any progress? *Language Learning (72),* https://doi.org/10.1111/lang.12475

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 3–8. https://doi.org/10.1037/0278-7393.31.1.3

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1051−1057. http://doi.org/10.1037/0278-7393.29.6.1051

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin*, *134*(2), 270–300. https://doi.org/10.1037/0033-2909.134.2.270

Peterson, L. R., Saltzman, D., Hillner, K., & Land, V. (1962). Recency and frequency in paired-associate learning. *Journal of Experimental Psychology*, *63*, 396–403. https://doi.org/10.1037/h0043571

Pinker, S. (1998). Words and rules. *Lingua*, *106*(1–4), 219–242. https://doi.org/10.1016/S0024-3841(98)00035-7

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

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*(8), 1917−1927. http://doi.org/10.3758/BF03192925

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437−447. http://doi.org/10.1016/j.jml.2009.01.004

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*(4), 857−866. http://doi.org/10.1002/tesq.252

Rogers, J., & Cheung, A. (2020a). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*, *24*, 616–641. http://doi.org/10.1177/1362168818805251

Rogers, J., & Cheung, A. (2020b). Does it matter when you review? Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition*. http://doi.org/10.1017/S0272263120000236

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

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

Schuetze, U. (2014). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research*, *19*(1), 28−42. http://doi.org/10.1177/1362168814541726

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, *19*(1), 107–122. https://doi.org/10.1002/acp.1066

Serrano, R. (2011). The time factor in EFL classroom practice. *Language Learning*, *61*(1), 117–145. https://doi.org/10.1111/j.1467-9922.2010.00591.x

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. http://doi.org/10.1002/tesq.445

Shintani, N. (2015). The effectiveness of processing instruction and production-based instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics*, *36*(3), 306–325. https://doi.org/10.1093/applin/amu067

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus productive-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, *63*(2), 296–329. https://doi.org/10.1111/lang.12001

Snoder, P. (2017). Improving English learners' productive collocation knowledge: The effects of involvement load, spacing, and intentionality. *TESL Canada Journal*, *34*(3), 140−164. http://doi.org/10.18806/tesl.v34i3.1277

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. http://doi.org/10.1111/lang.12236

Suzuki, Y. (2018). The role of procedural learning ability in automatization of L2 morphology under different learning schedules. *Studies in Second Language Acquisition*, *40*(4), 923–937. https://doi.org/10.1017/S0272263117000249

Suzuki, Y. (2019). Individualization of practice distribution in second language grammar learning: A role of metalinguistic rule rehearsal ability and working memory capacity. *Journal of Second Language Studies*, *2*(2), 170–197. http://doi.org/10.1075/bct.116.02suz

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

Suzuki, Y., & DeKeyser, R. (2017b). Exploratory research on second language practice distribution: An aptitude × treatment interaction. *Applied Psycholinguistics*, *38*(1), 27–56. https://doi.org/10.1017/S0142716416000084

Suzuki, Y., Nakata, T., & DeKeyser, R. M. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, *103*(3), 713–720. https://doi.org/10.1111/modl.12585

Suzuki, Y., Yokosawa, S., & Aline, D. (2020). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research*. http://doi.org/10.1177/1362168820913985

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837–848. https://doi.org/10.1002/acp.1598

Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 437–444. https://doi.org/10.1037//0278-7393.28.3.437

Toppino, T. C., & DiGeorge, W. (1984). The spacing effect in free recall emerges with development. *Memory and Cognition*, *12*(2), 118–122. https://doi.org/10.3758/bf03198425

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599. https://doi.org/10.1111/lang.12343

Ullman, M. T. (2015). The declarative/procedural model: A neurobiologically motivated theory of first and second language. In B. Van Patten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 135–158). New York: Routledge. https://doi.org/10.4324/9780429503986-7

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, *11*(5), 713–729. https://doi.org/10.1177/174569116659874

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Ö zsoy, B. (2008). Distributed rereading can hurt the spacing effect in text memory. *Applied Cognitive Psychology*, *22*(5), 685–695. https://doi.org/10.1002/acp.1388

Wickelgren, W. A. (1972). Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology*, *9*(4), 418–455. https://doi.org/10.1016/0022-2496(72)90015-6

Wilson, W. P. (1976). Developmental changes in the lag effect: An encoding hypothesis for repeated word recall. *Journal of Experimental Child Psychology*, *22*(1), 113–122. https://doi.org/10.1016/0022-0965(76)90094-1

Chapter 3: Does Spaced Practice Have the Same Effects on Different Second Language Vocabulary Learning Activities? Fill-in-the-blanks versus Flashcards

3.1 Introduction

There are many different activities that can be used to learn words. Webb and Nation (2017) described 23 approaches to developing vocabulary knowledge, while Morgan and Rinvolucri (2004) profiled 118 activities designed for word learning. For example, learners can memorize target words with their translations or synonyms using flashcards, match words to their meanings in matching activities, write target words in given sentences in fill-in-the-blanks exercises, and write original sentences using target words through sentence production tasks. With so many different approaches to learning vocabulary, it is important to understand the extent to which different approaches are effective.

It is important to consider the conditions within activities that contribute to learning in order to understand their effectiveness. The ways in which activities are performed provide for certain learning conditions (e.g., repetition, varied encounters and use, retrieval) that can influence the amount and quality of word learning (Webb & Nation, 2017). Numerous second language (L2) vocabulary studies have shown that encountering words repeatedly is an effective method of vocabulary learning (e.g., Horst, Cobb, & Meara, 1998; Pigada & Schmitt, 2006; Webb, 2007; Webb & Chang, 2015). Other investigations have evaluated repeated study practice (i.e., restudy) against repeated retrieval practice (e.g., Barcroft, 2007; Royer, 1973). Retrieval practice refers to the practice of testing the information or knowledge studied (Roediger & Guynn, 1996). Studies comparing restudy to repeated retrievals have shown that repeated retrieval practice promotes better L2 vocabulary learning and retention than repeated study. Further, both cognitive psychology and L2 vocabulary acquisition studies (e.g., Bahrick, 1979; Karpicke & Bauernschmidt, 2011; Nakata & Suzuki, 2019; Rogers & Cheung, 2020) support the idea that providing an interval between repetitions in learning (spaced practice) improves long-term retention of L2 vocabulary more than do repetitions that occur in immediate succession without any intervals (massed practice).

Spacing effect refers to the phenomenon that spaced practice enhances retention relative to massed practice (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

One limitation of L2 vocabulary learning studies of spaced practice is that studies have almost exclusively focused on paired-associate learning (i.e., learning from word pairs that consist of target words and their first language (L1) meanings, e.g., flashcards). Although learning from flashcards is considered to be a fast and efficient method of learning L2 vocabulary (Nation, 2013), when vocabulary is taught in the classroom, teachers are also likely to use many other vocabulary learning activities. This means that although previous studies have demonstrated the positive effects of spacing on L2 vocabulary learning, the effects cannot yet be generalized to different L2 vocabulary learning conditions. It is, therefore, valuable to look at the extent to which spacing contributes to vocabulary learning in different vocabulary learning activities.

The present study aimed to examine the effects of spacing on learning through fill-in-the-blanks activities and comparing this with learning through flashcards. Because fill-in-the-blanks activities are one of the most commonly used activities for vocabulary learning, this study may provide useful evidence about the degree to which spaced retrieval practice may contribute to vocabulary learning in other ways apart from flashcards.

3.2 Background

3.2.1 Effects of Spacing on L2 Vocabulary Learning

There have been many studies investigating the effects of spacing on L2 vocabulary learning (e.g., Bahrick, 1979; Nakata & Suzuki, 2019; Rogers & Cheung, 2020). Kim and Webb (in press) meta-analyzed forty-eight experiments from 37 studies investigating the effects of spaced practice on L2 learning and found large effects of spacing on L2 vocabulary learning and retention; spaced practice led to greater learning (measured immediately after the treatment, g = 0.76, 95% CI [0.26, 1.25]) and retention (measured one day or greater following the treatment, g = 1.15, 95% CI [0.81, 1.49]) than massed practice.

Most studies investigating the effects of spaced practice on deliberate L2 vocabulary learning have used flashcards or word pairs as the approach to learning. Only two studies have investigated the effects of spaced practice using other deliberate vocabulary learning activities (Bloom & Shuell, 1981; Rogers & Cheung, 2021). Bloom and Shuell (1981) compared spaced and massed practice on French word learning through multiple-choice, fill-in-the-blanks, and form recall (from L1 to L2) activities. Half the participants were placed in a spaced condition and completed three 10-minute worksheets over three days, with the other half assigned to a massed condition and completed all three worksheets on one day. The three worksheets were multiple-choice, fill-in-the-blanks, and form recall exercises. Bloom and Shuell observed no significant differences between the spaced and massed groups on an immediate posttest (mean percentage: spaced = 84.25%, massed = 80.6%). However, a significant advantage was found for the spaced condition on a 4-day delayed posttest (mean percentage: spaced = 75.20%, massed = 55.75%). This finding suggests that there may be a positive effect of spacing in other vocabulary learning conditions. However, because the participants learned the target words in each of the three different exercises, the study did not show the benefits of spacing pertaining to specific exercises.

Rogers and Cheung (2021) examined the effects of spacing using crossword puzzles as practice. This experiment consisted of three training sessions with three intact Chinese primary school classes. In each class, half the target words were subjected to a shorter spaced condition (1 day), and the other half were in a longer spaced condition (8 days). Each training session involved a PowerPoint presentation to teach the target words from a word list, followed by a crossword puzzle to practice the target words. Each PowerPoint slide was animated to present information in a sentence including the target word. Feedback from the teachers was provided after the puzzle practice. A posttest was administered 28 days after the last session. Rogers and Cheung found no significant difference between the shorter-spaced and the longer-spaced conditions.

A related line of research suggests that feedback timing, provided immediately after retrieval attempts or with a delay, may affect learning and memory (e.g., Butler, Karpicke, & Roediger, 2007; Guo, 2021; Roediger & March, 2005). Immediate feedback provided after each retrieval may be useful for learning because it can make learners fully process feedback

after both successful and unsuccessful retrievals (Butler & Roediger, 2007). The value of delayed feedback is that it can provide learners with more effortful learning circumstances, which may strengthen retention (Desirable difficulty framework, Bjork, 1994). In a meta-analysis of studies of L2 spaced practice, Kim and Webb (in press) examined whether feedback timing moderates the effects of spaced practice and found large effects of immediate feedback (g = 1.04, 95% CI [0.59, 1.49]) and delayed feedback (g = 0.64~2.34, 95% CI [0.15, 3.04]) for the retention of L2 vocabulary. However, all of the included studies of deliberate vocabulary learning involved learning through flashcards or word lists, and so the degree to which feedback timing affects vocabulary learning in most other conditions remains to be explored. Recently, Guo (2021) directly examined effects of feedback timing (immediate and delayed) in two different spaced conditions (1-day and 3-day intervals) from textbook glosses, followed by post-reading activities to provide the participants with extended exposures of the target words. Guo (2021) found the superiority of delayed feedback over immediate feedback for the retention of L2 vocabulary (measured two weeks after the treatment). Guo's findings together with earlier studies of paired associate learning indicate that feedback timing might be a useful variable to examine in relation to the efficacy of different L2 vocabulary learning activities.

3.3 The Current Study

The aim of the present study was to investigate whether spacing had a similar effect on the learning and retention of L2 vocabulary in fill-in-the-blank and flashcard activities. Fill-in-the-blanks was selected because it was found to be one of the most commonly used exercises for learning and teaching of L2 vocabulary in an analysis of 10 L2 English textbooks (see Appendix A for a compilation of activity types found in the L2 English textbooks in the online supplementary material). Flashcards was chosen as the comparison condition because it has been frequently used in studies of spaced practice. Participants performed either fill-in-the-blanks or flashcards under one of the two (massed and spaced) practice schedules. Posttest formats were matched to the practice formats (contextualized form recall and form recall) to ensure that the correspondence between vocabulary learning condition and test

format did not bias gains in word knowledge towards one condition. The current study addresses the following research questions.

1. To what extent is vocabulary learned through fill-in-the-blank and flashcard act ivities using different types of spacing?
2. To what extent do vocabulary learning gains differ across the learning conditio ns?
3. Does the correspondence between test format and vocabulary learning condition affect gains in word knowledge?
4. To what extent does feedback timing affect vocabulary learning in fill-in-the-bl ank and flashcard activities?

3.4 Method

3.4.1 Participants

The participants were 150 Korean students from six universities (68 male and 82 female, M_{age} = 21.5, *SD* = 1.4) in South Korea. All participants had studied English for a minimum of eight years. Seven were English majors (English education or English literature) and the remaining participants were majoring in academic disciplines. Prior to the experiment, the participants took the five sections (from 1,000 up to 5,000 frequency levels) of the Vocabulary Levels Test (VLT; Webb, Sasao, & Ballance, 2017) to measure their lexical knowledge. The average scores (standard deviation [SD]) of the participants on the VLT were 98% (4.1) at the 1000 word level, 93% (13.1) at the 2000 word level, 89% (14.7) at the 3000 word level, 80% (15.9) at the 4000 word level, and 78% (17.4) at the 5000 word level.

The participants were randomly assigned to a control and four experimental (two learning conditions x two spacing schedules) groups. Most spaced practice research (e.g., Karpicke & Bauerschmidt, 2011; Nakata, 2015) has included a massed group for the control,

but to control for learning in the presentation phase and testing effects, a no treatment control group (n = 30) was also included in this study. The four experimental groups completed the following conditions: fill-in-the-blanks with massed (n = 30) and spaced (n = 30), flashcards with massed (n = 30) and spaced (n = 30).

3.4.2 Target Items

Forty-eight low frequency English words from the most frequent 8,000 to 16,000 word families in Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) lists were selected as target items (see Appendix B). Low frequency English words were selected to increase the likelihood that participants were unfamiliar with the items. The target items consisted of 28 nouns and 20 verbs, following the 6:4 ratio of nouns to verbs in natural text (Webb, 2005). The average number of letters of the target words was 5.77 (SD = 1.10). The average concreteness score for the target words, based on the crowd-sourced norms by Brysbaert, Warriner, and Kuperman (2014), was 4.02 (SD = 0.75).

3.4.3 Instructional Treatment

Each participant had access to a computer with the software PsychoPy installed to present the treatments and collect data on learning and performance in the presentation phase, practice phase, and tests (pretest, immediate, and delayed).

3.4.3.1 Presentation Phase

The target words were presented in a common format utilized in dictionaries to introduce vocabulary. Each target word was presented in bold font followed by its part of speech and L1 definition for 10 seconds, and participants listened to its pronunciation once (see screenshot on the left, Figure 1). A sentence example containing the underlined target word

and the L1 translation of the sentence were then presented for 15 seconds (see screenshot on the right, Figure 1). Sentences used in the presentation phase were sourced from the COCA (https://www.english-corpora.org/coca/) with lower frequency words replaced with words from the most frequent 1,000 and 2,000 word families to increase the likelihood that all of the sentences would be easily understood (see Appendix C for sentences used in the presentation phase).

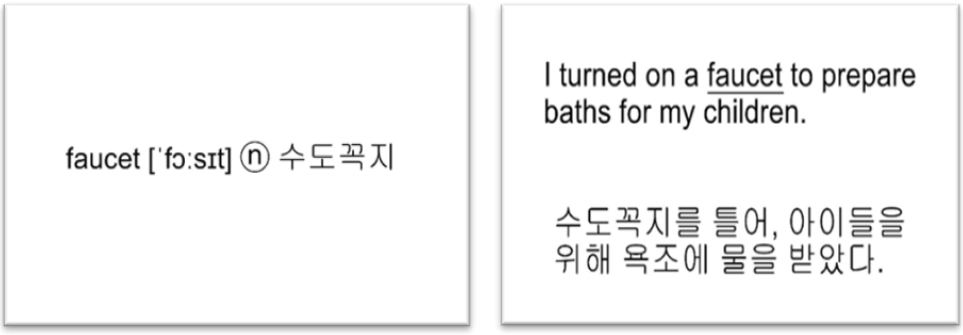


FIGURE 1 Screenshots of target word presentation during the treatment

3.4.3.2 Practice Phase: Fill-in-the-blanks Group

In the fill-in-the-blanks exercise, participants were asked to type the appropriate English target word at the bottom of the screen to complete the gap in the provided sentence (e.g., I heard the _____ run in the bathroom) (see screenshot on the left, Figure 2). Creating the fill-in-the-blank sentences for the target words involved two steps. First, sentences were reviewed in the COCA (https://www.english-corpora.org/coca/). Next, the words used in the sentences were simplified if they did not belong to the most frequent 1,000 and 2,000 word families (see Appendix D for the sentences used in the fill-in-the-blanks). Simplifying sentences made it more likely that the sentences were easily understood, and that learning was a function of completing the fill-in-the-blanks exercise.

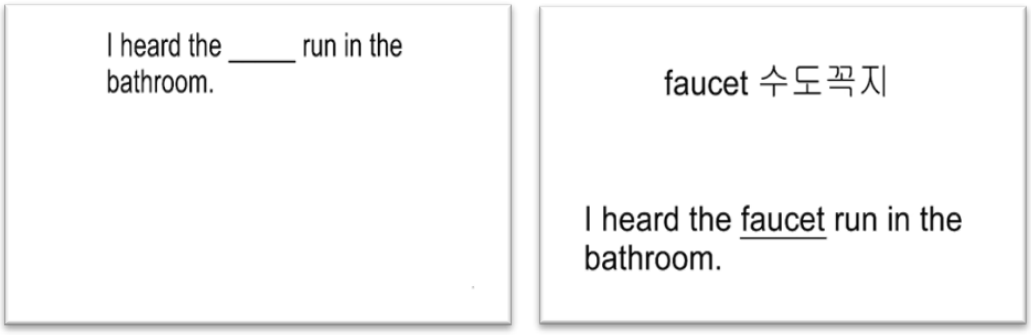


FIGURE 2 Screenshots of the fill-in-the-blanks question for the target item *faucet* (left) and feedback (right)

Nakata and Webb (2016) suggested that the correct answers should be made available within an exercise to provide learners with feedback on their responses. In this study, immediate feedback (feedback given after completing each item) was applied to half the target items (24 items), and the remaining items were placed under a delayed feedback condition (feedback given after completion of all 24 items). These items were counterbalanced between participants to ensure that any differences in learning within the two feedback conditions were not due to word-related variables (see Appendix E). The target word and its Korean definition and the sentence with the underlined target word were provided as feedback for 10 seconds (see screenshot on the right, Figure 2).

3.4.3.3 Practice Phase: Flashcards Group

Flashcards depicted L2 words and their L1 definitions (e.g., faucet – 수도꼭지). In the practice phase, the participants were asked to type the English target word corresponding to the Korean definition provided on the screen (e.g., 수도꼭지 = _____) (see screenshot on the left, Figure 3). The practice phase employed productive retrieval (i.e., retrieving the L2 word form). Productive retrieval is more demanding than receptive retrieval (i.e., retrieving the L1 meaning of the L2 word) but desirable because it may result in larger vocabulary gains

(e.g., Mondria & Wiersman, 2004; Webb, 2009). The target word and its Korean definition were provided as feedback for 10 seconds (see screenshot on the right, Figure 3) through immediate feedback for half the target items (24 items), while the remaining 24 items were placed under the delayed feedback condition (feedback given after completion of all 24 items). The type of feedback applied to the items was counterbalanced between participants. Research has suggested that time on task should be considered when comparing the effects of different treatments (e.g., Webb, 2005). Participants in this study completed each learning condition over the amount of time required for each exercise. However, the amount of time taken to complete each exercise was collected and included as a covariate.

FIGURE 3 Screenshots of the flashcard question (left) and feedback (right)

3.4.4 Spacing Schedules

Participants were randomly assigned to one of the two different types of spacing (massed and spaced). All participants learned the 48 items in the same manner in the presentation phase with a practice session in the assigned exercise (fill-in-the-blanks or flashcards) following the presentation phase. Participants in the massed condition retrieved the items five times using the assigned exercise within one session. Participants in the spaced (1-day interval) condition retrieved the items over five sessions (one retrieval attempt in each session) from Monday to Friday (one session per day) in the assigned exercise. Thus, the only difference between the two spacing schedules in the assigned exercise was the interval between retrieval attempts for target items.

3.4.5 Measurement

Three types of tests, form recall, contextualized form recall, and sentence production were administered in this study. Form recall and contextualized form recall test formats corresponded with the learning conditions: the form recall test simulated the flashcard condition, and the contextualized form recall test simulated the fill-in-the-blanks condition. The sentence production test format was selected as a neutral test that did not favor either of the two learning conditions. Nation and Webb (2011) suggested that when comparing multiple learning conditions, it is useful to measure knowledge with tests that are sensitive to the gains made in individual learning conditions, as well as tests that do not favor either condition to provide a more accurate assessment of learning. Participants could take as much time as they needed to type responses on all tests.

In the form recall test, participants were asked to type the target word corresponding to the Korean definition provided on the screen. The aim of the form recall test was to determine whether the participants could connect L2 form of the target words with their L1 meanings and write the target words correctly. Therefore, if there were any spelling mistakes (e.g., faucel, faus, or fouset for *faucet*), the responses were marked incorrect.

In the contextualized form recall test, a sentence with a blank was provided and the participants were asked to type the appropriate target word to complete the blank (see Appendix F for the test items used in the contextualized form recall test). The aim of the contextualized form recall test was to determine whether the participants could retrieve correct target word to fill the blanks in the given sentences. Therefore, correct target words with correct spellings and grammatical functions of target words were determining factors for a correct answer (it should be noted that the grammatical function of the target words and the responses required in the test were always the same; i.e., inflected forms were never required).

In the sentence production test, participants were asked to make a sentence including the target word that corresponded with the Korean definition provided on the screen. The aim

of this test was to determine whether the participants could make sentences using target words corresponding to the L1 meanings provided. When a participant produced a sentence such as “My favorite color is *mauve*”, “I like *azaleas*”, or “She loves *croon*ing”, in which the correct target words were produced with correct spellings and the sentence was comprehensible and grammatically correct, the responses were marked as correct. Responses that included incorrectly spelled target words, target words with incorrect grammatical functions, were incomprehensible (e.g., Employer *fawn* the boss), or incomplete (e.g., A *weasel* is), were marked as incorrect.

3.4.5.1 Pretest

One week before the treatment, the form recall, sentence production, and contextualized form recall tests were administered in that order as the pretest. All 48 target items were included in each of these three tests. When a participant wrote a synonym rather than a target word (e.g., writing *wild pig* rather than *boar* or writing *shout* rather than *holler*), the participant was asked if he or she knew any other words that corresponded with the Korean definition provided to ensure that the recall formats did not underestimate knowledge.

3.4.5.2 Posttest

Posttests were administered immediately and 2 weeks after the treatment. The 48 target items were divided into six sets of eight items; half the items (24 items, three sets of 8 items) were tested on an immediate posttest, and the other half (24 items, three sets of 8 items) were tested on a delayed posttest (see Appendix E). Suzuki (2017) reported that when a study administered a posttest more than once, the first posttest can be regarded as a learning session. In this study, therefore, the items tested on the immediate and delayed posttests were different to ensure that measurement on the immediate and delayed posttests was for an equal number of learning sessions (frequency of retrieval practice). Each set of 8 items was randomly assigned to each of the three test formats (form recall, contextualized form recall,

and sentence production tests) in each of the posttests (see Appendix G). The order of the test items was randomized between tests for each participant to reduce the possibility of an order effect. The delayed posttest was conducted with no prior notice.

3.4.6 Procedure

All the participants were informed about the research procedure and completed a consent form after volunteering to take part in the study. In the initial session one week before the treatment, all participants completed the pretest and VLT. To ensure that participants had no knowledge of target items, only the data from the 150 participants who scored incorrectly on all target items on the pretest were included in the study.

The four experimental groups underwent three phases: presentation, practice, and testing. The control group only undertook the presentation and testing phases. Learning target items occurred through the presentation and practice phases. In the presentation phase, all groups learned the 48 target words. In the practice phase, the participants in the experimental groups learned target items 5 times in their assigned exercises. This practice provided the participants with opportunities to retrieve the words they had learned in the presentation phase.

After the presentation phase for the control group and practice phase for the experimental groups, the participants answered 10 2-digit additions (e.g., 56+78 = ?) as a filler task. The filler items were used as recency buffers during the treatment (e.g., Karpicke & Roediger, 2007). Following the filler task, the groups took the immediate posttest. The participants of all groups took the delayed posttest two weeks after the immediate posttest. Table 1 summarizes the procedures for all groups in this study.

Table 1 Procedures of the current study

All (one control and four experimental) groups

Pre-meeting	Pretest, VLT		
	Control (1 group)	Massed (2 groups)	Spaced (2 groups)
	Presentation phase Learning words **Testing phase** Immediate posttests (at the end of the session)	**Presentation phase** Learning words **Practice phase** *5* sessions of assigned exercise (within a session) **Testing phase** Immediate posttests (at the end of the session)	**Presentation phase** Learning words **Practice phase** 5 sessions of assigned exercise (1-day interval) **Testing phase** Immediate posttests (at the end of the last session)
Two weeks after the treatment	**Testing phase** 14-day delayed posttests	**Testing phase** 14-day delayed posttests	**Testing phase** 14-day delayed posttests

Note. 2 groups include fill-in-the-blanks and flashcard learning conditions.

3.4.7 Data Analysis

To examine the effects of spacing on vocabulary learning through different learning conditions, the scores from the immediate and delayed posttests were analyzed separately using a logistic mixed-effects model fit by maximum likelihood using the Ime4 software package in R (Bates, Mächler, Bolker, & Walker, 2015). The dependent variable was a binary

response (correct/incorrect). Fixed-effect predictors were learning condition (control, fill-in-the-blanks, flashcards) and spacing type (massed, spaced). First, the initial model started with intercept-only random models with learning condition and spacing type as fixed effects and time on task as a covariate. Second, interactions among the two fixed effects and one covariate (time on task) were added to the initial model. The alpha level of statistical significance was set at less than .05. Post hoc tests were conducted using the R package (Ismeans; Lenth, 2016) to compare the differences between groups. Effect sizes of the group effects (learning condition and spacing type) were calculated and interpreted based on Plonsky and Oswald's (2014) benchmark (small: $0.40 \leq$ Cohen's $d < 0.70$; medium: $0.70 \leq d < 1.00$; large: $1.00 < d$ for between-participants contrasts).

To determine whether the correspondence between test format and learning condition affects gains in vocabulary knowledge, the binary data from immediate and delayed posttests were analyzed separately using a logistic mixed-effects model, conducted with learning condition (fill-in-the-blanks, flashcards) and spacing type (massed, spaced) at subject level and test formats (form recall, contextualized form recall, sentence production) at item level, and time on task as a covariate.

To determine the extent to which feedback timing affects vocabulary learning in different vocabulary learning conditions, the binary data from immediate and delayed posttests were analyzed separately using a logistic mixed-effects model, conducted with learning condition (fill-in-the-blanks, flashcards) and spacing type (massed, spaced) at subject level and feedback timing at item level, and time on task as a covariate. The interactions between spacing and feedback timing were added.

3.5 Results

None of the participants demonstrated prior knowledge of any of the target words on pretests (form recall, contextualized form recall, and sentence production).[1] Note that the words tested on the immediate and delayed posttests were different (knowledge of 24 items from the treatment were assessed on the immediate posttest and the other 24 items were assessed on

the delayed posttest). Cronbach's alpha was .81 or higher (.81-.94) for all dependent measures (form recall, contextualized form recall, and sentence production) on the immediate and delayed posttests, indicating good reliability. Table 2 presents means (M), standard deviations (SD), and 95% confidence intervals (CIs) of the test formats on both immediate and delayed posttests in all five conditions.

3.5.1 To What Extent is Vocabulary Learned Through Fill-in-the-blank and Flashcard Activities Using Different Types of Spacing?

3.5.1.1 Immediate Posttest

Participants in the fill-in-the-blanks massed condition had mean scores of 6.57, 6.83, and 3.10 on the form recall, contextualized form recall, and sentence production immediate posttest, respectively, for a total mean score of 16.50 out of 24 on the three test formats combined. The mean gains from the pretest to the immediate posttest for the fill-in-the-blanks massed condition were statistically significant on each immediate posttest format and the total mean gain on the three test formats combined (*ps* < .001; see Appendix 2I). Participants in the fill-in-the-blanks spaced condition had mean scores of 7.60, 6.83, and 2.97 in the form recall, contextualized form recall, and sentence production immediate posttest for a total mean score of 17.40 out of 24. The mean gains from the pretest to the immediate posttest for the fill-in-the-blanks spaced condition were statistically significant on each immediate posttest format and the total mean gains on the three test formats combined (*ps* < .001; see Appendix 2I).

Participants in the flashcard massed condition had mean scores of 7.07, 5.50, and 3.50 in the form recall, contextualized form recall, and sentence production immediate posttest, respectively, for a total mean score of 16.07 out of 24 on the three test formats combined. The mean gains from the pretest to the immediate posttest in the flashcard massed condition were statistically significant on each immediate posttest format and the three test formats combined (*ps* < .001; see Appendix 2I). Participants in the flashcard spaced condition had mean scores of 7.17, 5.30, and 2.80 on the form recall, contextualized form recall, and sentence production immediate posttest for a total mean score of 15.27 out of 24. The mean gains from

the pretest to the immediate posttest in the flashcard spaced condition were statistically significant on each immediate posttest format and the three test formats combined (*ps* < .001; see Appendix 2I).

Participants in the no treatment control group had mean scores of 1.33, 0.67, and 0.70 on the form recall, contextualized form recall, and sentence production immediate posttest for a total score of 2.7 out of 24 on the three test formats combined. The mean gains from the pretest to the immediate posttest in the control group were statistically significant on the form recall, contextualized form recall, and sentence production immediate posttest and the three test formats combined (*ps* ≤ .001; see Appendix 2I).

3.5.1.2 Delayed Posttest

Participants in the fill-in-the-blanks massed condition had mean scores of 2.63, 2.27, and 1.07 on the form recall, contextualized form recall, and sentence production delayed posttest for a total mean score of 5.97 out of 24. The mean gains from the pretest to the delayed posttest in the fill-in-the-blanks massed condition were statistically significant on each delayed posttest format and the three test formats combined (*ps* ≤ .001; see Appendix 2I). Participants in the fill-in-the-blanks spaced condition had mean scores of 5.10, 5.00, and 2.70 on the form recall, contextualized form recall, and sentence production delayed posttest for a total mean score of 12.80 out of 24. The mean gains from the pretest to the delayed posttest in the fill-in-the-blanks spaced condition were statistically significant on each delayed posttest format and the three test formats combined (*ps* < .001; see Appendix 2I).

Participants in the flashcard massed condition had mean scores of 1.90, 1.87, and 0.93 on the form recall, contextualized form recall, and sentence production delayed posttest for a total mean score of 4.70 out of 24. The mean gains from the pretest to the delayed posttest in the flashcard massed condition were statistically significant on each delayed posttest format and the three test formats combined (*ps* < .001; see Appendix 2I). Participants in the flashcard spaced condition had mean scores of 5.47, 3.00, and 1.83 on the form recall, contextualized form recall, and sentence production delayed posttest for a total mean score of

10.30 out of 24. The mean gains from the pretest to the delayed posttest in the flashcard spaced condition were statistically significant on each delayed posttest format and the three test formats combined (*ps* < .001; see Appendix 2I).

Participants in the control group had mean scores of 0.13, 0.17, and 0.20 on the form recall, contextualized form recall, and sentence production delayed posttest for a total score of 0.5 out of 24. The mean decay in knowledge from the pretest to the delayed posttest for the control group were statistically significant on form recall and sentence production delayed posttest formats and the three test formats combined (*ps* ≤ .05), but not statistically significant on contextualized form recall delayed posttest format (*p* = .08) (see Appendix 2I).

3.5.2 To What Extent do Vocabulary Learning Gains Differ Across the Learning Conditions?

3.5.2.1 Immediate Posttest

When examining the results on the immediate posttest with all three test formats combined (scores out of 24), the results revealed that the four experimental groups contributed to significantly greater gains than the control group (*ps* < .001; see Appendix 2J). The comparisons between the four experimental groups showed that the fill-in-the-blanks spaced condition had statistically greater gains than the flashcard massed ($z = 2.07$, $p = .04$) and spaced conditions ($z = 2.29$, $p = .02$), but the fill-in-the-blanks spaced condition was as effective as the fill-in-the-blanks massed condition ($z = 1.19$, $p = .24$). There was no significant difference between the fill-in-the-blanks massed and flashcard massed conditions ($z = 0.05$, $p = .62$), nor between the fill-in-the-blanks massed and flashcard spaced conditions ($z = 1.14$, $p = .26$). There was also no significant difference between the flashcard massed and spaced conditions ($z = 0.80$, $p = .42$).

Form recall test. When examining the results of the form recall test format (scores out of 8) in the immediate posttest, the comparisons between the four experimental groups showed that the fill-in-the-blanks spaced condition had statistically greater gains than the fill-

in-the-blanks massed and flashcard massed conditions, and that the fill-in-the-blanks massed condition had statistically greater gains than the flashcard massed condition (*ps* ≤ .05). However, there was no significant difference in gains between the fill-in-the-blanks spaced and flashcard spaced conditions ($z = 1.33$, $p = .18$), between the fill-in-the-blanks massed and flashcard spaced conditions ($z = -1.63$, $p = .10$), or between the flashcard massed and spaced conditions ($z = -0.31$, $p = .75$) (see Appendix 2K).

Contextualized form recall test. When examining the results of the contextualized form recall test format (scores out of 8) in the immediate posttest, the comparisons between the four experimental groups showed that the fill-in-the-blanks massed and spaced conditions had statistically greater gains than the flashcard massed and spaced conditions (*ps* < .001), and there was no significant difference between flashcard massed and spaced conditions ($z = 0.44$, $p = .66$) (see Appendix 2K).

Sentence production test. When examining the results of the sentence production test format (scores out of 8) in the immediate posttest, the comparisons between the four experimental groups showed that there was no significant difference in gains between the fill-in-the-blanks massed and spaced, between flashcard massed and spaced, between the fill-in-the-blanks massed and flashcard massed, between the fill-in-the-blanks massed and flashcard spaced, and between flashcard massed and spaced conditions (*p* ≥ .15; see Appendix 2K).

3.5.2.2 Delayed Posttest

When examining the results on the delayed posttest with all three test formats combined (scores out of 24), the results revealed that the four experimental groups contributed to significantly greater gains than the control group (*ps* < .001; see Appendix 2J). The comparisons between the four experimental groups showed that the fill-in-the-blanks spaced condition had statistically greater gains than the fill-in-the-blanks massed ($z = 4.40$, $p < .001$) and flashcard massed conditions ($z = 5.74$, $p < .001$). Similarly, the flashcard spaced condition had statistically greater gains than the fill-in-the-blanks massed ($z = 4.13$, $p < .001$) and flashcard massed conditions ($z = -3.08$, $p < .001$). No significant difference was found

between the fill-in-the-blanks and flashcard spaced conditions ($z = 1.80$, $p = .07$). There was also no significant difference between the fill-in-the-blanks and flashcard massed conditions ($z = 1.07$, $p = .28$).

Form recall test. When examining the results of the form recall test format (scores out of 8) in the delayed posttest, the comparisons between the four experimental groups showed that the flashcard spaced condition had statistically greater gains than the fill-in-the-blanks and flashcard massed conditions (*ps* < .001). The fill-in-the-blanks spaced condition had statistically greater gains than the fill-in-the-blanks and flashcard massed conditions (*ps* < .001). No significant difference was found between the fill-in-the-blanks and flashcard spaced conditions ($z = -0.65$, $p = .52$), nor between the fill-in-the-blanks and flashcard massed conditions ($z = 1.61$, $p = .11$) (see Appendix 2L).

Contextualized form recall test. When examining the results of the contextualized form recall test format (scores out of 8) in the delayed posttest, the comparisons between the four experimental groups showed that the fill-in-the-blanks spaced condition had statistically greater gains than the fill-in-the-blanks massed, flashcard spaced, and flashcard massed conditions (*ps* < .001). The flashcard spaced condition had statistically greater gains than the flashcard massed condition ($z = -2.10$, $p = .04$). However, no significant difference was found between the flashcard and fill-in-the-blanks massed conditions ($z = 0.45$, $p = .44$), nor between the flashcard spaced and fill-in-the-blanks massed conditions ($z = -1.19$, $p = .24$) (see Appendix 2L).

Sentence production test. When examining the results of the sentence production test format (scores out of 8) in the delayed posttest, the comparisons between the four experimental groups showed that the fill-in-the-blanks spaced condition had statistically greater gains than the fill-in-the-blanks and flashcard massed conditions (*ps* < .001). The flashcard spaced condition had statistically greater gains than the flashcard massed condition ($p = .04$). However, there was no significant difference between the fill-in-the-blanks and flashcard spaced conditions ($z = 1.66$, $p = .10$). In addition, no significant difference was found between the fill-in-the-blanks and flashcard massed conditions ($z = 0.35$, $p = .72$), nor

between the fill-in-the-blanks massed and flashcard spaced conditions (z = -1.59, p = .11) (see Appendix 2L).

3.5.3 Does the Correspondence Between Test Format and Vocabulary Learning Condition Affect Gains in Word Knowledge?

When comparing the gains across the three test formats (form recall, contextualized form recall, sentence production) in the immediate and delayed posttests, the correspondence between test format and learning condition affected gains in word knowledge for fill-in-the-blanks but not flashcards (see Appendix 2M). Note that here we combined massed and spaced learning for each activity (fill-in-the-blanks and flashcards).

3.5.3.1 Form Recall Test

When examining the results of the form recall test format (scores out of 8) in the immediate posttest, the fill-in-the-blanks condition had a mean score of 7.08 and the flashcard condition had a mean score of 7.12, with no statistically significant difference between the gains (z = 0.19, p = .85). In the delayed posttest, the fill-in-the-blanks condition had a mean score of 3.87 and the flashcard condition had a mean score of 3.68, with no statistically significant difference between the gains (z = -0.42, p = .68).

3.5.3.2 Contextualized Form Recall Test

When examining the results of the contextualized recall test format (scores out of 8) in the immediate posttest, the fill-in-the-blanks condition had a mean score of 6.83 and the flashcard condition had a mean score of 5.40. The analyses revealed that fill-in-the-blanks contributed to statistically greater gains than flashcards (z = -4.88, p < .001). In the delayed

Table 2 Descriptive statistics for the three tests on the immediate and delayed posttests

		Form recall test				Contextualized form recall test				Sentence production test			
				95% CI				95% CI				95% CI	
		M	*SD*	Lower	Upper	*M*	*SD*	Lower	Upper	*M*	*SD*	Lower	Upper
Control (n= 30)	Immediate Posttest	1.33	1.03	0.97	1.67	0.67	0.71	0.43	0.93	0.70	1.09	0.33	1.13
	Delayed Posttest	0.13	0.35	0.03	0.27	0.17	0.53	0.00	0.37	0.20	0.48	0.07	0.40
Flashcard Massed (n=30)	Immediate Posttest	7.07	0.69	6.83	7.30	5.50	1.38	5.00	5.97	3.50	1.89	2.83	4.20
	Delayed Posttest	1.90	1.61	1.37	2.50	1.87	1.59	1.30	2.40	0.93	1.29	0.53	1.40
FlashcardSpaced (n=30)	Immediate Posttest	7.17	1.60	6.53	7.63	5.30	2.10	4.47	5.97	2.80	1.81	2.17	3.50
	Delayed Posttest	5.47	2.22	4.67	6.20	3.00	2.41	2.13	3.90	1.83	1.93	1.17	2.57
Fill-in Massed (n=30)	Immediate Posttest	6.57	1.19	6.10	7.00	6.83	1.44	6.30	7.33	3.10	1.83	2.43	3.73
	Delayed Posttest	2.63	1.85	2.03	3.27	2.27	2.32	1.53	3.23	1.07	1.74	0.50	1.73
Fill-in Spaced (n=30)	Immediate Posttest	7.60	0.72	7.33	7.83	6.83	1.12	6.40	7.20	2.97	0.89	2.67	3.27
	Delayed Posttest	5.10	2.19	4.30	5.77	5.00	2.64	4.00	5.83	2.70	2.09	2.07	3.47

posttest, the fill-in-the-blanks condition had a mean score of 3.63 and the flashcard condition had a mean score of 2.43, with significantly greater gains made through learning with fill-in-the-blanks than flashcards (z = -2.61, p = .01).

3.5.3.3 Sentence Production Test

When examining the results of the sentence production test format (scores out of 8) in the immediate posttest, the fill-in-the-blanks condition had a mean score of 3.03 and the flashcard condition had a mean score of 3.15. No statistically significant difference was found between the two learning conditions (z = -1.65, p = .10). In the delayed posttest, the fill-in-the-blanks condition had a mean score of 1.88 and the flashcard condition had a mean score of 1.38, with significantly greater gains made through learning with fill-in-the-blanks than flashcards (z = -0.51, p = .61).

3.5.4 To What Extent Does Feedback Timing Affect Vocabulary Learning in Fill-in-the-blank and Flashcard Activities?

Table 3 summarizes the immediate and delayed posttest results as a function of feedback timing. In this study, feedback timing was a within-participants variable. Immediate feedback (feedback given after completing each item) was applied to half the target items (24 items), and the remaining items were placed under a delayed feedback condition (feedback given after completion of all 24 items). Half the items (12 items) in the immediate feedback condition were tested in the immediate posttest, and the other 12 items in the immediate feedback condition were tested in the delayed posttest. Half the items (12 items) in the delayed feedback condition were tested in the immediate posttest, and the other 12 items in the delayed feedback condition were tested in the delayed posttest (see Method).

3.5.4.1 Immediate Posttest

Participants in the fill-in-the-blanks massed condition had mean scores of 7.93 and 7.90 out of 12 with immediate feedback and delayed feedback on the immediate posttest with all three test formats combined. Participants in the fill-in-the-blanks spaced condition had mean scores of 8.83 and 8.50 with immediate feedback and delayed feedback on the immediate posttest with all three test formats combined. Participants in the flashcard massed condition had mean scores of 8.10 and 7.98 with immediate feedback and delayed feedback on the immediate posttest with all three test formats combined. Participants in the flashcard spaced condition had mean scores of 7.87 and 7.43 with immediate feedback and delayed feedback on the immediate posttest with all three test formats combined.

The logistic results revealed that feedback timing did not affect vocabulary learning gains in the immediate posttest (z = -0.07, p = .95; see Appendix 2N). When examining the effect of feedback timing in each learning condition (massed and spaced conditions for each activity were combined), the effect of feedback timing was not significant with fill-in-the-blanks (z = -1.11, p = .27) nor with flashcards (z = -0.04, p = .97). There was no significant interaction between feedback timing and spacing type in the immediate posttest results ($ps \geq .69$). Overall, given that feedback timing did not significantly affect vocabulary learning in the fill-in-the-blanks or flashcard conditions and that there was no interaction between feedback timing and spacing type, it might be reasonable to assume that feedback timing had no effect on immediate posttest results irrespective of learning condition or spacing type.

3.5.4.2 Delayed Posttest

Participants in the fill-in-the-blanks massed condition had mean scores of 3.10 and 2.87 out of 12 with immediate feedback and delayed feedback on the delayed posttest with all three test formats combined. Participants in the fill-in-the-blanks spaced condition had

mean scores of 6.47 and 6.33 with immediate feedback and delayed feedback on the delayed posttest with all three test formats combined. Participants in the flashcard massed condition had mean scores of 2.67 and 2.03 with immediate feedback and delayed feedback on the delayed posttest with all three test formats combined. Participants in the flashcard spaced condition had mean scores of 5.27 and 5.00 with immediate feedback and delayed feedback on the delayed posttest with all three test formats combined.

The logistic results revealed that feedback timing did not affect vocabulary learning gains in the delayed posttest ($z = 0.34$, $p = .73$; see Appendix 2N). When examining the effect of feedback timing in each learning condition (massed and spaced conditions for each activity were combined), the effect of feedback timing was not significant with fill-in-the-blanks ($z = -0.51$, $p = .61$) nor with flashcards ($z = 0.08$, $p = .94$). There was no significant interaction between feedback timing and spacing type in the delayed posttest ($ps \geq .36$). Overall, given that feedback timing did not significantly affect vocabulary learning in the fill-in-the-blanks or flashcard conditions and that there was no significant interaction between feedback timing and spacing type, it might be reasonable to assume that feedback timing had no effect on delayed posttest results irrespective of learning condition or spacing type.

Table 3 Descriptive statistics for feedback timing

	Immediate posttest				Delayed posttest			
	Immediate feedback		Delayed feedback		Immediate feedback		Delayed feedback	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Flashcard Massed (*N* = 30)	8.10	1.81	7.97	1.73	2.67	2.01	2.03	1.96
Flashcard Spaced (*N* = 30)	7.87	2.32	7.43	2.78	5.27	2.45	5.00	2.79

Total (*N* = 60)	7.98	2.06	7.70	2.31	3.97	2.58	3.52	2.82
Fill-in-the-blanks Massed (*N* = 30)	7.93	2.70	7.90	2.58	3.10	3.00	2.87	2.80
Fill-in-the-blanks Spaced (*N* = 30)	8.83	1.39	8.50	1.20	6.47	3.17	6.33	2.92
Total (*N* = 60)	8.38	2.18	8.20	2.02	4.78	3.50	4.60	3.33

Note. The maximum score is 12 for each cell.

3.6 Discussion

The first research question examined the extent to which vocabulary is learned through the fill-in-the-blanks and flashcards using different types of spacing. The results showed that the sizes of the gains in the four experimental groups were very large on the immediate posttest (*d* = 6.34 and 13.14, 95% CI [5.10, 15.55] for the fill-in-the-blanks massed and spaced; *d* = 7.83 and 4.68, 95% CI [3.70, 9.32] for the flashcards massed and spaced) and delayed posttest (*d* = 1.56 and 3.25, 95% CI [0.98, 4.02] for the fill-in-the-blanks massed and spaced; *d* = 1.90 and 2.91, 95% CI [1.29, 3.64] for the flashcards massed and spaced; see Appendix 2I). These findings contrast previously observed effects of vocabulary learning activities (e.g., Webb, Yanagisawa, & Uchihara, 2020, for a review). A meta-analysis conducted by Webb et al. (2020) examined the extent to which L2 vocabulary is learned from the most frequently used vocabulary learning activities. Webb et al. found mean effect sizes (i.e., effect size of proportion of the target words learned) of 0.18 (95% CI [-0.35, 0.72]) and 0.66 (95% CI [0.50, 0.81]) for fill-in-the-blank and flashcard activities on form recall immediate posttests and 0.18 (95% CI [-0.65, 1.01]) and 0.32 (95% CI [0.15, 0.48]) on form recall delayed posttests (measured 4-14 days after engaging in activities). In Webb et al.'s meta-analysis, very small effects for fill-in-the-blanks activities were observed and they were statistically unstable (the confidence intervals passed zero). The findings of the current study, however, found large gains with fill-in-the-

blanks, suggesting that spacing may increase vocabulary learning through the fill-in-the-blanks activity. Webb et al. (2020) also observed a greater effect of flashcards on vocabulary learning than fill-in-the-blanks. Our findings, however, showed that fill-in-the-blanks was as effective or more effective than flashcards in spaced conditions, suggesting that both activities may be affected similarly by spacing. Given the abundant evidence for spaced practice effects with flashcards in L2 vocabulary learning (e.g., Karpicke & Bauernschmidt, 2011; Kim & Webb, in press, for a review; Nakata & Suzuki, 2019), our findings are important because they provide evidence that spaced practice may contribute to vocabulary learning in other ways apart from flashcards.

The second research question compared vocabulary learning gains across the four learning conditions. When comparing massed and spaced conditions in each activity, there were no significant differences between fill-in-the-blanks massed and spaced conditions nor between flashcard massed and spaced conditions for initial learning. However, spaced conditions contributed to significantly greater gains than massed conditions for retention with both activities (d = 1.24, 95% CI [0.69, 1.79] for the fill-in-the-blanks, d = 1.30, 95% CI [0.74, 1.85] for the flashcards; see Appendix 2J). These results indicate that spacing had the same effect across both activities. This is pedagogically valuable because L2 classroom textbooks tend to present target words within units rather than across units. This may represent more condensed and massed presentations of target words for learning rather than spaced presentations, which reduces the potential for vocabulary learning gains. When comparing vocabulary learning gains across activities, the results indicate that the fill-in-the-blanks spaced condition contributed to significantly greater gains than the flashcard spaced condition on initial learning, but these two conditions were similarly effective for retention. These findings suggest that different vocabulary learning activities may be affected similarly by spacing. There are many vocabulary learning conditions (e.g., 23 different activities described by Webb & Nation, 2017; 118 activities profiled by Morgan & Rinvolucri, 2004). Although there have been many studies investigating the effects of spaced practice in paired-associate learning (e.g., flashcards), there is almost no research in relation to any other activities. This is problematic because earlier studies showed that spaced practice has medium-to-large effects on vocabulary learning and

retention (e.g., Cepeda et al., 2006; Donovan & Radosevich, 1999; Kim & Webb, 2022). Further research investigating different vocabulary learning activities might help to promote improved materials design.

Taken together, the findings of the current study suggest that we should consider the value of spaced practice beyond flashcards. Flashcards is a common and efficient activity, but it represents a relatively small aspect of language learning primarily focused on vocabulary learning. Spaced practice research has been useful, but the degree to which it is meaningful might have been constrained by the lack of research beyond flashcards. If spacing does have a positive effect on other vocabulary learning conditions, L2 classroom materials in which words are often learned in more condensed and massed presentations within the units of textbooks may be less than optimal and reduce the potential for vocabulary learning gains.

In answer to the third research question, the results showed that while fill-in-the-blanks contributed to significantly greater learning gains than flashcards on the contextualized form recall posttests, there were no significant differences found between the two activities on the form recall posttests. These findings suggest that the correspondence between learning condition and test format affected gains in word knowledge for fill-in-the-blanks but not flashcards. This is surprising because many earlier studies revealing learning and testing correspondence effects have involved the decontextualized recall format (e.g., Barcroft, 2004; Griffin & Harley, 1996; Mondria & Wiersma, 2004) used in flashcard conditions. Learning and testing correspondence effects are explained by transfer appropriate processing theory (Morris, Bransford, & Franks, 1977), which suggests that memory performance is enhanced when the processes engaged during learning match testing. The current study expands on earlier studies by comparing decontextualized vocabulary learning (flashcards) with contextualized vocabulary learning (fill-in-the-blanks) with assessment occurring in decontextualized form recall (corresponding to flashcards), contextualized form recall (corresponding to fill-in-the-blanks), and a neutral (sentence production) format. Our findings indicate that there was a greater transfer appropriate processing effect through contextualized vocabulary learning than decontextualized vocabulary learning. The reason why transfer appropriate processing

was found with fill-in-the-blanks but not flashcards may be related to the overlap between the psychological conditions that contribute to learning within the activities and the test formats. The psychological conditions that contribute to learning include retrieval and varied encounters (encountering a word in different contexts) in fill-in-the-blanks, but only retrieval in flashcards (Webb & Nation, 2017). Thus, the inclusion of retrieval in both fill-in-the-blanks and flashcards may have had a positive impact on form recall test performance for both conditions. However, the inclusion of varied encounters in only fill-in-the-blanks may have a positive impact on the contextualized form recall for that learning condition.

The last research question looked at the extent to which feedback timing affects vocabulary learning through fill-in-the-blanks and flashcards. In the current study, feedback timing did not affect vocabulary learning in either learning condition. Although immediate feedback showed higher scores than delayed feedback in both activities on the immediate and delayed posttests, the differences were not statistically significant. These results are not consistent with the earlier studies. Kim and Webb (2022) meta-analyzed earlier L2 studies to examine the effects of spaced practice and found large effects of immediate feedback (g = 1.04, 95% CI [0.59, 1.49]) and delayed feedback (g = 0.64~2.34, 95% CI [0.15, 3.04]) in L2 vocabulary learning on delayed posttests (a delay of 1 day or greater following the treatment), although the effect of immediate feedback was smaller than that of delayed feedback.

The difference in results for feedback timing between this and earlier studies may be due to methodological differences. For example, Guo (2021) found greater effects for delayed feedback in a classroom-based study with paper-and-pencil tasks, in contrast to the current study which was a computer-based laboratory study. Delayed feedback provided in classroom-based settings with paper-and-pencil tasks and computer-based setting may lead to different recall rates, because paper-and-pencil tasks may provide students with chances to look over all of their responses. This might not be the case with computer-based delayed feedback. Second, different materials (Kulik & Kulik, 1988, for a review) may account for the inconsistent results. While Guo (2021) involved vocabulary learning from marginal glosses and post-reading activities, the current study involved learning from fill-in-the-

blanks and flashcards. The difference in findings between studies indicates that the effects of feedback timing may depend on the types of learning task. Third, as for operationalization of immediate and delayed feedback (i.e., feedback timing differs between studies, e.g., Metcalfe, Kornell, & Finn, 2009), Guo (2021) included delayed feedback provided 2-3 days after a test, whereas the current study provided delayed feedback after the completion of 24 items (half of the target words). Thus, in Guo's study feedback may function as another learning opportunity, rather than error correction. Although Guo conducted a classroom-based study, providing feedback 2-3 days after an activity may not be typical in classrooms. Providing feedback after completion of all responses in the current study may have greater ecologically validity; in classroom textbooks a fill-in-the-blank activity tends to consist of three to eight questions allowing teachers and students to correct or check answers within a relatively short period of classroom time.

If we consider the current results to be ecologically valid, it may be useful to consider feedback timing based on the learning conditions. Immediate feedback may be more useful with flashcards, because it is easier to manipulate than delayed feedback. Delayed feedback may be more suitable for more typical paper and pencil activities such as fill-in-the-blanks that involves multiple questions. Since the current study showed learning condition to be an important factor in L2 vocabulary learning ($ps \leq .05$; see Appendix 2N) rather than feedback timing, other learning conditions (e.g., sentence writing, multiple-choice) may lead to different results regarding the impact of feedback timing. There are many vocabulary learning activities, and there would be value in further exploring the effects of feedback timing with other vocabulary learning conditions.

3.7 Conclusion

The current study indicates that spacing may increase vocabulary learning gains through fill-in-the-blanks in the same way as flashcards. The effects of spaced practice were greater for fill-in-the-blanks than flashcards immediately after the treatment and spaced practice was more effective than massed practice for both activities two weeks after the

treatment. Many empirical studies have revealed the benefits of spaced practice for paired-associate learning tasks (e.g., Bahrick, 1979; Karpicke & Bauernschmidt, 2011; Kim & Webb, in press, for a review). However, there is almost no research in relation to any of other learning activities. Inside and outside the classrooms, teachers and students use many activities for word learning, and textbooks include a variety of activities to promote vocabulary learning. The findings of the current study provide evidence that the benefits of spaced practice may apply to other deliberate vocabulary learning conditions. With respect to instructional practice, these findings may be informative for teachers, students, and material designers. Spaced practice may provide a means to increase vocabulary learning both inside and outside the classroom. Materials could be designed to include exercises in which target words are learned in a more spaced sequence both between and within the units of textbooks, to increase the potential for vocabulary learning gains.

The current study also indicates that contextualized vocabulary learning (fill-in-the-blanks) may lead to greater gains across test formats than decontextualized vocabulary learning (flashcards). The findings may be important for both teachers and students regarding learning and testing words.

3.8 Note

1 The data of 13 participants who knew some of the target words (*faucet*, *stammer*, and *boar*) on the pretest was excluded to ensure that all participants had no prior knowledge of target items. At pretest, therefore, all the participants scored on all of form recall (0/48), contextualized form recall (0/48), and sentence production (0/48) pretests and a total average score of 0/48 on the three test formats.

3.9 References

Baddeley, A. (1999). *Human memory: Theory and practice* (Revised ed.). UK: Psychology Press.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*(3), 296−308. http://doi.org/10.1037/0096-3445.108.3.296

Barcroft, J. (2004). Effects of sentence writing in second language lexical acquisition. *Second Language Research, 20*(4), 303−334. http://doi.org/10.1191/0267658304sr233oa

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*(1), 35−56. http://doi.org/10.1111/j.1467-9922.2007.00398.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using Ime4. *Journal of Statistical Software, 67*(1), 1−48. http://doi.org/10.18637/jss.v067.i01

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185−205). Cambridge, MA: The MIT Press.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research 74*(4), 245−248. . http://doi.org/10.1080/00220671.1981.10885317

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46,* 904−911. https://doi.org/10.3758/s13428-013-0403-5

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273−281. http://doi.org/10.1037/1.76-898X.13.4.273

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

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

Cohen, J. (1988). *Statistical power analysis for the Behavioral Science* (2nd ed). Lawrence Erlbaum Associates Publishers.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*(5), 795–805. https://doi.org/10.1037/0021-9010.84.5.795

Griffin, G. F., & Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics, 17,* 443−460.

Guo, L (2021). Effects of the initial test interval and feedback timing on L2 vocabulary retention. *The Language Learning Journal, 49*(3), 382−398. http://doi.org/10.1080/09571736.2018.1551416

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.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250−1257. http://doi.org/10.1037/a0023436

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*(2), 151−162. http://doi.org/10.1016/j.jml.2006.09.004

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning.* http://doi.org/10.1111/lang.12479

Kulik, J. A., & Kulik, C. –L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58,* 79–97. http://doi.org/10.3102/00346543058001079

Lenth, R. V. (2016). Least-squares means: The r package lsmeans. *Journal of Statistical Software, 69,* 1−33. http://doi.org/10.18637/jss.v.069.i01

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37,* 1077−1087. http://doi.org/10.3758/MC.37.8.1077

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive and receptive + productive L2 vocabulary learning: What difference does it make? In B. Laufer (Ed.),

Vocabulary in a second language: Selection, acquisition, and testing (pp.79−100). Amsterdam, the Netherlands: Benjamins.

Morgan, J., & Rinvolucri, M. (2004). *Vocabulary*. Oxford: Oxford University Press.

Morris, D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519−533. http://doi.org/10.1016/S0022-5371(77)80016-9

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. http://doi.org/10.1017/S0272263114000825

Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research, 37*(2), 233−260. http://doi.org/10.1177/0267658320927764

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. http://doi.org/10.1017/S0272263118000219

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition, 38*(3), 523−552. https://doi.org/10.1017/S0272263115000236

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved May 3, 2019, from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle.

Pigada, M., & Schmitt, N., (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language, 18*(1), 1−28.

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

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437−447. http://doi.org/10.1016/j.jml.2009.01.004

Roediger, H. L., & Guynn, M. J. (1996). Retrieval processes. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 197−236). San Diego: Academic Press.

Roediger, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 23−47). New York: Psychology Press.

Roediger, H. L., & March, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(5), 1155−1159. http://doi.org/10.1037/0278-7393.31.5.1155

Rogers, J., & Cheung, A. (2020). Input spacing and the learning of L2 vocabulary in a classroom context. Language Teaching Research, 24, 616–641. http://doi.org/10.1177/1362168818805251

Rogers, J., & Cheung, A. (2021). Does it matter when you review? Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition, 43(5), 1138–1156.* http://doi.org/10.1017/S0272263120000236

Royer, J. M. (1973). Memory effects for test like events during acquisition of foreign language vocabulary. *Psychological Reports, 32*, 195−198. http://doi.org/10.2466/pr0.1973.32.1.195

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. http://doi.org/10.1111/lang.12236

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

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46−65. http://doi.org/10.1093/applin/aml048

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. RELC Journal, 30(3), 360−376. http://doi.org/10.1177/0033688209343854

Webb, S., & Chang, A. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, *19*, 667−686. http://doi.org/10.1177/1362168814559800

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

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated vocabulary levels test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, *168*(1), 34−70. http://doi.org/10.1075/itl.168.1.02w

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, *104*(4), 715−738. http://doi.org/10.1111/modl.12671

Chapter 4: When Should We Learn Second Language Words in Sentence Production Activities? Comparing Spaced and Massed Learning

4.1 Introduction

There are many activities designed for learning words in the classroom. For example, teachers can use flashcards to help students memorize target words and their meanings, fill-in-the-blanks for writing appropriate target words in given sentences, matching activities for connecting target words to their meanings, and sentence production tasks for using target words in sentences. Many different vocabulary learning activities have been developed and discussed in the research literature (118 activities for word learning, Morgan & Rinvolucri, 2004; 23 approaches for developing vocabulary knowledge, Webb & Nation, 2017), with studies demonstrating that vocabulary can be explicitly learned through activities with the size of gains varying across different activities (e.g., Webb, Yanagisawa, & Uchihara, 2020, for a review).

The ways in which activities are performed provide learners with certain learning conditions that may contribute to vocabulary learning (Webb & Nation, 2017). Two variables that have been found to affect the learning and retention of words are frequency of encounters and retrieval practice. Research examining the frequency of encounters tends to indicate that the more that words are studied or encountered, the more likely learning is to occur (e.g., Nakata, 2017; Webb, 2007). Research investigating retrieval practice (i.e., testing the knowledge studied) has shown that compared to restudy (i.e., learning words and then restudying them), retrieval practice more enhances retention than restudy (e.g., Barcroft, 2007; Royer, 1973). Furthermore, several studies have demonstrated that *spaced practice* (i.e., providing an interval between repeated practice) improves learning and retention relative to *massed practice*, in which repeated practice occurs in immediate succession without any intervals (e.g., Kim & Webb, 2022a, for a review).

Despite positive effects of spaced practice on L2 vocabulary learning, the abundant extant research has centered mainly on paired-associate word learning (e.g., flashcards), in

which learners are asked to recall target words and their first language [L1] meanings (Kim & Webb, 2022a). Although paired-associate learning is an effective method of learning words (Webb et al., 2020), it is one of many approaches to deliberately learning words. Earlier studies have shown positive effects of spaced practice on L2 vocabulary learning and retention (Kim & Webb, 2022a), but the effects cannot yet be generalized to other deliberate vocabulary learning conditions.

The present study attempted to examine whether spaced practice had a similar effect on the learning of L2 vocabulary in sentence production and flashcard activities. Sentence production is one of the most frequently used word-focused activities for L2 vocabulary learning (Webb et al., 2020). Determining the extent to which spacing may contribute to vocabulary learning in different learning activities such as sentence production has pedagogical value because it may help teachers and learners to optimize vocabulary learning gains. Comparing the gains in vocabulary learning through sentence production and flashcards may provide some indication of the degree to which spacing effects found through paired-associate learning conditions may be generalized to other vocabulary learning activities.

4.2 Background

4.2.1 Spaced Practice and L2 Vocabulary Learning

Spaced practice can refer to two types of spaced learning conditions. First, repeated learning sessions may be separated by time as in a 3-day interval schedule within which encounters with a given item are spaced by intervals of 3 days. Second, spaced practice may be separated by the number of words studied between encounters with each target word. For example, a group of three words implies a spacing of two words between opportunities to retrieve a particular word, such as apple, banana, orange, apple, banana, orange, apple, banana, orange. In contrast, massed practice involves learning words in a sequence with no items in between occurrences (e.g., apple, apple, apple, banana, banana, banana, orange, orange, orange). *Spacing effect* refers to the phenomenon where spaced

practice promotes better learning and more enhanced retention than massed practice (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

A great deal of research has demonstrated that spaced practice leads to greater learning and retention of L2 vocabulary in comparison to massed practice (e.g., Karpicke & Bauernschmidt, 2011; Nakata & Suzuki, 2019). Kim and Webb (2022a) meta-analyzed forty-eight experiments from 37 studies investigating the effects of spaced practice on L2 learning and found that spaced practice contributed to greater gains than massed practice in L2 vocabulary learning (measured by immediate posttests, g = 0.76, 95% CI [0.26, 1.25]) and retention (measured one-day or greater following the treatment, g = 1.15, 95% CI [0.81, 1.49]). However, most spaced practice research on L2 vocabulary learning was limited to paired-associate learning (e.g., flashcards). Although spaced practice research has been useful, the degree to which it is meaningful might have been constrained by a lack of research beyond flashcards.

There have been three studies investigating the effects of spaced practice using other deliberate vocabulary learning activities (Kim & Webb, 2022b; Bloom & Shuell, 1981; Rogers & Cheung, 2021). Bloom and Shuell (1981) compared the effects of massed and spaced practice on French word learning through multiple-choice, fill-in-the-blanks, and form recall (from L1 to L2) activities. In the massed condition, participants completed all three activities on one day. In the spaced condition, the participants completed the three activities over three days. Bloom and Shuell found no significant differences between the conditions on an immediate posttest (mean percentage: massed = 80.6%, spaced = 84.25%), but a significant effect of spaced practice was found on a 4-day delayed posttest (massed = 55.75%, spaced = 75.20%). These results indicate that other deliberate vocabulary learning activities may be positively affected by spacing. However, Bloom and Shuell involved word learning in each of the three different activities, and did not investigate the benefits of spacing pertaining to specific activities. Rogers and Cheung (2021) compared two different types of spaced practice (1-day versus 8-day) on English word learning with Chinese primary school students. In the practice and testing (4-week delayed posttest) sessions, crossword puzzles were used. Half the words were subjected to a shorter spaced (1-day) condition, and the other half were in a longer spaced (8-day)

condition. Rogers and Cheung (2021) found no significant difference between the two conditions. Recently, Kim and Webb (2022b) compared fill-in-the-blanks and flashcard activities to examine the effect of massed and spaced practice on vocabulary learning. Kim and Webb found that the effects of spaced practice were greater for fill-in-the-blanks than flashcards on an immediate posttest (d = 0.61, 95% CI [0.09, 1.12]) and that spaced practice was more effective than massed practice for both activities on a 2-week delayed posttest (d = 1.24, 95% CI [0.69, 1.79] for the fill-in-the-blanks, d = 1.30, 95% CI [0.74, 1.85] for the flashcards). These results suggest that spacing may lead to vocabulary learning in other ways apart from flashcards. However, Kim and Webb's study is the only one to investigate the extent to which spacing affects vocabulary learning in other ways relative to flashcards, and therefore further research is warranted.

4.2.2 Sentence Production Activities and L2 Vocabulary Learning

Many studies have demonstrated the effects of sentence production activities for L2 vocabulary learning. Webb (2005, experiment 2) found benefits of sentence writing in promoting both receptive and productive knowledge of vocabulary (41%-54% on immediate posttests). Keating (2008) also found effects of sentence writing for the learning of receptive and productive knowledge of words (64% and 46% gains on L2-L1 word recall immediate and delayed posttests; 42% and 21% gains on L1-L2 sentence translation immediate and delayed posttests). In addition, Javanbakht (2011) found positive effects of sentence writing in developing knowledge of form-meaning connection (84% and 66% gains on immediate and delayed posttests). Pichette, De Serres, and Lafontaine (2012) also found that writing words in sentences facilitates learning and retention of words (25% and 11% gains on immediate and delayed posttests). Barcroft (2004, experiment 1) compared the effects of sentence writing with word-picture pair learning and found that word-picture pairs led to greater learning than sentence writing. However, Barcroft (2004, experiment 1) did not control the number of repetitions (writing one sentence versus viewing word-picture pairs 4 times) and the words were assessed in a form recall test (i.e., a picture was given, and the learners were asked to write a target word corresponding to the picture),

which matched the word-picture pair learning condition. Folse (2006) compared the effects of learning L2 vocabulary from one fill-in-the-blank exercise, three fill-in-the-blank exercises, and a sentence writing exercise. Folse found that three fill-in-the-blank exercises significantly outperformed single fill-in-the-blank and sentence writing exercises but no difference between one fill-in-the-blank exercise and writing sentences was found.

Taken together, the results of Javanbakht (2011), Keating (2008), Pichette et al. (2012), and Webb (2005 experiment 2) have shown positive effects for sentence writing on vocabulary learning in comparison to reading conditions. However, Barcroft (2004, experiment 1) and Folse (2006) failed to find greater effects of sentence writing when compared to other word-focused activities. However, to accurately gauge the effects of sentence production in relation to other word-focused activities, studies of L2 vocabulary learning need to control frequency and learning and testing correspondence (Nation & Webb, 2011), because research has consistently shown that frequency of encounters (e.g., Nakata, 2017; Webb, 2007), and learning and testing correspondence (Morris, Bransford, & Franks, 1977) affect gains. Furthermore, given large effects of spacing in repeated practice with paired-associate learning conditions (e.g., flashcards) on the learning and retention of L2 vocabulary, comparing the gains in spaced vocabulary learning through sentence production and flashcards may provide evidence that the degree to which spacing effects observed through paired-associate learning conditions may be generalized to other vocabulary learning conditions. Determining whether spacing affects sentence production activities would also be pedagogically valuable because it may help to reveal new ways to increase vocabulary learning.

4.2.3 Effects of Feedback Timing on L2 Vocabulary Learning

The timing of feedback—whether feedback is provided immediately or with a delay—has been found to affect learning and memory in cognitive psychology research (e.g., Butler, Karpicke, & Roediger, 2007; Metcalfe, Kornell, & Finn, 2009; Roediger & March, 2005). Immediate feedback provided after each retrieval (i.e., testing) can reduce the negative effects of tasks that could possibly provide learners with erroneous information (i.e.,

multiple-choice questions) (Roediger & March, 2005). When delayed feedback is provided, incorrect responses might be forgotten over time, and the correct responses may be easily learned (Butler et al., 2007; Metcalfe et al., 2009).

Research examining whether feedback timing moderates the effects of spaced practice found large effects of immediate feedback (g = 1.04, 95% CI [0.59, 1.49]) and delayed feedback (g = 0.64~2.34, 95% CI [0.15, 3.04]) for the retention of L2 vocabulary (measured one day or greater after the treatment; Kim & Webb, 2022a). However, the studies included in Kim and Webb's meta-analysis only involved paired-associate word learning (e.g., flashcards, word list).

The only study to directly examine the effectiveness of feedback timing (immediate and delayed) in spaced (1-day versus 3-day) L2 vocabulary learning apart from paired-associate learning conditions was conducted by Guo (2021). Guo investigated how feedback timing affects vocabulary learning from textbook glosses, followed by post-reading activities. Guo found that delayed feedback contributed to significantly greater gains than immediate feedback in the retention of L2 vocabulary (measured by a 5-day delayed posttest). The degree to which feedback timing affects vocabulary learning in other learning conditions remains to be explored.

4.3 The Current Study

This study investigated how spacing in sentence production and flashcard activities affected L2 vocabulary learning and retention. Participants completed either sentence production or flashcard activities under one of two (massed and spaced) practice schedules. Posttest formats were matched to the learning conditions (sentence production and form recall), and a neutral assessment task (fill-in-the-blanks) was also included. This study also investigated whether feedback timing (feedback provided immediately or with a delay) moderated vocabulary learning in the two activities. This following research questions were addressed.

1. To what extent is vocabulary learned through sentence production and flashcard activities using different types of spacing?

2. To what extent does vocabulary learning differ across the learning conditions?

3. Does the correspondence between vocabulary learning condition and test format affect gains in word knowledge?

4. To what extent does feedback timing affect vocabulary learning in sentence production and flashcard activities?

4.4 Method

4.4.1 Participants

150 Korean students (76 male and 74 female, M_{age} = 21.2, *SD* = 1.1) from six universities in South Korea participated in this study. Nine participants were English majors and the remaining participants were majoring in other academic disciplines. All participants had studied English for a minimum of eight years, and took the Vocabulary Levels Test (VLT; Webb, Sasao, & Ballance, 2017) before the experiment. The average vocabulary scores (Standard deviation) of the participants were 98% (4.2) at the 1000 word level, 93% (13.4) at the 2000 word level, 88% (15.2) at the 3000 word level, 80% (16.3) at the 4000 word level, and 76% (18.6) at the 5000 word level. The participants were randomly assigned to one of 5 treatment groups, one control and four experimental (two learning conditions x two spacing schedules) groups. A no treatment control group (*n* = 30) was included in this study in order to obtain a more accurate assessment of spacing effects in the learning conditions. The four experimental groups were massed (*n* = 30) and spaced (*n* = 30) sentence production, and massed (*n* = 30) and spaced (*n* = 30) flashcards.

4.4.2 Target Items

The target items were forty-eight low frequency English words from the most frequent 8,000 to 16,000 word families in Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) lists. Low frequency English words were selected to increase the likelihood that the participants were not familiar with the items (see Appendix B). The target items included 28 nouns and 20 verbs, following the 6:4 ratio of nouns to verbs in natural text (Webb, 2005).

4.4.3 Instructional Treatment

PsychoPy was used to present the treatments and collect data on learning and test performance: present target words in the presentation phase, the exercises (sentence production, flashcards) in the practice phase, and the tests (pretest, immediate, and delayed).

4.4.3.1 Presentation Phase

The target words were presented onscreen in a dictionary format. Each target word was presented in bold font followed by its part of speech and Korean definition for 10 seconds, and participants listened to its pronunciation once (see screenshot on the left, Figure 1). A sentence example including the target word underlined and the Korean translation of the sentence were then presented for 15 seconds (see screenshot on the right, Figure 1). Sentences used in the presentation phase were taken from the COCA (https://www.english-corpora.org/coca/), and lower frequency words were replaced with words from the most frequent 1,000 and 2,000 word families to increase the likelihood that all of the sentences would be easily understood (see Appendix C).

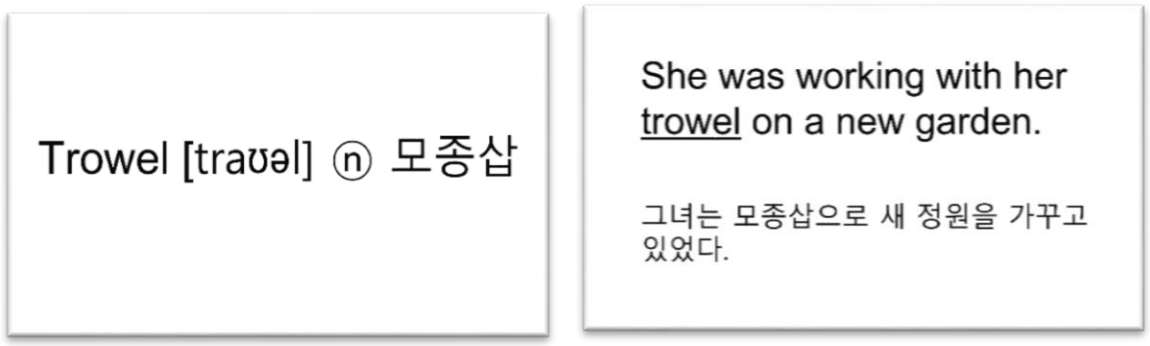


FIGURE 1 Screenshots of target word presentation during the treatment

4.4.3.2 Practice Phase: Flashcards Group

The majority of earlier studies examining the effects of spaced practice employed paired-associate learning tasks (e.g., flashcards) as the learning condition and demonstrated positive effects (e.g., Kim & Webb, 2022). Therefore, flashcards was used for comparison to sentence production. As shown in the left panel in Figure 2, the participants were first given the following instructions in their L1, "Type the English target word corresponding to the Korean definition provided on the screen". The participants were then presented with a screen, which was accompanied by a Korean definition (e.g., 모종삽). The participants were given as much as time they needed to type the English target word corresponding to the Korean definition provided on the screen. After pressing enter, the target word and its Korean definition were provided as feedback for 10 seconds through immediate feedback for half the target items (24 items), while the other 24 items were placed under the delayed feedback condition (feedback provided after completion of all 24 items). The type of feedback applied to the items was counterbalanced between participants (see Appendix E). Earlier studies have suggested that time on task is a considerable factor in completing the effects of different learning conditions (e.g., Webb, 2005), but it was important that the participants in this study complete each learning

condition over the amount of time required for each exercise. The amount of time taken to compete each exercise was, therefore, collected and included as covariate.

4.4.3.3 Practice Phase: Sentence Production Group

The sentence production exercise had two features. First, similar to the flashcard practice phase, productive retrieval (i.e., retrieving the L2 word form), was required. Second, the target words were expected to be used in context. As shown in the right panel in Figure 2, the participants were first instructed in their L1 to "Make a sentence in English that includes the target word corresponding to the Korean definition provided on the screen". The participants were then presented with the Korean definition (e.g., 모종삽). The participants were given as much as time they needed to type each sentence. After pressing enter, the target word and its Korean definition were provided as feedback for 10 seconds. Immediate feedback (feedback provided immediately after each response) was applied to half the target items (24 items), and the remaining items were placed under a delayed feedback condition (feedback provided after completion of all 24 items). These items were counterbalanced between participants to ensure that any differences in learning within the two feedback conditions were not due to word-related variables (see Appendix E).

4.4.4 Spacing Schedules

Participants were randomly assigned to one of the two (massed and spaced) practice schedules. After the presentation phase, participants in the massed condition learned the target words in the practice phase by retrieving the words five times using the assigned exercise (sentence production or flashcards) within one session. Participants in the spaced condition retrieved the words over five sessions (one retrieval attempt in each session) from Monday to Friday (one session per day). The only difference between the two practice schedules in the assigned exercise was the interval between retrieval attempts for target words.

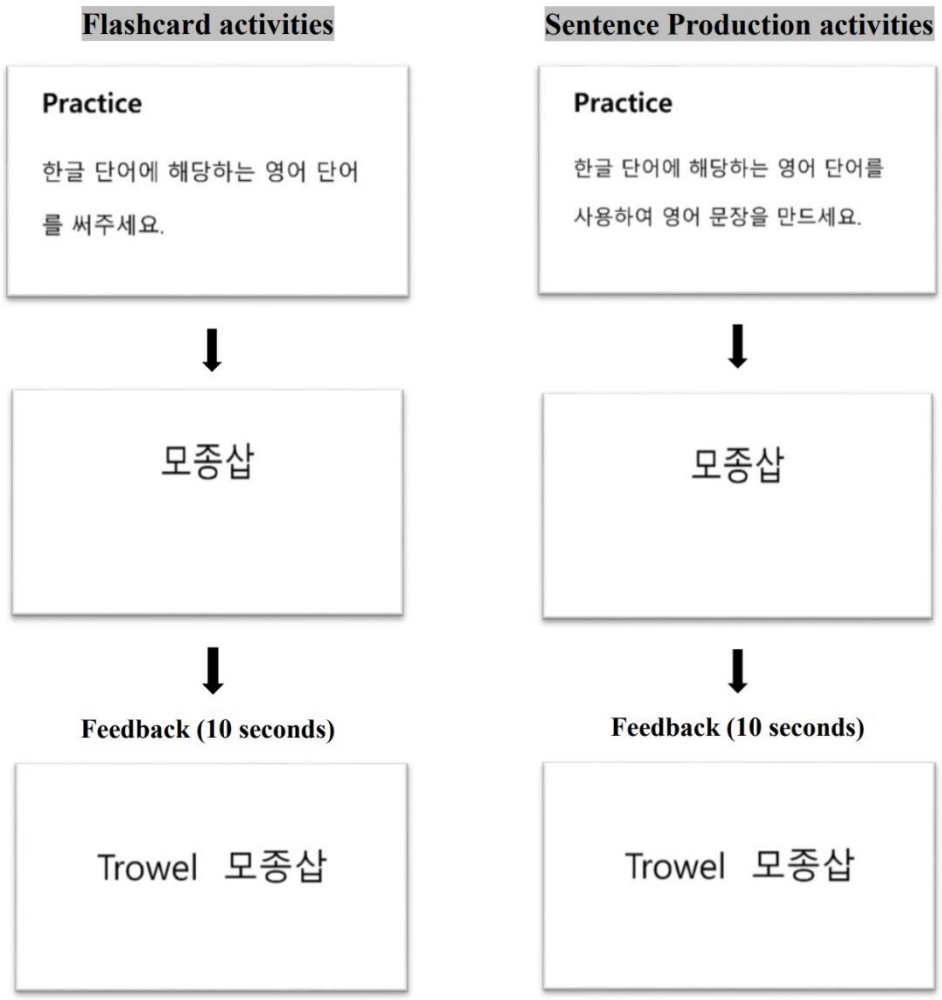


FIGURE 2 A sample display of flashcard and sentence production activities in practice phase (for target word *trowel*)

4.4.5 Measurement

Form recall, sentence production, and contextualized form recall tests were administered in this study. Form recall and sentence production test formats matched the learning conditions: the form recall test corresponded with the flashcards, and sentence production

test corresponded with the sentence production activity. The contextualized form recall test format corresponded with a fill-in-the-blanks task (i.e., participants were asked to type the appropriate English target word at the bottom of the screen to complete the gap in the provided sentence; see Appendix F). The contextualized form recall test format was selected as a neutral test that did not favor either of the two learning conditions.

In the form recall test, participants were asked to type the English target words corresponding to the Korean definitions provided on the screen. In the sentence production test, participants were asked to make a sentence including the target word that corresponded with the Korean definition provided on the screen. In the contextualized form recall test, a sentence with a blank was provided and the participants were asked to type the appropriate target word to complete the blank. The sentences in the contextualized form recall test were taken from the COCA (https://www.english-corpora.org/coca/) with lower frequency words replaced with words from the most frequent 1,000 and 2,000 word families to increase the chances that the sentences would be understood (see Appendix F). Participants could take as much time as they needed to type responses on all tests.

4.4.5.1 Pretest

Before the treatment, knowledge of the 48 target items was assessed on each of the form recall, sentence production, and contextualized form recall tests in that order as the pretest. When a participant wrote a synonym rather than a target word (e.g., writing *sing* rather than *croon*), the participant was asked if he or she knew any other words that corresponded with the Korean definition provided to ensure that the recall tests did not underestimate knowledge.

4.4.5.2 Posttest

The posttests were administered immediately and 2 weeks after the treatment. The 48 target items were divided into six sets of eight items; half the items (24 items, three sets of

8 items) were tested on the immediate posttest, and the remaining items (24 items, three sets of 8 items) were tested on the delayed posttest (see Appendix E). Suzuki (2017) mentioned that when a study administered an outcome test more than once, the first test can be regarded as a learning session. In this study, the test items on the immediate and delayed posttests were different (24 items for each posttest) so that there were an equal number of learning sessions (frequency of retrieval practice), and there could not be a learning effect from taking the immediate posttest. Each set of 8 items was randomly assigned to each of the three test formats (form recall, sentence production, and contextualized form recall) in each of the posttests (immediate and delayed posttests) so that knowledge of a target item was only evaluated on a single test. This ensured that there could not be a learning effect from taking the different test formats on the immediate and delayed posttests. The order of the test items was randomized between tests for each participant to reduce the possibility of an order effect (see Appendix G). The delayed posttest was administered with no prior notice.

4.4.6 Procedure

Prior to the treatment, participants took the pretest and VLT in the initial session. All participants were informed about the research procedure and completed a consent form. To ensure that participants had no knowledge of target items, only data from participants who scored 0 on the pretest were included in this study. The four experimental groups underwent three phases: presentation, practice, and testing. The control group only undertook the presentation and testing phases.

In the presentation phase, the control and four experimental groups learned the 48 target items. In the practice phase, the participants in the experimental groups practiced the items (i.e., retrieving the items they had learned in the presentation phase) 5 times in their assigned exercises.

After the presentation and/or practice phase, all participants answered 10 2-digit additions (e.g., 82+39 = ?) as filler items, which were used as recency buffers during the

treatment (e.g., Karpicke & Roediger, 2007). The control group took the immediate posttest after the presentation phase. The experimental groups took the immediate posttest after the last practice. All groups took the delayed posttest two weeks after the immediate posttest. Table 1 summarizes the procedure.

Table 1 Procedures of the current study

	All groups (one control and four experimental)		
Pre-meeting	Pretest, VLT		
	Control (1 group)	Massed (2 groups)	Spaced (2 groups)
	Presentation phase Learning words **Testing phase** Immediate posttests (at the end of the session)	**Presentation phase** Learning words **Practice phase** *5* sessions of assigned exercise (within a session) **Testing phase** Immediate posttests (at the end of the session)	**Presentation phase** Learning words **Practice phase** 5 sessions of assigned exercise (1-day interval) **Testing phase** Immediate posttests (at the end of the last session)
Two weeks after the treatment	**Testing phase** 2-week delayed posttests	**Testing phase** 2-week delayed posttests	**Testing phase** 2-week delayed posttests

Note. 2 groups include sentence production and flashcard conditions

4.4.7 Scoring

Scoring for all three tests included the following criteria: First, the target words needed to be spelled correctly. The reason for this was that one aspect of the participants' task in all experimental groups was to produce the correct written forms of target words. Second, the target words needed to include the correct grammatical function. This was because the responses required in the contextualized form recall test did not require inflected or derived forms of target words, and the use of target words in sentence production tasks involves producing words with their correct grammatical forms. Therefore, responses such as *trowels* in "My mother uses a ___ to do some flower gardening" in the contextualized form recall test, and if "She has many *trowel* for garden", were marked incorrect. Third, the responses produced on the sentence production test needed to be complete sentences that were comprehensible. Responses such as "A *trowel* is", and "Flower *trowel*" and "*trowel* to have this up" were marked incorrect.

The responses on the form recall and contextualized form recall pre- and posttests were first scored as 1 (correct) or 0 (incorrect) by the PsychoPy software based on answers (target words with correct spellings) compiled by the authors. Responses that were marked incorrect by the PsychoPy were manually checked by the authors. Responses produced on the sentence production test were manually scored by the authors.

4.4.8 Data Analysis

The immediate and delayed posttest scores were analyzed separately using a logistic mixed-effects model fit by maximum likelihood with binomial logit functions through the lme4 software package in R 4.1.1 (Bates, Mächler, Bolker, & Walker, 2015). The dependent variable was a binary response (correct/incorrect). Fixed-effect predictors were learning condition (control, sentence production, flashcards) and spacing type (massed, spaced). Learning condition and spacing type were conducted at subject level, and test formats (form recall, sentence production, contextualized form recall) and feedback timing (immediate and delayed) were conducted at item level. The initial model included

intercept-only random models with learning condition and spacing type as fixed effects and time on task as a covariate, and interactions among the fixed effects and one covariate (time on task) were added to the initial model. The alpha level of statistical significance was set at less than .05. To compare the differences between groups, post hoc tests were conducted using the R package (Ismeans; Lenth, 2016). Effect sizes of the comparisons between groups were calculated and interpreted based on Plonsky and Oswald's (2014) benchmark (small: 0.40 ≤ Cohen's d < 0.70; medium: 0.70 ≤ d < 1.00; large: 1.00 < d for between-participants contrasts).

4.5 Results

None of the participants in this study demonstrated prior knowledge of any of the target words on pretests (form recall, sentence production, and contextualized form recall). Cronbach's alpha was .86 or higher (.86-.88) for all dependent measures (form recall, sentence production, and contextualized form recall) on the immediate and delayed posttests, indicating good reliability. Table 2 presents means (M), standard deviations (SD), and 95% confidence intervals (CIs) on both immediate and delayed posttest in all five conditions.

4.5.1 Vocabulary Learning Through Sentence Production and Flashcard Activities Using Different Types of Spacing

4.5.1.1 Immediate Posttest

Massed sentence production had mean scores of 6.50, 3.70, and 5.70 on the form recall, sentence production, and contextualized form recall immediate posttests, respectively, for a total mean score of 15.90 out of 24 on the three test formats combined. The mean gains from the pretest to the immediate posttest for massed sentence production were statistically significant on each immediate posttest format and the total mean gains on the three test formats combined (ps < .001). Spaced sentence production had mean scores of 7.27, 3.57,

and 5.63 on the form recall, sentence production, and contextualized form recall immediate posttests for a mean score of 16.47 out of 24. The mean gains from the pretest to the immediate posttest for spaced sentence production were statistically significant on each immediate posttest format and the three test formats combined (*ps* < .001) (see Appendix 3H).

Massed flashcards had mean scores of 7.07, 3.50, and 5.50 in the form recall, sentence production, and contextualized form recall immediate posttest for a total mean score of 16.07 out of 24. The mean gains from the pretest to the immediate posttest for massed flashcards were statistically significant on each immediate posttest format and the three test formats combined (*ps* < .001). Spaced flashcards had mean scores of 7.17, 2.80, and 5.30 on the form recall, sentence production, and contextualized form recall immediate posttest for a total mean score of 15.27 out of 24. The mean gains from the pretest to the immediate posttest for spaced flashcards were statistically significant on each immediate posttest format and the three test formats combined (*ps* < .001) (see Appendix 3H).

The no treatment control group had mean scores of 1.33, 0.70, and 0.67 on the form recall, sentence production, and contextualized form recall immediate posttest for a total score of 2.7 out of 24. The mean gains from the pretest to the immediate posttest for the control group were statistically significant on the form recall, contextualized form recall, and sentence production immediate posttest and the three test formats combined (*ps* ≤ .001) (see Appendix 3H).

4.5.1.2 Delayed Posttest

Massed sentence production had mean scores of 1.53, 0.87, and 1.30 on the form recall, sentence production, and contextualized form recall delayed posttest for a total mean score of 3.70 out of 24. The mean gains from the pretest to the delayed posttest for massed sentence production were statistically significant on the each delayed posttest format and the three test formats combined (*ps* ≤ .001). Spaced sentence production had mean scores of 4.67, 2.60, and 3.53 on the form recall, sentence production, and contextualized form

recall delayed posttest for a total score of 10.80 out of 24. The mean gains from the pretest to the delayed posttest for spaced sentence production were statistically significant on each format and the three test formats combined (*ps* < .001) (see Appendix 3H).

Massed flashcards had mean scores of 1.90, 0.93, and 1.87 on the form recall, sentence production, and contextualized form recall delayed posttest for a total mean score of 4.70 out of 24. The mean gains from the pretest to the delayed posttest for massed flashcards were statistically significant on each format and the three test formats combined (*ps* < .001). Spaced flashcards had mean scores of 5.47, 1.83, and 3.00 on the form recall, sentence production, and contextualized form recall delayed posttest for a total mean score of 10.30 out of 24. The mean gains from the pretest to the delayed posttest in the flashcard spaced condition were statistically significant on each format and the three test formats combined (*ps* < .001) (see Appendix 3H).

The control group had mean scores of 0.13, 0.20, and 0.17 on the form recall, sentence production, and contextualized form recall delayed posttest for a total score of 0.50 out of 24. The mean decay in knowledge from the pretest to the delayed posttest for the control group were statistically significant on form recall and sentence production delayed posttest formats and the three test formats combined (*ps* ≤ .05), but not statistically significant on contextualized form recall delayed posttest format ($z = 1.74$, $p = .08$) (see Appendix 3H).

4.5.2 Comparisons of Vocabulary Learning Gains Across the Learning Conditions

4.5.2.1 Immediate Posttest

Results for the three test formats combined (scores out of 24) revealed that the four experimental groups contributed to significantly greater gains than the control group on the immediate posttest (*ps* < .001; see Appendix 3I). The comparisons between the four experimental groups showed no significant differences between the four experimental groups. Spaced sentence production was as effective as massed sentence production (*z* =

0.53, p = .59), massed flashcards (z = 0.43, p = .67), and spaced flashcards (z = 1.04, p = .30). Similarly, massed sentence production was as effective as massed flashcards (z = -0.19, p = .85), and spaced flashcards (z = 0.57, p = .57). No significant difference was found between the massed and spaced flashcard conditions (z = -0.80, p = .42). The results of the individual test format (scores out of 8) in the immediate posttest are reported in Appendix 3J.

4.5.2.2 Delayed Posttest

Results revealed that the four experimental groups contributed to significantly greater gains than the control group on the delayed posttest with three test formats combined (scores out of 24) (ps < .001; see Appendix 3I). The comparisons between the four experimental groups showed that spaced sentence production had statistically greater gains than massed sentence production (z = 5.42, p < .001) and massed flashcards (z = 4.89, p < .001), but spaced sentence production was as effective as spaced flashcards (z = -0.39, p = .70). Spaced flashcards had statistically greater gains than massed sentence production (z = -5.13, p < .001) and massed flashcards (z = 4.57, p < .001). There was no significant difference between massed sentence production and massed flashcards (z = -1.07, p = .29). The results of the individual test format (scores out of 8) in the delayed posttest are reported in the Appendix 3J.

4.5.3 Vocabulary Learning Gains and Test Formats

To answer this question the scores of the massed and spaced groups were combined for each of the two learning conditions.

Table 2 Descriptive statistics for the three tests on the immediate and delayed posttests

		Form recall test				Contextualized form recall test				Sentence production test			
				95% CI				95% CI				95% CI	
		M	*SD*	Lower	Upper	*M*	*SD*	Lower	Upper	*M*	*SD*	Lower	Upper
Control (n = 30)	Immediate Posttest	1.33	1.03	0.97	1.67	0.67	0.71	0.43	0.93	0.70	1.09	0.33	1.13
	Delayed Posttest	0.13	0.35	0.03	0.27	0.17	0.53	0.00	0.37	0.20	0.48	0.07	0.40
Flashcard Massed (n = 30)	Immediate Posttest	7.07	0.69	6.83	7.30	5.50	1.38	5.00	5.97	3.50	1.89	2.83	4.20
	Delayed Posttest	1.90	1.61	1.37	2.50	1.87	1.59	1.30	2.40	0.93	1.29	0.53	1.40
Flashcard Spaced (n = 30)	Immediate Posttest	7.17	1.60	6.53	7.63	5.30	2.10	4.47	5.97	2.80	1.81	2.17	3.50
	Delayed Posttest	5.47	2.22	4.67	6.20	3.00	2.41	2.13	3.90	1.83	1.93	1.17	2.57
Sentence production Massed (n = 30)	Immediate Posttest	6.50	1.17	6.07	6.87	5.70	1.69	5.10	6.23	3.70	2.22	2.93	4.47
	Delayed Posttest	1.53	1.93	0.87	2.30	1.30	1.62	0.73	1.90	0.87	1.07	0.53	1.27
Sentence productionSpaced (n = 30)	Immediate Posttest	7.27	1.48	6.70	7.73	5.63	2.11	4.97	6.37	3.57	1.89	2.83	4.27
	Delayed Posttest	4.67	2.07	3.93	5.40	3.53	2.19	2.73	4.33	2.60	1.71	1.93	3.27

4.5.3.1 Form Recall Test

Sentence production and flashcards had mean scores of 6.88 and 7.12 on the form recall test format immediate posttest (scores out of 8), respectively, with no statistically significant difference between the gains (p = .31). In the delayed posttest, sentence production and flashcards produced mean scores of 3.10 and 3.68, respectively, with no statistically significant difference between the gains (p = .22).

4.5.3.2 Sentence Production Test

Sentence production and flashcards produced mean scores of 3.63 and 3.15 on the sentence production immediate posttest (scores out of 8), respectively, with no statistically significant difference between the gains (p = .18). In the delayed posttest, sentence production and flashcards led to mean scores of 1.73 and 1.38, respectively, with no statistically significant difference between the gains (p = .26).

4.5.3.3 Contextualized Form Recall Test

Sentence production and flashcards contributed to mean scores of 5.67 and 5.40 on the contextualized form recall immediate posttest (scores out of 8), respectively, with no statistically significant difference between the gains (p = .42). In the delayed posttest, sentence production and flashcards had mean scores of 2.42 and 2.43, respectively, with no statistically significant difference between the gains (p = .98).

Taken together, when comparing the gains across the three test formats in the immediate and delayed posttests, the correspondence between learning condition and test format did not affect vocabulary learning gains (see Appendix 3K).

4.5.4 Effects of Feedback Timing on Vocabulary Learning

Note that immediate feedback was applied to half the target items (24 items), and the remaining 24 items were placed under a delayed feedback condition. Half the items (12 items) in each feedback condition were tested in the immediate posttest, and the other 12 items in each feedback condition were tested in the delayed posttest. Table 3 summarizes the immediate and delayed posttest results for feedback timing.

4.5.4.1 Immediate Posttest

Massed sentence production with immediate feedback and delayed feedback contributed to mean scores of 8.50 and 7.40 out of 12 on the immediate posttest with three test formats combined. Spaced sentence production with immediate feedback and delayed feedback led to mean scores of 8.43 and 8.03 on the immediate posttest. Massed flashcards with immediate and delayed feedback contributed to mean scores of 8.10 and 7.98 on the immediate posttest. Spaced flashcards with immediate and delayed feedback led to mean scores of 7.87 and 7.43 on the immediate posttest.

When collapsing sentence production and flashcards conditions to see how feedback timing affected learning, the logistic model results showed that feedback timing significantly affected vocabulary learning gains in the immediate posttest (z = -2.40, p = .02; see Appendix 3L). In each learning condition, although immediate feedback showed higher mean scores than delayed feedback in both sentence production and flashcard conditions, the effect of feedback timing was significant with sentence production (z = -2.04, p = .05) but not with flashcards (z = -0.04, p = .97) (see Appendix 3M). In the sentence production conditions, immediate feedback led to significantly greater gains than delayed feedback, with a small effect size (d = 0.36, 95% CI [0.00, 0.72]). There was no significant interaction between feedback timing and spacing type for sentence production (z = 1.40, p = .16) nor for the flashcards (z = -0.40, p = .69).

4.5.4.2 Delayed Posttest

Massed sentence production with immediate and delayed feedback led to mean scores of 1.63 and 2.23 out of 12 on the three delayed test formats combined, respectively. Spaced sentence production with immediate and delayed feedback contributed to mean scores of 5.50 and 5.30 on the delayed posttest. Massed flashcards with immediate and delayed feedback led to mean scores of 2.67 and 2.03 on the delayed posttest. Spaced flashcard condition had mean scores of 5.27 and 5.00 with immediate feedback and delayed feedback on the delayed posttest.

The logistic results showed that feedback timing did not significantly affect vocabulary learning gains in the delayed posttest (z = -0.29, p = .77; see Appendix 3L). In each learning condition, although immediate feedback showed higher mean scores than delayed feedback, the effect of feedback timing was not significant with sentence production (z = 1.41, p = .16) nor with flashcards (z = 0.08, p = .94). There was no significant interaction between feedback timing and spacing type for the sentence production (z = -1.35, p = .18) nor for flashcards (z = -0.91, p = .36).

Table 3 Descriptive statistics for feedback timing

	Immediate posttest				Delayed posttest			
	Immediate feedback		Delayed feedback		Immediate feedback		Delayed feedback	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Flashcard Massed (*N* = 30)	8.10	1.81	7.97	1.73	2.67	2.01	2.03	1.96
Flashcard Spaced (*N* = 30)	7.87	2.32	7.43	2.78	5.27	2.45	5.00	2.79
Total (*N* = 60)	7.98	2.06	7.70	2.31	3.97	2.58	3.52	2.82

Sentence production Massed (N = 30)	8.50	2.24	7.40	2.18	1.63	2.28	2.23	2.05
Sentence production Spaced (N = 30)	8.43	1.18	8.03	2.66	5.50	2.84	5.30	2.63
Total (N = 60)	8.47	1.71	7.72	2.43	3.57	3.21	3.77	2.80

Note. The maximum score is 12 for each cell.

4.6 Discussion

The present study investigated the extent to which vocabulary is learned through sentence production and flashcards using different practice schedules. The results revealed that gains for the four experimental groups were very large on the immediate posttest (d = 5.64 and 5.47, 95% CI [4.37, 6.76] for massed and spaced sentence production; d = 7.83 and 4.68, 95% CI [3.70, 9.32] for massed and spaced flashcards) and delayed posttest (d = 1.41 and 3.05, 95% CI [0.84, 3.79] for massed and spaced sentence production; d = 1.90 and 2.91, 95% CI [1.29, 3.64] for massed and spaced flashcards; see Appendix S6). These results contrast previously observed effects of vocabulary learning activities (e.g., Webb et al., 2020). Webb et al.'s (2020) meta-analysis examined the extent to which L2 vocabulary is learned from the most commonly used word-focused activities and found mean effect sizes (effect size of proportion of the target words learned) of 0.37 (95% CI [0.10, 0.62]) and 0.66 (95% CI [0.50, 0.81]) for writing and flashcard activities on form recall immediate posttests and 0.18 (95% CI [-0.15, 0.52]) and 0.32 (95% CI [0.15, 0.48]) on form recall delayed posttests (measured 4-14 days after engaging in activities). Webb et al. observed small effects for writing on both immediate (effect size = 0.37) and delayed (effect size = 0.18) posttests, and the effects were statistically unstable in the delayed posttest (the CI passed zero). However, the current study found large effect sizes with sentence production. Webb et al. (2020) also found larger effect sizes for flashcards (0.66 and 0.32 on immediate and delayed posttests) in vocabulary learning than writing (0.37

and 0.18 on immediate and delayed posttests). The current study, however, found greater effects of sentence production on vocabulary learning than flashcards in the spaced condition. This suggests that spacing may increase vocabulary learning through sentence production activities. Kim and Webb (2022b) examined the effects of spacing on L2 vocabulary learning through a fill-in-the-blanks exercise and also found positive effects of spacing with fill-in-the-blanks. Together these findings provide evidence that spacing may contribute to vocabulary learning in a variety of word-focused activities.

When comparing massed and spaced conditions in each activity, there were no statistically significant differences between massed and spaced sentence production conditions nor between massed and spaced flashcard conditions for initial learning of L2 vocabulary. However, spaced conditions had statistically greater gains than massed conditions for retention for both sentence production (d = 1.61, 95% CI [1.03, 2.19]) and flashcards (d = 1.30, 95% CI [0.74, 1.85]). When comparing learning gains across activities, spaced sentence production was as effective as spaced flashcards for both initial learning and retention of L2 vocabulary. This provides more evidence that different vocabulary learning activities may be affected similarly by spacing. There are many classroom activities for vocabulary learning (Webb & Nation, 2017; Morgan & Rinvolucri, 2004). Further research investigating the extent to which spacing promotes vocabulary learning in different activities might help to optimize the teaching and learning of L2 vocabulary.

A secondary aim of the current study was to investigate learning and testing correspondence effects on vocabulary learning through sentence production and flashcards. There were no significant differences found between the two activities on any of the three (form recall, sentence production, and contextualized form recall) posttests. These findings contrast earlier findings which revealed learning and testing correspondence effects (e.g., Barcroft, 2004). Barcroft (2004) compared the effects of sentence writing with word-picture pair learning and found that word-picture pairs led to better performance than sentence writing in L2 vocabulary learning and retention measured by a form recall immediate and 2-day delayed posttests (i.e., a picture was given and learners were asked to recall the word corresponding to the picture). The superiority of

the word-picture pair learning over the sentence writing on the form recall tests was supported by transfer appropriate processing theory (Morris et al., 1977), which suggests that better retention occurs when processes engaged during learning match testing. The current study expands on earlier studies by including two assessments (form recall and sentence production tests) that are sensitive to the gains made in individual learning conditions, as well as assessment (contextualized form recall test) that did not favor either condition. The current study found no transfer appropriate processing effect in either sentence production or flashcard activities. The reason why no transfer appropriate processing effect was found may be that both sentence production and flashcard activities include retrieval (Webb & Nation, 2017), which may have had a positive impact on the form recall test for both activities. Another reason may be that although in sentence production the processes engaged during learning matched the sentence production test, learners in both sentence production and flashcards may rely on their prior knowledge to help them to create sentences. This suggests that although processes during learning match testing, no additional support to gain knowledge of how to use vocabulary during learning may not allow learners to successfully use new words in sentences.

The final research question examined the extent to which feedback timing affects vocabulary learning. The results showed that feedback timing significantly affected vocabulary learning in sentence production but not in flashcards. These findings are not consistent with earlier studies. Kim and Webb (2022a) meta-analyzed the effects of feedback timing on L2 spaced vocabulary learning through paired-associate learning and found large effects of immediate feedback (g = 1.04, 95% CI [0.59, 1.49]) and delayed feedback (g = 0.64~2.34, 95% CI [0.15, 3.04]) on delayed posttests (measured 1 day or greater after the treatment). This suggests that the effects of feedback timing may differ across vocabulary learning conditions. Guo (2021) examined the effects of feedback timing on L2 vocabulary learning through glosses and post-reading activities and found that delayed feedback had greater gains than immediate feedback. Kim and Webb (2022b) examined the role of feedback timing on L2 vocabulary learning through fill-in-the-blanks and found no significant difference between immediate and delayed feedback. In the current study, however, immediate feedback in sentence production led to significantly

higher scores than delayed feedback in the immediate posttest (d = 0.36, 95% CI [0.00, 0.72]).

The difference in results for feedback timing between the current study and earlier studies may also be related to methodological differences such as different timings of feedback between studies (e.g., Metcalfe et al., 2009) or experimental settings (e.g., Butler et al., 2007). For example, in Guo (2021), delayed feedback was provided 2-3 days after testing, while in the present study, and Kim and Webb's (2022b), delayed feedback was provided after the completion of half of the target words (24 words) in the learning conditions. Furthermore, Guo (2021) conducted a classroom-based study with pencil-and-paper tasks, while Kim and Webb's (2022b) and the current study are computer-based studies. Classroom-based studies with paper-and-pencil tasks may not precisely control feedback timing because learners may look over all of their responses, leading to differing feedback timings among participants. Because there are many vocabulary learning exercises, there is a need for further research investigating the effects of feedback timing across activities. It should also be noted that in the current study the target word and its Korean definition were provided as feedback in both learning conditions. While this is typical for flashcards, there are many ways in which feedback could be provided for sentence production. It would also be useful to conduct further research investigating whether the positive effect of spacing can be replicated when different types of feedback are provided.

4.7 Conclusion

The current study examined how spacing in sentence production and flashcards affected L2 vocabulary learning and retention. The results showed that spacing had similar effects for both activities on L2 vocabulary learning. Although sentence production is a frequently used activity for learning, vocabulary learning gains from sentence production activities are typically small (Webb et al. 2020). The findings of the current study suggest that spaced practice provides a means to increase the potential for vocabulary learning through sentence production activities. Thus, spacing sentence production activities in course

books and classroom-based learning programs may help to increase vocabulary knowledge.

4.8 References

Barcroft, J. (2004). Effects of sentence writing in second language lexical acquisition. *Second Language Research, 20*(4), 303−334. http://doi.org/10.1191/0267658304sr233oa

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*(1), 35−56. http://doi.org/10.1111/j.1467-9922.2007.00398.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using Ime4. *Journal of Statistical Software, 67*(1), 1−48. http://doi.org/10.18637/jss.v067.i01

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research 74*(4), 245−248. . http://doi.org/10.1080/00220671.1981.10885317

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*, 273−281. http://doi.org/10.1037/1.76-898X.13.4.273

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

Folse, K. S. (2006). The effects of type of written exercise on L2 vocabulary retention. *TESOL Quarterly, 40*(2), 273−293. http://doi.org/10.2307/40264523

Guo, L (2021). Effects of the initial test interval and feedback timing on L2 vocabulary retention. *The Language Learning Journal, 49*(3), 382−398. http://doi.org/10.1080/09571736.2018.1551416

Javanbakht, Z. O. (2011). The impact of tasks on male Iranian elementary EFL learners' incidental vocabulary learning. *Language Education in Asia, 2*(1), 28−42. http://doi.org/10.5746/LEiA/11/V2/I1/A03/Javanbakht

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250−1257. http://doi.org/10.1037/a0023436

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. http://doi.org/10.1177/1362168808089922

Kim, S. K., & Webb, S. (2022a). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning.* Early view. http://doi.org/10.1111/lang.12479

Kim, S. K., & Webb, S. (2022b). Does spaced practice have the same effects on different second language vocabulary learning activities? Fill-in-the-blanks versus flashcards. [Manuscript submitted for publication]. Faculty of Education, University of Western Ontario.

Lenth, R. V. (2016). Least-squares means: The r package lsmeans. *Journal of Statistical Software, 69*, 1−33. http://doi.org/10.18637/jss.v.069.i01

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37*, 1077−1087. http://doi.org/10.3758/MC.37.8.1077

Morgan, J., & Rinvolucri, M. (2004). *Vocabulary*. Oxford: Oxford University Press.

Morris, D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519−533. http://doi.org/10.1016/S0022-5371(77)80016-9

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition, 39*(4), 653−679. http://doi.org/10.1017/S0272263116000280

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. http://doi.org/10.1017/S0272263118000219

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved May 3, 2019, from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle.

Pichette, F., De Serres, L., & Lafontaine, M. (2012). Sentence reading and writing for second language vocabulary acquisition. *Applied Linguistics, 33*(1), 66−82. http://doi.org/10.1093/applin/amr037

Roediger, H. L., & March, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(5), 1155−1159. http://doi.org/10.1037/0278-7393.31.5.1155

Rogers, J., & Cheung, A. (2021). Does it matter when you review? Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition, 43(5), 1138–1156.* http://doi.org/10.1017/S0272263120000236

Royer, J. M. (1973). Memory effects for test like events during acquisition of foreign language vocabulary. *Psychological Reports, 32*, 195−198. http://doi.org/10.2466/pr0.1973.32.1.195

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. http://doi.org/10.1111/lang.12236

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

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46−65. http://doi.org/10.1093/applin/aml048

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

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated vocabulary levels test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics, 168*(1), 34−70. http://doi.org/10.1075/itl.168.1.02w

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal, 104*(4), 715−738. http://doi.org/10.1111/modl.1267

Chapter 5: Conclusion

This chapter reviews the results of the three studies presented in this thesis, followed by methodological and pedagogical implications for L2 vocabulary learning. It also presents the limitations of the studies and provides suggestions for future research.

5.1 Review of the Findings

5.1.1 Summary of Study 1

Study 1 (Chapter 2) systematically reviewed 48 experiments from 37 L2 studies of spaced practice to provide a more reliable estimate of its effect on L2 learning. There was also a secondary aim of determining the extent to which spaced practice effects are moderated by different variables (age, learning target, number of sessions, type of practice, activity type, provision of feedback, feedback timing, frequency of practice, and retention interval). The results showed medium-to-large effects of spaced practice for immediate L2 learning (g = 0.58, 95% CI [0.13, 1.00]) and longer-term retention (g = 0.80, 95% CI [0.44, 1.17]) over massed practice. Shorter spacing was as effective as longer spacing for immediate L2 learning, but longer spacing was more effective than shorter spacing for longer-term retention (g = 0.40, 95% CI [0.16, 0.64]). However, there was no significant difference found in learning gains between equal and expanding spacing conditions for L2 learning. Variability in spaced practice effects across studies was explained by several methodological variables such as number of sessions, type of practice, and retention interval. Spaced practice effects on L2 vocabulary learning were more pronounced when spacing was within a single training session than between multiple training sessions. Greater effects of longer spacing on retention were observed when it involved test-restudy trials than when it involved study-only trials. Effects of expanding spacing were greater than equal spacing when the retention interval (i.e., the interval between the last practice and the final test) was longer.

Study 1 showed significant effects of spaced practice on L2 vocabulary, grammar, and pronunciation learning, but most studies investigating spaced practice effects have examined L2 vocabulary learning (k = 33 for vocabulary learning, k = 12 for grammar learning, and k = 4 for pronunciation learning). In the studies of L2 intentional vocabulary learning included in this meta-analysis, the majority of studies involved paired-associate learning tasks (e.g., flashcards) as the activity type. This indicated a need for more research on the effects of spaced practice on L2 vocabulary learning through other activities. Investigating the effects of spaced practice on L2 vocabulary learning with other activities would be pedagogically valuable because it may help teachers and students to optimize vocabulary learning gains.

5.1.2 Summary of Study 2

Study 2 (Chapter 3) examined how spacing in fill-in-the-blanks and flashcard activities affected L2 vocabulary learning and retention. Greater effects of spaced practice were observed for fill-in-the-blanks than flashcards on immediate learning of vocabulary. In addition, spaced practice was more effective than massed practice for both activities in the retention of vocabulary (measured two weeks after the treatment). Regarding learning and testing correspondence effects, results showed that the correspondence between learning condition and test format affected vocabulary learning for fill-in-the-blanks but not flashcards. Fill-in-the-blanks had greater gains than flashcards in the contextualized form recall test but there was no difference found in the gains between fill-in-the-blanks and flashcards in the form recall test. This suggests that contextualized vocabulary learning (fill-in-the-blanks) may contribute to greater learning gains across test formats than decontextualized vocabulary learning (flashcards). Regarding feedback timing, results showed that when feedback was provided did not have an impact on vocabulary learning in either learning condition.

The findings of Study 2 indicated that spaced practice may have positive effects on vocabulary learning in activities apart from flashcards. This reveals that spaced practice

might be more effectively used with other activities to increase L2 vocabulary learning potential.

5.1.3 Summary of Study 3

Study 3 (Chapter 4) examined how spacing effected the learning and retention of L2 vocabulary in sentence production and flashcard activities. The results of Study 3 showed that spacing had a similar effect on both activities. Spaced practice was as effective as massed practice with both activities for immediate learning of vocabulary. Furthermore, spaced practice was more effective than massed practice for retention with both activities. Regarding learning and testing correspondence effects, there was no differences found between the two activities on any of the three (form recall, sentence production, and contextualized form recall) tests. Regarding feedback timing, immediate feedback led to greater gains than delayed feedback in the immediate learning of vocabulary through sentence production activities.

Taken together, the findings of Study 3 indicate that spacing had similar effects on both activities. This suggests that spacing can be used to increase vocabulary learning gains with sentence production activities in the same way as flashcards.

5.2 General Implications

The current research aimed to examine overall effects of spaced practice on L2 learning, followed by investigating whether spacing works with other activities for L2 vocabulary learning. The current research may be able to provide several implications based on the results from the three studies.

5.2.1 Methodological Implications

The studies in this thesis have several implications for researching L2 vocabulary learning. First, there is a need to investigate a greater number of activities to gain a better understanding of the extent to which the effects of spaced practice differ across learning conditions. The majority of earlier L2 spaced practice studies demonstrating positive effects of spacing tended to involve deliberate vocabulary learning through paired-associate learning task (e.g., flashcards, word list) with small number of studies (e.g., Macis, Sonbul, & Alharbi, 2021; Nakata & Elgort, 2021; Serrano & Huang, 2018) also showing positive effects of spaced practice on incidental learning of L2 vocabulary (e.g., learning words through reading or listening). A lack of research beyond deliberate vocabulary learning through paired-associate learning tasks may have constrained the degree to which spaced practice effects are meaningful. Findings of Studies 2 and 3 suggested that spacing may increase vocabulary learning through different learning conditions (fill-in-the-blanks and sentence production activities) in the same way as flashcards but the sizes of effects were different. It is important for researchers to be aware that the effects of spaced practice across learning conditions may vary.

Second, it is important for researchers to control for possible confounding variables affecting the effects of spaced practice when comparing different learning conditions. Earlier empirical studies and reviews have shown that learner-related variables such as learners' aptitude (Suzuki& DeKeyser, 2017) or methodological variables such as task difficulty (Donovan & Radosevich, 1999) and retention interval (interval between last learning session and final test, Cepeda et al., 2006) may affect the contributions of spaced practice. The current findings indicated that the effects of feedback timing may also differ across learning conditions, regardless of whether practice is spaced or not. It is also important for researchers to be aware that immediate and delayed feedback timing may have differing effects on learning and retention across conditions.

It is also important for researchers to consider different types of test formats when comparing different learning conditions to provide a more sensitive and accurate assessment of learning. Many earlier studies comparing learning conditions have used decontextualized recall test formats that corresponded with one of two learning conditions revealing learning and testing correspondence effects (memory performance is enhanced

when the processes engaged during learning match testing, Morris et al., 1977). Findings in this thesis suggested a greater transfer appropriate processing effect through contextualized vocabulary learning than decontextualized vocabulary learning. When contextualized and decontextualized learning conditions are compared, it is important for researchers to consider test formats that are sensitive to learning conditions as well as add another test that does not favour either of the learning conditions to accurately evaluate the effectiveness of learning conditions and interpret the findings of studies appropriately.

5.2.2 Pedagogical Implications

The findings of the three studies in this thesis also have important implications for L2 vocabulary teaching and learning. First, the results of Study 1 indicate that introducing greater spacing of target words between activities may be very important for L2 vocabulary learning outcomes, especially for enhancing retention. This suggests that teachers and students may be able to use spaced practice as a means to increase the potential for vocabulary learning gains inside and outside the classroom. For example, it may be more useful for teachers to revisit taught words across lessons rather than within lessons. Similarly, students should be aware of the value of using a spaced schedule for self-testing to better remember the words that were studied. For example, it may be more useful for students to test studied words every day similar to the spaced conditions in Studies 2 and 3 rather than several times within a day (similar to the massed conditions in these studies). In addition, it would be useful for teachers to schedule activities designed to evaluate students' knowledge of studied words over a course in a manner that includes sufficient spacing to ensure that students have opportunities to more effectively evaluate and further develop their knowledge of these words. Moreover, students should be made aware of the pedagogical value of spaced self-practice (i.e., practicing studied words with activities) at home.

Materials could also be effectively designed to include activities in which target words are learned in a more spaced sequence within **and** between the units of textbooks. When scheduling materials for a course, teachers may be able to introduce activities for

studied words across lessons to have students revisit the words over a course rather than simply within units or lessons. Several activities for studied words may be scheduled as additional practice (e.g., homework, assignment) by teachers. It may also be useful for teachers to encourage students to use a spacing schedule when they make their own study plans.

The results of Study 2 indicate that fill-in-the-blanks (contextualized vocabulary learning) led to greater gains across test formats (form recall and contextualized form recall tests) than flashcards (decontextualized vocabulary learning). This suggests that teachers and students should be aware that practicing studied words through context-based activities may contribute to greater vocabulary learning than practicing studied words and their meanings through flashcards or word lists.

The results of Studies 2 and 3 also indicate that effects of feedback timing may differ across learning conditions. These findings suggest that providing feedback immediately after each response may have important consequences for L2 vocabulary learning outcomes, regardless of spacing schedules (i.e., whether practice is massed or spaced). Teachers may need to consider feedback timing based on learning conditions (i.e., activity type). For example, when teachers use computer-assisted flashcard activities, either immediate or delayed feedback may be provided in flashcard learning. When teachers and students use paper-based flashcards, immediate feedback may be easier to implement manually than delayed feedback. Fill-in-the-blanks typically involve multiple questions (3-8 question items for each fill-in-the-blanks exercise). Either immediate feedback (providing feedback immediately after each question) or delayed feedback (providing feedback after completion of all 3-8 questions) may be used in the fill-in-the-blanks activities. For sentence production activities, teachers may be able to provide feedback immediately after a student produces each sentence including the target word. Since the amount of time to complete sentence production activities may be longer than when they complete flashcards or fill-in-the-blanks, providing feedback immediately after each sentence response may help students fully understand feedback after both successful and unsuccessful retrievals of words as well as both grammatically correct and incorrect sentences, rather than providing delayed feedback (e.g., Butler & Roediger, 2007).

Therefore, teachers may be advised to let students produce sentences including a target words in class and provide feedback immediately after completion of each sentence produced by students.

5.3 Limitations and Future Directions

It is important to note several limitations of the three studies in this thesis. Study 1 (meta-analysis) revealed that many L2 studies of spaced practice had investigated L2 vocabulary learning but that the majority of these studies involved paired-associate learning. Although studies 2 and 3 have shown that fill-in-the-blanks and sentence production activities are also affected positively by spacing, there are many other activities used in L2 classrooms. Therefore, more research investigating effects of spaced practice with different learning conditions is still needed. Studies 2 and 3 involved two different spacing schedules (massed [no interval] and spaced [1-day interval]). Revisiting the words that are learned may be possible within and between units of course books, or across courses. To increase ecological validity, comparing massed practice to longer spaced practice (e.g., massed versus 1-week spaced) or shorter spaced practice with longer spaced practice (e.g., 1-day spaced versus 1-week spaced) would be useful. Given that effects of feedback timing may differ across learning conditions, it would be useful for future studies to examine the effects of feedback timing with other learning activities (e.g., multiple-choice and matching exercises). It should also be noted that in Study 3 the target word and its Korean definition were provided as feedback in both learning conditions (sentence production and flashcards activities). While this is common for flashcards, there are many ways in which feedback could be provided for sentence production. It would be useful to conduct further research investigating whether the positive effect of spacing can be replicated when different types of feedback are provided. It would also be useful for future studies to examine the extent to which other learner-related variables (e.g., aptitude, vocabulary size, language background) moderate the effects of spaced practice in different vocabulary learning conditions. Understanding how learner differences affect learning through spaced practice may help teachers to more effectively select activities or spacing schedules.

5.4 Conclusion

The purpose of this thesis was to investigate the effects of spaced practice in order to optimize L2 vocabulary learning gains. The studies in this thesis highlighted the value of spaced practice, and revealed that its effects can be generalized beyond flashcards. Because there is little research that has attempted to examine the effects of spaced practice on vocabulary learning through other activities, there is much that remains to be explored. Further research investigating the effects of spaced practice with different spacing schedules under different learning conditions may be a useful follow-up to the studies in this thesis.

5.5 References

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research 74*(4), 245−248. . http://doi.org/10.1080/00220671.1981.10885317

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

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

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. https://doi.org/10.1037/0021-9010.84.5.795

Macis, M., Sonbul, S., & Alharbi, R. (2021). The effect of spacing on incidental and deliberate learning of L2 collocations. *System, 103*. Early view. http://doi.org/10.1016/j.system.2021.102649

Morris, D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519−533. http://doi.org/10.1016/S0022-5371(77)80016-9

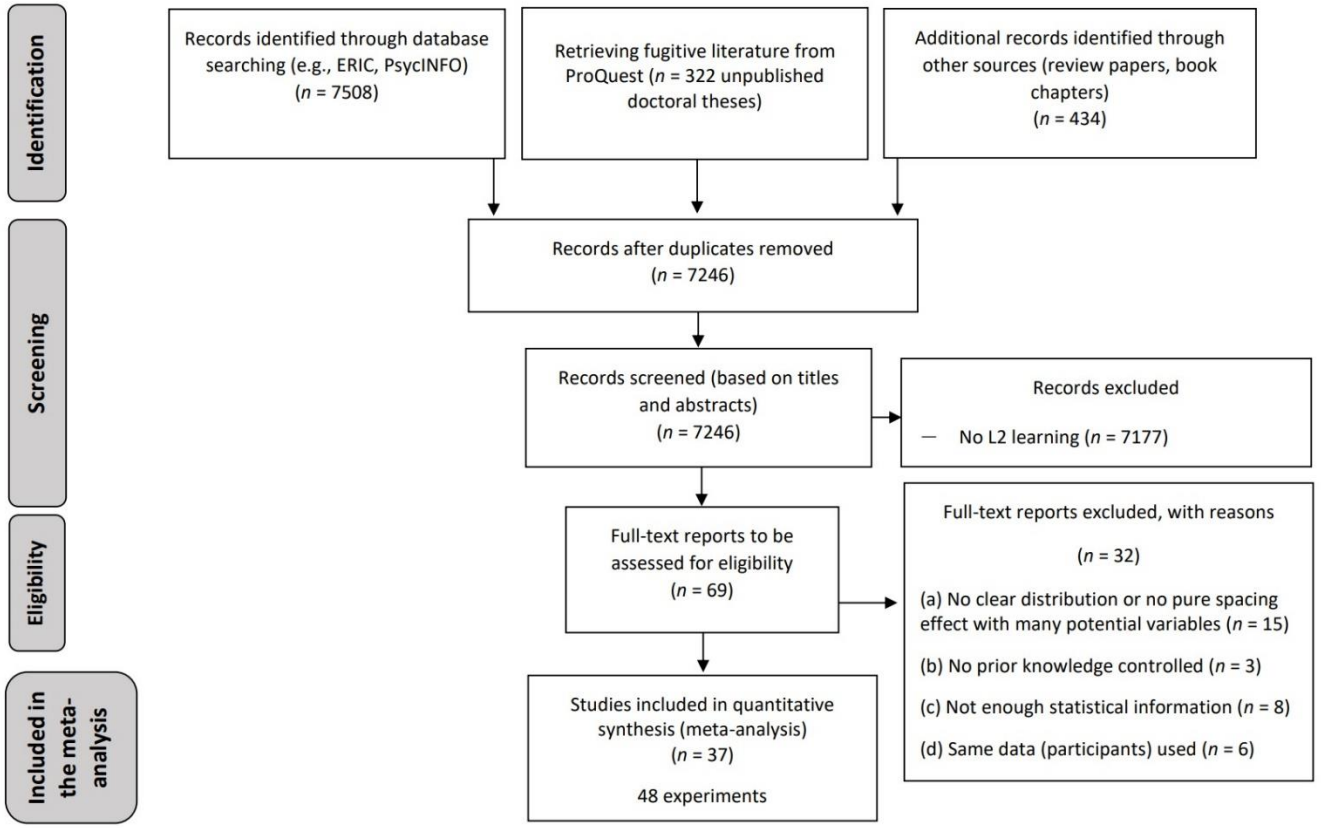
Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, *37*(2), 233−260. http://doi.org/10.1177/0267658320927764

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. http://doi.org/10.1002/tesq.445

Suzuki, Y., & DeKeyser, R. (2017b). Exploratory research on second language practice distribution: An aptitude × treatment interaction. *Applied Psycholinguistics*, *38*(1), 27–56. https://doi.org/10.1017/S0142716416000084

Appendices for Study 1

Appendix S1: PRISMA Flow Diagram



PRISMA flow diagram depicting study inclusion criteria (Moher, Liberati, Tetzlaff, & Altman, 2009) provides the number of included and excluded references and the reason why we excluded some of them. A total of 69 studies were assessed for eligibility, 37 satisfied all criteria (See Eligibility, PRISMA flow diagram). During this stage, 32 studies were excluded, with reasons: (a) no clear distribution or no clear effect of spaced practice with a number of potential variables (e.g., Collins, Halter, Lightbown, & Spada, 1999; Freed, Segalowitz, & Dewey, 2004; Lapkin, Hart, & Harley, 1998; Lightbown & Spada, 1994; Mashhadi, Farvardin, & Mozaffari, 2017; Namaziandost, Nasri, Rahimi Esfahani, & Keshmirshekan, 2019; Namaziandost, Rahimi Esfahani, & Hashemifardnia, 2018; Schuetze & Weimer-Stuckmann, 2011; Serrano, 2011; Serrano & Munoz, 2007) or with different research focus (benefits of CALL vocabulary program, Miles & Kwon, 2008; the effect of distribution of cumulative versus non-cumulative retrieval practice, Nakata, Tada, McLean, & Kim, 2020; comparing different within-session spacing conditions manipulated in relearning sessions with 1-week intersessional interval to examine relearning effect (benefits and costs), Rawson, Vaughn, Walsh, & Dunlosky, 2018; context of learning, Finkbeiner & Nicol, 2003; Schneider, Healy, & Bourne, 1998); (b) not clear if participants' prior knowledge of target items was controlled (Küpper-Tetzel, Erdfelder, & Dickhaeuser, 2014; Lee & Choe, 2014, Experiment 1; Suzuki & Sunada, 2020; see the examples below describing the level of prior knowledge of target structures (or rules) in grammar and pronunciation studies and how inclusion/exclusion criteria were applied to the studies); (c) not enough statistical information to calculate effect sizes (Bahrick, 1979; Bahrick, Bahrick, & Bahrick, 1993; Bahrick & Hall, 2005, Experiment 2; Bahrick & Phelps, 1987; Pyc & Rawson, 2012a, Experiments 1 and 2; Pyc & Rawson, 2012b, Experiment 2; Schneider, Healy, & Bourne, 2002, Experiments 1 and 2; Tsai, 1927, Experiments 2 and 3); and (d) same participants were used (Kanayama & Kasahara, 2017; Li, 2017; Nakata, 2013; Pan, Lovelett, Phun, & Rickard, 2019; Suzuki, 2018, 2019; Suzuki & DeKeyser, 2017b).

Examples describing the level of prior knowledge of target items and how inclusion/exclusion criteria were applied

Study	Learning target	The level of prior knowledge of target items	Inclusion/Exclusion
Bird (2010)	Grammar	A pretest (error correction) was given on the first day of the course. Because no significant difference between groups on the pretests, the pretest mean scores were collapsed across groups.	**Included:** -Controlled prior knowledge of target items by conducting a pretest and no statistically significant difference between groups -Pretest score was not used as a covariate
Miles (2014)	Grammar	Pretests (editing and translation) were given, and initial analyses indicated that no significant differences were found on the pretests between groups.	**Included:** -Controlled prior knowledge of target items by conducting a pretest and no statistically significant difference between groups -Pretest score was not used as a covariate
Rogers (2015)	Grammar	A pretest (grammaticality judgment test) was given a week before the treatment. Initial analyses indicated that there was no significant difference between groups on the pretest scores.	**Included:** -Controlled prior knowledge of target items by conducting a pretest and no statistically significant difference between groups -Pretest score was not used as a covariate
Suzuki (2017)	Grammar	Participants had no prior knowledge of the target pronunciation rules (Spanish).	**Included:** no prior knowledge of the target grammatical rules
Suzuki & DeKeyser (2017a)	Grammar	Pretests (Time 1) was conducted before the treatment. All the Japanese consonant verbs were unknown to participants as shown by the pretest scores.	**Included:** -Controlled prior knowledge of target items by conducting a pretest. -Pretest score was not used as a covariate - Given the small percentage of valid responses for the pretest (Time 1), temporal

			measure of the rule application test at the pretest were not included in the subsequent analyses
Kasprowitz, Marsden, & Sephton (2019)	Grammar	Pretests (sentence-picture matching, acceptability judgement test) were given a week before the treatment. Given the difference observed in group scores at the sentence-matching pretest, the model was rerun with pretest as a control variable, rather than as part of the independent variable time. However, no significant main effect for pretest was observed. Therefore, the pretest was not included as a control variable in subsequent models. In the acceptability judgement pretest, a small difference between shorter spacing and control group's pretest scores (control group was higher) was found, although the CIs for both groups' effect sizes crossed zero, suggesting that this effect was not reliable.	**Included:** -Pretests were conducted, and not effect for the sentence-picture matching pretest was observed. A small difference but not reliable effect for the acceptability judgment pretest was found between control and shorter groups. However, our meta-analysis did not include the control group data when comparing to shorter group. Our meta-analysis used data from shorter (3.5 day) and longer (7-day) spacing groups for longer vs. shorter comparison category. -Pretest score was not used as a covariate
Nakata & Suzuki (2019b)	Grammar	Pretests (grammaticality judgment tests) were conducted. *****Additional analysis was conducted by us for our meta-analysis with the	**Included:** -Accuracy pretest scores were not included as a covariate. Only d-prime pretest scores and others (treatment duration, proficiency test scores) were included as covariates.

		descriptive data for the accuracy scores provided in Nakata and Suzuki's (2019b) study: no significant difference was found between blocked and interleaved practice groups (p = .35)	
Pan, Tajran, Lovelett, Osuna, & Richard (2019)	Grammar	Participants had no prior knowledge of the target pronunciation rules (Spanish).	**Included:** no prior knowledge of the target grammatical rules
Suzuki & Sunada (2020)	Grammar	Pretests (accuracy, fluency tests) were conducted (Accuracy test, Cronbach's alpha = .58 due to the lower accuracy rates before the treatment). -Accuracy scores on the pretest, which were standardized to reduce collinearity, were included as a covariate in models. *****Additional analysis was conducted by us for our meta-analysis with the descriptive data for the accuracy scores provided in Suzuki and Sunada's (2020) study: the difference between input-blocked and output-blocked groups was statistically significant (p = .05); and the difference between input-blocked and output-interleaved	**Excluded:** (1) No clear whether the pretest scores between groups were significantly different (no information provided). (2) Additional analysis (conducted by us) showed significant difference between groups on the pretests.

		groups was statistically significant (*p* = .03).	
Suzuki, Yokosawa, & Aline (2020)	Grammar	A pretest (sorting-questions test) was conducted before the experiment to control for prior knowledge of target structure (declarative knowledge of relative clauses), and no significant difference was found between groups.	**Included:** -Controlled prior knowledge of target items by conducting a pretest and no statistically significant difference between groups -Pretest score was used as a covariate
Carpenter & Mueller (2013)	Pronunciation	Participants had no prior knowledge of the target pronunciation rules (French).	**Included:** no prior knowledge of the target pronunciation rules
Li & DeKeyser (2019)	Pronunciation	A pretest (oral word naming) was conducted, but authors mentioned that participants had no prior knowledge of a tonal language such as Mandarin or Cantonese.	**Included:** -Pretest score was not used as a covariate

Notes. Having posttest scores adjusted for the pretest scores through ANCOVAs is methodologically preferable (Dimitrov & Rumrill, 2003). However, if a study that used the pretest scores as a covariate but found a significant difference between groups on the pretest, the study was excluded in the current meta-analysis.

References

*The studies included in the current meta-analysis are listed in Appendix S10.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*(3), 296−308. http://doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4*(5), 316−321. http://doi.org/10.1111/j.1467-9280.1993.tb00571.x

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52,* 566−577. http://doi.org/10.1016/j.jml.2005.01.012

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13*(2), 344−349. http://doi.org/10.1037/0278-7393.13.2.344

Collins, L., Halter, R. H., Lightbown, P. M., & Spada, N. (1999). Time and the distribution of time in L2 instruction. *TESOL Quarterly, 33*(4), 655–680. http://doi.org/10.2307/3587881.

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work: Journal of Prevention, Assessment & Rehabilitation, 20,* 159-165.

Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied Psycholinguistics, 24,* 369–383. http://doi.org/10.1017/S0142716403000195

Freed, B. F., Segalowitz, N., & Dewey D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion program. *Studies in Second Language Acquisition, 26*(2), 275–301. http://doi.org/10.1017/S0272263104262064

Kanayama, K., & Kasahara, K. (2017). What spaced learning is effective for long-term L2 vocabulary retention? *ARELE: Annual Review of English Language Education in Japan, 28,* 113–128. http://doi.org/10.20581/arele.28.0_113

Küpper-Tetzel, C. E., Erdfelder, E., & Dickhaeuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. Instructional Science, 42(3), 373–388. http://doi.org/10.1007/s11251-013-9285-2

Lapkin, S., Hart, D., & Harley, B. (1998). Case study of compact core French models: Attitude and achievement. In S. Lapkin (Ed.), *French second language education in Canada: Empirical studies* (pp.3–30). Toronto: University of Toronto Press.

Lee, E., & Choe, M. H. (2014). The effect of spaced repetitions on Korean elementary students' L2 English vocabulary learning. *Studies in English Education, 19*(1), 55−75.

Li, M. (2017). *Temporal distribution of practice and individual differences in the acquisition and retention of L2 Mandarin tonal word production* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Lightbown, P. M., & Spada, N. (1994). An innovative program for primary ESL students in Quebec. *TESOL Quarterly, 28*(3), 563−579. http://doi.org/10.2307/3587308

Mashhadi, A., Farvardin, M. T., & Mozaffari, A. (2017). Effects of spaced and massed distribution instruction on EFL learners' recall and retention of grammatical structures. *Teaching English Language, 11*(2), 57–75. http://doi.org/10.22132/TEL.2017.53183

Miles, S., & Kwon, C-J. (2008). Benefits of using CALL vocabulary programs to provide systematic word recycling. *English Teaching, 63*(1), 199–216.

Moher D., Liberati A., Tetzlaff J., & Altman D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med 6*(7): e1000097. http://doi.org/10.1371/journal.pmed1000097

Namaziandost, E., Nasri, M., Rahimi Esfahani, F., & Keshmirshekan, M. H. (2019). The impacts of spaced and massed distribution instruction on EFL learners' vocabulary learning. *Cogent Education, 6*:1661131, 1–10. http://doi.org/10.1080/2331186X.2019.1661131

Namaziandost, E., Rahimi Esfahani, F., & Hashemifardnia, A. (2018). The comparative effect of spacing instruction and massed instruction on intermediate EFL learners' reading comprehension. *SAGE Open, 8*(4), 1–8. http://doi.org/10.1177/2158244018811024

Nakata, T. (2013). *Optimising second language vocabulary learning from flashcards* (Unpublished doctoral dissertation). Victoria University of Wellington, New Zealand.

Nakata, T., Tada, S., McLean, S., & Kim, Y. A. (2020). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly*. http://doi.org/10.1002/tesq.596

Pan, S. C., Lovelett, J. T., Phun, V., & Rickard, T. C. (2019). The synergistic benefits of systematic and random interleaving for second language grammar learning. *Journal of Applied Research in Memory and Cognition, 8*(4), 450–462. http://doi.org/10.1016/j.jarmac.2019.07.004

Pyc, M. A., & Rawson, K. A. (2012a). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition, 40*, 976–988. doi:10.3758/s13421-012-0200-x

Pyc, M. A., & Rawson, K. A. (2012b). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 737 –746. http://doi.org/10.1037/a0026166

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. J*ournal of Experimental Psychology: Applied, 24*(1), 57–71. http://doi.org/10.1037/xap0000146

Schneider, V. I., Healy, A. F., & Bourne, L. E. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 77–90). Mahwah, NJ: Erlbaum.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under different conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46*(2), 419–440. http://doi.org/10.1006/jmla.2001.2813

Schuetze, U., & Weimer-Stuckmann, G. (2011). Retention in SLA lexical processing. *CALICO Journal, 28*(2), 460–472. http://doi.org/10.1558/cj.28.2.460-472

Serrano, R. (2011). The time factor in EFL classroom practice. *Language Learning*, *61*(1) 117−145. http://doi.org/10.1111/j.1467-9922.2010.00591.x

Serrano, R., & Munoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System, 35*(3), 305−321. http://doi.org/10.1016/j.system.2007.02.001

Suzuki, Y. (2018). The role of procedural learning ability in automatization of L2 morphology under different learning schedules. *Studies in Second Language Acquisition, 40*(4), 923−937. http://doi.org/10.1017/S0272263117000249

Suzuki, Y. (2019). Individualization of practice distribution in second language grammar learning: A role of metalinguistic rule rehearsal ability and working memory capacity. *Journal of Second Language Studies, 2*(2), 170–197. http://doi.org/10.1075/bct.116.02suz

Suzuki, Y., & DeKeyser, R. (2017b). Exploratory research on second language practice distribution: An aptitude x treatment interaction. *Applied Psycholinguistics, 38*(1), 27−56. http://doi.org/10.1017/S0142716416000084

Suzuki, Y., & Sunada, M. (2020). Dynamic interplay between practice type and practice schedule in a second language. *Studies in Second Language Acquisition, 42*(1), 169–197. http://doi.org/10.1017/S0272263119000470

Tsai, L. S. (1927). The relation of retention to the distribution of relearning. *Journal of Experimental Psychology,* 10(1), 30–39. http://doi.org/10.1037/h0071614

Appendix S2: Category Criteria

1. **Spaced vs. Massed category**: When the study time devoted to any given item is subject to interruptions of intervening items or intervening time, the learning is *spaced*. In contrast, when the treatment without any interruptions of intervening items or intervening time, the learning is considered *massed*.
    - Blocking corresponds to massed practice while interleaving is equivalent to spaced practice (Nakata & Suzuki, 2019b; Suzuki, Yokosawa, & Aline, 2020). In blocked practice, learners practice all items from one concept at a time before the learners move on to the next set of another concept. In contrast, in the interleaved practice, learners practice multiple (different types of) concepts simultaneously. For example, studies comparing blocked to interleaved practice such as Nakata and Suzuki (2019b), Suzuki *et al.* (2020), and Pan, Tajran, Lovelett, Osuna, & Rickard (2019, Experiments 1 and 2) were categorized into the spaced vs. massed comparison.

2. **Longer vs. Shorter category**: When a measurable time interval separates study episodes for a given item is shorter than the time interval treated in the other experimental group, the learning is considered *shorter* and the other group considered *longer*. For example, a study involved 1-day spaced practice and 3-day spaced practice groups, the 1-day group is considered a shorter spacing group and 3-day group is considered a longer spacing group. However, there are some studies that used terms differently.
    - Suzuki and DeKeyser (2017a) defined a one-day interval practice as a massed practice. In the current meta-analysis, it was considered a shorter spacing schedule.

- Serrano and Huang (2018) compared intensive practice (one-day interval) with spaced practice (seven-day interval). An intensive practice was considered shorter spaced practice, and spaced practice was considered longer spaced practice.
- Snoder (2017) compared intensive learning schedule (Day 1, Day 2, and Day 4) with expanding learning schedule (Day 1, Day 7, and Day 16). Both learning schedules are expanding spacing conditions. To have comparable groups for this category, the intensive learning schedule was considered a shorter spaced practice, and the expanding learning schedule was considered a longer spaced practice.
- Nakata and Suzuki (2019a) used the terms "massed" and "spaced" in their article. Nakata and Suzuki (2019a) involved the "massed" condition is massed in the sense that semantically related target items are repeated without any intervals (e.g., raccoon, weasel, otter), not in the sense that the same target item is repeated without any intervals (e.g., raccoon, raccoon, raccoon). Since the massed condition is not (pure) massed condition in Nakata and Suzuki (2019a), the condition was considered a shorter spaced practice. Spaced condition was then considered a longer spaced practice.
- Carpenter and Mueller (2013) and Pan *et al.* (2019, Experiments 3 and 4) did not provide pure massed (blocked) / spaced (interleaved) conditions. For example, Carpenter and Mueller (2013, p. 673) included 8 distinct French pronunciation rules. 32 French words were (4 words per rule) were presented to each participant according to a blocked versus interleaved schedule. The items for each rule were randomly assigned for each participant to a predetermined sequence of blocked (B) or interleaved (I) groups of items in the order BIIBBIIB: a participants saw and heard a blocked group of 4 words that represented a single rule, followed by an interleaved group of 4 words that each represented a different rule. In Pan *et al.* (2019, Experiments 3 and 4), participants learned Spanish verbs in the preterite (P) and imperfect (I) past tenses in blocked practice (session 1: PPPP / session 2: IIII) or interleaved

practice (session 1: PPPIIIPI / session 2: PI) with 7-day intersession interval (between two sessions). In the current meta-analysis, therefore, the practice that was not manipulated pure blocked, it was considered a shorter spaced practice, and the interleaved practice was considered a longer spaced practice. That is, Carpenter and Mueller's (2013) and Pan *et al.* (2019, Experiments 3 and 4) studies were categorized in the longer vs. shorter comparison.

3. **Equal vs. Expanding category**: Spaced practice with equal, uniform, or fixed intervals was considered an equal spaced practice, and spaced practice with gradually increasing intervals was considered an expanding spaced practice. For example, Çekiç and Bakla (2019) used three spacing schedules defined as fixed spacing (once a week for nine weeks; nine sessions in total), spaced massing with fixed intervals (on the first, second, and third weeks; three sessions in total), and spaced massing with expanding intervals (on the first, third, and seventh weeks; three sessions in total). To have comparable groups for the equal and expanding spacing category, spaced massing conditions with fixed intervals and expanding intervals were selected from this study.
4. Some studies (e.g., Karpicke & Bauernschmidt, 2011; Nakata, 2015a) included multiple spaced schedules (e.g., massed, short, medium, long, equal, and expanding), which were separately categorized to avoid overlaps in the number of participants in each category. For example, Nakata (2015a) involved massed, absolute spacing (short, medium, and long), and relative spacing (equal and expanding) schedules. In his study, absolute spacing schedules followed a between-participants design, and relative spacing schedules followed a within-participants design. The massed condition and short spacing condition from the absolute spacing schedule were categorized into the spaced vs. massed comparison. Meanwhile, the short and long spacing conditions from the absolute spacing schedule were placed

under the longer vs. shorter comparison. The equal and expanding conditions from each absolute spacing schedule (i.e., equal and expanding spacing from the short, medium, and long spacing schedules) were classified under the equal vs. expanding comparison.

Appendix S3: Coding Scheme

Table S3.1

Coding Scheme

Variables	Values				
Learner					
Age	Young (Primary, Secondary school)		Adult (University, Others)		
Methodology					
Learning target	Vocabulary	Grammar	Pronunciation		
Number of sessions	Single session	Multiple sessions			
Type of practice	Test-restudy (all) trial	Test-restudy (not recalled) trial	Study trial	Test trial	Study-test trial
Activity type	Paired associate	Comprehension activities	Production activities	Combined activities	

Provision of feedback	Absence	Presence
Feedback timing	Immediate	Delayed
Frequency of practice		
Retention interval		

Notes. Variables without labelled values are continuous, non-categorical, or open-ended. Test-restudy (all) trial = A test trial was followed by a restudy trial for all target items. Test-restudy (not recalled) trial = A test trial was followed by a restudy trial for the items that were not recalled by participants. See Tables S4.2 and S4.3 in Appendix S4 for the coding details for *Type of practice* variable. Combined activities = both comprehension and production-based activities were provided during the practice session(s). See Tables S4.4 and S4.5 in Appendix S4 for the coding details for *Activity type*.

Appendix S4: Details of the Studies Included in the Meta-Analysis

Table S4.1.

Details of Variables of the Studies Included in the Meta-Analysis

Study	Age	LT	Number of sessions	Posttest format	Feed back	Feedback timing	FP	Immediate posttest	Length of RI
Bloom & Shuell, 1981	Secondary	V	Multiple	Form recall (Productive)	-	Self-correction (possibly) but no timing information provided	3	Yes	4 days
Pashler *et al*., 2003 (Ex 1)	University	V	Single	Meaning recall (Receptive)	Yes	Immediate: if a response was incorrect, the learner was shown the L1 translation	2	No	1 day
Bahrick & Hall, 2005 (Ex 1)	University	V	Multiple	Meaning recall (Receptive)	Yes	Delayed: after retrieval practice (test trial), word pairs that had not been correctly recalled were presented again	4	No	14days
Pyc & Rawson, 2007 (Ex 1 / Ex2)	University	V	Single	Meaning recall (Receptive)	Yes	Immediate: response was given and presented together with target for 4s	3	Yes	-
Cepeda *et al*., 2009 (Ex1)	University	V	Multiple	Meaning recall (Receptive)	Yes	Immediate: response was given immediate after each item	2	No	10 days
Pyc & Rawson, 2009 (Ex 1 / Ex2)	University	V	Single	Form recall (Productive)	Yes	Delayed: not explicitly given, but students were informed that only items that were incorrectly retrieved would receive a restudy trial	1~10 (not clear)	No	7 days

Bird, 2010	University (19-23 years)	G	Multiple	Grammaticality judgement test (Receptive)	Yes	Delayed: after 30 min the transparency was presented on the overhead projector, and participants were given the correct answers as well as brief explanations of why each verb phrase in each sentence was correct or incorrect and how to form the correct sentence	5	No	7 days / 60 days
Karpicke & Bauernschmidt, 2011	University	V	Single	Meaning recall (Receptive)	No		4	No	7 days
Gerbier & Koenig, 2012 (Ex 1)	University	V	Multiple	Meaning recall (L2-L1): participants were asked to say aloud the L1 associated with each pseudoword (Receptive)	-		4	No	2 days
Gerbier & Koenig, 2012 (Ex 2)	University	V	Multiple	Meaning recall: writing down the L1 associated word with each pseudoword (Receptive)	-		4	No	2 days
Carpenter & Mueller, 2013 (Ex 1)	University	P	Single	Rule-correction pronunciation test (Multiple-choice): listening to each	-		1	Yes	-

				of the 3 recordings for each word and choose which one was correct (Receptive)					
Carpenter & Mueller, 2013 (Ex 2)	University	P	Single	Multiple-choice test (Receptive)	-		1	Yes	-
Carpenter & Mueller, 2013 (Ex 3)	University	P	Single	Participants were asked to pronounce each word out loud (Productive)	-		1	Yes	-
Carpenter & Mueller, 2013 (Ex 4)	University	P	Single	Multiple-choice test (Receptive)	-		1	Yes (5 min)	-
Kang *et al*., 2014	Other (average = 36.4 years)	V	Multiple	Meaning recall (Receptive)	Yes	Immediate: The intact Japanese-English pair was presented for 2s	3	No	56 days
Lee & Choe, 2014 (Ex2)	Primary	V	Multiple	Form recall (Productive) / Meaning recognition (Receptive)	-		4	Yes	5 weeks (Ex 2)
Miles, 2014	University	G	Multiple	Error correction task (Receptive) / translation task: to translate L1	No		2	Yes	35 days

				sentences into English (Productive)					
Schuetze, 2014 (Ex 1 / Ex 2)	University (17-24 years)	V	Multiple	Form recall (Productive)	No		4 (Ex1) / 5 (Ex2)	No	1 day / 4 weeks / 8 weeks
Gerbier, Toppino, & Koenig, 2015 (Ex 1a)	University	V	Multiple	Meaning recall: writing down the L1 associated with each pseudoword (Receptive)	Yes	Delayed: After each test trial, word pairs were projected on the screen for 25 sec each	3	No	2 days
Gerbier, Toppino, & Koenig, 2015 (Ex 1b)	University	V	Multiple	Meaning recall (Receptive)	Yes	Delayed: After each test trial, word pairs were projected on the screen for 25 sec each	3	No	6 days
Gerbier, Toppino, & Koenig, 2015 (Ex 1c)	University	V	Multiple	Meaning recall (Receptive)	Yes	Delayed: After each test trial, word pairs were projected on the screen for 25 sec each	3	No	13 days
Nakata, 2015a	University	V	Single	Receptive meaning recall / Productive form recall	Yes	Immediate: After each response, target English word, L1 translation, and learners' response were shown for 5 seconds	4	Yes	7 days
Rogers, 2015	University (19.5 years)	G	Multiple	Form recognition: grammaticality judgement test (Receptive)	Yes	Immediate: Yes/no comprehension check question answers were given after each	5	Yes	42 days

Kanayama & Kasahara, 2016	University	V	Multiple	Meaning recall (Receptive)	No		4	Yes	21 days
Lotfolahi & Salehi, 2016	Primary (8~12 years)	V	Multiple	Meaning recall (Receptive)	Yes	Delayed: word-pairs and a sample sentence for each word were given after test trial practice	1	No	7 days / 35 days
Nakata & Webb, 2016 (Ex 1 and 2)	University	V	Single	Meaning recall (Receptive) / form recall (Productive)	Yes	Immediate: answer was given after each item	5 (Ex1) / 4 (Ex2)	Yes	7 days
Khoii & Abed, 2017	Secondary	V	Multiple	Vocabulary Knowledge Scale (VKS): meaning recall (Receptive)	Yes	Immediate: whenever an error was made, teacher first let students know and let them do self-correct. Also, peer- or teacher feedback if there is an error	3	Yes	-
Snoder, 2017	Secondary	V (collocations)	Multiple	form recall: participants were asked to complete the target collocation by filling in a gap in the English translation of the Swedish cue (Productive)	-		2	No	21 days
Suzuki, 2017	University	G (nonsense)	Multiple	form recall / productive (rule application) / productive (form recall from pictures)	Yes	Immediate: experimenter provided a recast as a form of feedback to incorrect responses	27: (4 times for vocab+ 4 times for gramma	No (Monitoring test 2 can be an immediate posttest)	7days / 28 days

Suzuki & DeKeyser, 2017a	University	G	Multiple	productive (rule application) / productive (form recall: picture sentence completion)	Yes	Immediate: recast feedback was given when participants produced an incorrect form of target verb	r+1 for monitoring) x 3 sessions 6: 2(comprehension)+2 (picture description) + 2(video description)	Yes	7 days / 28 days
Serrano & Huang, 2018	Secondary	V	Multiple	meaning recognition: matching (Receptive)	No		5	Yes	4 days (shorter group)/ 28 days (longer group)
Çekiç & Bakla, 2019	Secondary	V	Multiple	Receptive: MC (immediate) / VKS (delayed)	Yes	Immediate: Answers to multiple-choice questions for reading comprehension were immediately given.	2	Yes	7 days
Kasprowicz *et al*., 2019	Primary	G	Multiple	Sentence-picture matching; written acceptability judgement test (Receptive)	Yes	Immediate: Correct and incorrect responses were indicated aurally by different sounds and visually via the progress. Learners also received a short explanation	3	Yes	42 days
Koval, 2019	University	V	Single	Form-meaning mapping): MC/	-		4	Yes	44~78hr = 2.5 days

				Matching (Receptive)					
Li & DeKeyser, 2019	Other (18-41 yrs)	P	Multiple	Oral picture naming/written picture naming/oral word naming (Productive)	Yes	Immediate	3	No	7 days / 28 days
Nakata & Suzuki, 2019a	University	V	Single	Paired associate (Receptive)	Yes	Immediate: given after each response (for 5 seconds)	3	Yes	7 days
Nakata & Suzuki, 2019b	University	G	Single	Grammaticality judgement test (Receptive)	Yes	Immediate: given after each response / metalinguistic explanation of the target structure were provided as feedback for 12 seconds	10	Yes	7 days
Pan, Tajran, Lovelett, Osuna, & Rickard, 2019 (Ex 1/ Ex 2)	University	G	Single	Fill-in-the-blank multiple-choice (Receptive)	Yes	Immediate: correct answer including suffix, tense name, and relevant pronoun was provided on each test trial	2	No	2 days
Pan, Tajran, Lovelett, Osuna, & Rickard, 2019 (Ex 3/ Ex 4)	University	G	Single	Fill-in-the-blank multiple-choice (Receptive)	Yes	Immediate: correct answer including suffix, tense name, and relevant pronoun was provided on each test trial	2	No	7 days
Koval (2020)	University	V	Single	Form-recognition/translation test (L2-L1) / Matching	Yes	Immediate: word pair was presented after each retrieval	6	Yes	14 days

Nakata & Elgort, 2021	University	V (pseudo words)	Single	(Receptive): Meaning recall, form recognition (MC)	Yes	Immediate: correct meaning of the target pseudoword was given	3	Yes	2 days
Rogers & Cheung, 2020a	Primary	V	Multiple	meaning recognition: multiple-choice (Receptive)	-		2	No	28 days
Rogers & Cheung, 2020b	Primary	V	Multiple	Form recall: Crossword puzzle production test (Productive)	Yes	Delayed: feedback from teachers was given after practice	2	No	28 days
Suzuki, Yokosawa, & Aline, 2020	University	G	Single	Oral description (Productive): describing pictures using appropriate relative pronouns or the relative adverb (e.g., *where*)	Yes	Immediate: a correct answer was provided both visually and aurally and the example sentence remained on the screen for 8s	10	Yes	7 days

Notes. LT = Learning target, V = Vocabulary, G = Grammar, P = Pronunciation, FP = Frequency of practice, RI = Retention interval. Frequency of (repeated) practice reported in this table is the number of repetitions reported in each original study, and some of the frequency numbers coded for this meta-analysis are different according to the number of immediate and delayed posttests (see Table S9.1, Appendix S9).

Table S4.2.

Coding for variable Type of practice

Code	Type of practice	*k*	Study
1	Test-restudy (all) trial	24	Bloom & Shuell, 1981; Pyc & Rawson, 2007 (Ex 1 & 2); Cepeda *et al*., 2009 (Ex 1); Bird, 2010 ; Kang *et al*., 2014; Gerbier, Toppino, & Koenig, 2015 (Ex 1a, 1b, & 1c) ; Nakata, 2015a; Rogers, 2015; Kanayama & Kasahara, 2016; Lotfolahi & Salehi, 2016; Nakata & Webb, 2016 (Ex 1 & 2); Li & DeKeyser, 2019; Nakata & Suzuki, 2019a, 2019b; Pan, Tajran, Lovelett, Osuna, & Rickard, 2019 (Ex 1, 2, 3, & 4); Koval, 2020; Nakata & Elgort, 2021; Rogers & Cheung, 2020b; Suzuki, Yokosawa, & Aline, 2020
2	Test-restudy (not recalled) trial	6	Pashler *et al*., 2003 (Ex1); Bahrick & Hall, 2005 (Ex1); Pyc & Rawson, 2009 (Ex 1 & 2); Karpicke & Bauernschmidt, 2011; Khoii & Abed, 2017; Suzuki, 2017
3	Study trial	6	Carpenter & Mueller, 2013 (Ex 1, 2, 3, & 4); Gerbier & Koenig, 2012 (Ex 1 & 2); Lee & Choe, 2014 (Ex 2); Snoder, 2017; Rogers & Cheung, 2020a; Koval, 2019
4	Test trial	8	Miles, 2014; Schuetze, 2014 (Ex 1 & 2)
5	Study-test trial	3	Serrano & Huang, 2018; Çekiç & Bakla, 2019; Kasprowicz, Marsden, & Sephton, 2019

Notes. k = Number of study experiments. Suzuki & DeKeyser (2017a) was excluded to code for this moderator variable Type of practice, because Suzuki and DeKeyser (2017a) involved test-restudy (not recalled) trial in the production task and test-restudy (all) trial in the narrative task (see Table S4.3 below for the details).

Description of the coding *Type of practice*

- Test-restudy (all) trial = A practice that involves a test trial, followed by a restudy trial for all target items was coded as test-restudy (all) trial. A test trial, followed by feedback for all target items in the practice session was also considered test-restudy (all) trial.
- Test-restudy (not recalled) trial = A practice that involves a test trial, followed by a restudy trial for only the items that were not recalled by participants was coded as test-restudy (not recalled) trial. Also, a practice that provided feedback for the items that were incorrect from the test trial was coded as test-restudy (not recalled) trial.
- Study trial = A practice that involves a study-only trial was coded as study trial.
- Test trial = A practice that involves a test-only trial without feedback was coded as test trial.
- Study-test trial = A practice that involves a study trial, followed by a test trial was coded as study-test trial. A practice that involved study-test trials, followed by feedback for either all items or incorrect (not recalled) items was also coded as study-test trial.

Table S4.3.

Details of Coding Variable "Type of Practice" for the Studies Included in the Meta-Analysis

Study	Task used for practice session(s)	Trial type
Test-restudy (for all items) trial		
Bloom & Shuell (1981)	A series of three written exercises was given for use during class study periods. (1) Multiple-choice: participants were asked to choose a correct French word associated with English word given, (2) fill-in exercise: participants were to write	Test (possibly self-correction with the word list given) trials

	in French the name of the occupation described in a sentence, and (3) written practice: participants were to write the French word for each occupation given in English. The list of word pairs was given to study only during class and collected at the end of each day's work.	
Pyc & Rawson (2007, Ex 1 and 2)	Swahili word was presented alone, and participants were asked to enter the English translation in a text box provided below the Swahili word. After 8 sec, the response box was removed from the screen, and the Swahili and English words were presented together for 4 sec (regardless of whether the response was correct or not).	Test-restudy (all items) trials
Cepeda *et al.* (2009, Ex 1 & 2)	Each Swahili word was presented, and participants were asked to type the English word. After each retrieval, Swahili-English word pair appeared for 5 sec.	Test-with-feedback (all) trials
Bird (2010)	Worksheets of (simple present/past perfect) sentences were given to participants. Participants read each sentence and judge whether the sentence was grammatically correct. If a sentence was judged incorrect, participants were to rewrite the sentence to make it grammatically correct. After 30 min, participants were given the correct answers as well as brief (oral) explanations of why each verb phrase in each sentence was correct or incorrect and how to form the correct sentence.	Test-with-feedback (all) trials
Kang *et al.* (2014)	The target word was presented alone for 6 sec, and during that time participants were asked to retrieve and type in the L1 meaning. After 6 sec had elapsed, the word pair would be presented for 2 sec.	Test-with-feedback (all) trials
Gerbier, Toppino, & Koenig (2015, Ex 1a, 1b, and 1c)	Participants were asked to recall the correct member of the pair. After each test trial, word pairs were shown on the screen for 25 sec each.	Test-restudy (all) trials

Nakata (2015a)	Participants were asked to type the target word corresponding to the L1 translation provided. After each response, the target word, L1 translation, and participants' response were shown for 5 sec per response as feedback.	Test-with-feedback (all) trials
Rogers (2015)	Each stimulus sentence was displayed onto the white board for 7 sec. Following this, the sentence was replaced by a Yes/No comprehension check question. Participants answered the question by ticking one of two boxes on an answer sheet. After each practice session, a teacher displayed the correct answers using the projector.	Study-test (with feedback for all items) trials
Kanayama & Kasahara (2016)	Participants were asked to write down the meaning of each word in L1. In the subsequent session, participants were given a word list and a blank sheet of paper in case participants wished to memorize the target words by writing them. It was followed by an immediate recall test.	Test-restudy (all) trials
Lotfolahi & Salehi (2016)	Participants were asked to write down the meaning of each L2 word in L1. Afterwards, all word pairs and a sample sentence for each one were given. Teacher molded the word pairs and sample sentences, and participants then repeated them chorally. In addition, children were given five minutes to practice the meaning of L2 words. Finally, participants were given four minutes to practice writing down the meaning of each L2 word.	Test-restudy (all) trials
Nakata & Webb (2016, Ex 1 and 2)	(Ex 1) In the second and third encounters, target items were practised in a receptive recall format. Participants were asked to translate target L2 words into L1. In the fourth and fifth encounters, target items were practised in a productive recall format. Participants were presented with the L1 meanings and asked to type the corresponding L2 translations. After each response, the target word, L1 meaning, and learners' response were shown for 5 seconds as feedback.	Test-with-feedback (all) trials

	(Ex 2) Participants were presented with the L1 meanings and asked to type the corresponding L2 translations. After each response, the target word, L1 meaning, and learners' response were shown for 5 seconds as feedback.	
Li & DeKeyser (2019)	(Tone identification practice) Participants heard audio recordings of target monosyllabic words, one at a time, and were asked to choose the correct tone on a paper sheet with the target monosyllables on it. The experimenter provided feedback for each trial. (Tone production practice) Participants were presented with the monosyllabic words on the screen, one at a time, and were asked to pronounce the words. The feedback for each trial was given.	Test-with-feedback (all) trials
Nakata & Suzuki (2019a)	Participants were presented with an L2 target word and asked to type in the corresponding L1 translation. After each response, the correct answer was provided as feedback for 5 seconds.	Test-with-feedback (all) trials
Nakata & Suzuki (2019b)	Participants were presented with a sentence where a verb or verb phrase was replaced with a blank together with four options. Participants were instructed to choose the most appropriate verb or verb phrase to complete the sentence. To make the intended meaning clear, the L1 translation of each sentence was also provided. After each response, the correct answer and metalinguistic explanation of the target structure were provided for 12 sec.	Test-with-feedback (all) trials
Pan, Tajran, Lovelett, Osuna, & Rickard (2019, Ex 1, 2, 3, and 4)	(Tense rule practice) Participants made a yes/no judgment as to whether each presented sentence reflected the tense that participants had learned. (Verb suffixes practice) Participants typed the proper suffix of the verb given into the sentence. A summary slide was presented after each phase and correct answer feedback was provided after each practice trial.	Test-with-feedback (all) trials

Koval (2020)	Participants were asked to say the L1 translation aloud for the L2 target word. Their responses were audio-recorded. The word pair was presented immediately after each retrieval.	Test-with-feedback (all) trials
Nakata & Elgort (2021)	The participants practiced guessing the meaning of the pseudoword and type their answer either in their L1 or L2. Correct meaning of the pseudoword was presented as feedback in the form of L1 (Japanese) translation equivalent and L2 (English) synonym.	Test-with-feedback (all) trials
Rogers & Cheung (2020b)	Participants performed two short crossword puzzles. Feedback from teachers was given after practice.	Test-with-feedback (all) trials
Suzuki, Yokosawa, & Aline (2020)	All lexical items necessary for output practice (oral description) were shown in the picture on a screen. Participants were then asked to describe pictures that appeared on a screen using target features. After that, correct answer was provided both visually and aurally and the example sentence was also given.	Test-with-feedback (all) trials
Test-restudy (for not recalled items) trial		
Pashler *et al.* (2003, Ex 1)	The Eskimo word was presented, and participants were asked to type the English word. If a response was incorrect, the correct L1 translation (English word) was displayed for 5 sec.	Test-with-feedback (for incorrect ones) trials
Bahrick & Hall (2005, Ex 1)	Each Swahili word was presented, and participants were asked to type the associated English word on the keyboard. After test trial, word pairs that had not been correctly recalled were presented again.	Test-restudy (restudy for not recalled ones) trials
Pyc & Rawson (2009, Ex 1 and 2)	Participants were presented with an English word, and they had to recall the Swahili word. After test trial, word pairs that had not been correctly recalled were presented again for restudy.	Test-restudy (restudy for not recalled ones) trials

Karpicke & Bauernschmidt (2011)	Swahili word was presented, and participants were asked to type the English word. If a participant recalled a word, the word was no longer presented in subsequent trial. If a participant failed to recall a word, the word pair was presented again during the next cycle of the trial. However, all the target words were repeatedly tested during the three sessions (trials) regardless of whether it was recalled or not.	Test-restudy (restudy for not recalled ones) trials
Khoii & Abed (2017)	Each word pair was shown again, and participants were asked to make sentences orally using each target item. Whenever an error was made, the teacher alerted the students by using a facial expression, making a hand gesture, or saying "Can you say that again?" in order to have them self-correct. Peer- and teacher-feedback were also used when self-correction did not occur.	Test-with-feedback (for incorrect ones) trials
Suzuki (2017)	Participants saw an animation video in which a man performed the action of the verbs. (Step 1) Each video clip showed an uninflected verb in the top right corner for the entire duration of the video clip (i.e., 8 seconds), and participants practiced using morphological rules while seeing the lexical items. (Steps 2 and 3) Each video clip showed the animation without presenting an uninflected verb for the first 4 seconds, and the verb appeared in the right top corner as a hint for the last 4 seconds. Participants had to orally describe the animation using the present progressive form of the verb. Experimenter provided a recast as a form of feedback to incorrect responses. (Step 4) same procedure in the steps 2 and 3, but the presentation order was changed. Afterwards, monitoring test 1 (form recall test, rule application, and form recall from pictures; same as the posttests) were conducted. No feedback was given. In the next training sessions, vocabulary practice with explicit explanations was provided, followed by the same procedure (steps 1, 2, 3, and 4) was given. Monitoring test 2 was provided at the end of the training session.	Test-with-feedback (for incorrect ones) trials

Study trial

Carpenter & Mueller (2013, Ex 1, 2, 3, and 4)	Participants saw and heard the 32 French words in either blocked or interleaved condition. Each word was presented on the screen for 4s (each of the 32 words was presented only once).	Study trials
Gerbier & Koenig (2012, Ex 1 and 2)	Participants were required to learn word/pseudoword pairs.	Study trials
Lee & Choe (2014, Ex 2)	Participants read and listened to the target words. After that, participants practiced writing the words. Finally, participants practiced speaking the words in a structured conversation.	Study trials
Snoder (2017)	(in the second and third treatment) Participants read six short texts containing the 14 target collocations and answered questions on the texts. In the next practice session, participants reread three short texts containing 14 of the target collocations and wrote new titles for these texts. Participants then studied the 14 other target collocations again.	Study trials
Rogers & Cheung (2020a)	Verbal drilling and exercises: Teacher pronounced the target item and had the students repeat the target item aloud in unison (i.e., choral drilling). The teacher followed up the pronunciation drills with either crossword puzzles (Classes 1 and 2) or word search (Classes 3 and 4).	Study trials
Koval (2019)	Participants read L2 sentences, followed by comprehension questions. However, none of the comprehension questions contained a target word translation to avoid causing additional processing of the target word or its meaning.	Study trials

Test trial		
Miles (2014)	To review what participants have learned, participants performed some activities (sentence completion, error correction, and translation L1 sentences into L2).	Test trials
Schuetze (2014, Ex 1 and 2)	Participants were asked to write down the target words they saw and heard on a piece of paper (form recall).	Test trials
Study-test trial		
Serrano & Huang (2018)	Participants reread a passage, while listening to it at the same time. After reading while listening, the participants were given the reading comprehension activities (true/false, multiple-choice, and fill-in-the-blank) together with the glossary to refer to when necessary. In the comprehension questions, participants can revisit each target word at least once.	Study-test trials
Çekiç & Bakla (2019)	Participants reread short passages. They were able to see the definition of the word and hear the related definition or synonym when they clicked on the speaker icon next to each word written in bold. Each passage was accompanied by two multiple-choice reading comprehension questions. All the comprehension questions stimulated participants to process target words. Upon answering these questions, the participants were given immediate feedback.	Study-test (with feedback for all items) trials
Kasprowicz, Marsden, & Sephton (2019)	Participants performed a series of mini games. In each mini game, explicit information of one pair of French verb inflections was given (as initial learning for target feature), followed by reading and listening activities (with questions) in	Study-test (with feedback for incorrect ones) trials

	which participants were required to notice the feature and connect it with the meaning. Following incorrect answers, a short explanation was given.	
Excluded study: both test-restudy (all) and (not recalled) trials		
*Suzuki & DeKeyser (2017a)	(Comprehension practice) Cards that had the same pictures as the ones used during the vocabulary training session were laid out on the table. The experimenter read aloud the sentence that described the action in one of the pictures, and participants were asked to pick up the corresponding card as soon as possible. (Production task) Picture matching came next. The roles were reversed from those in the comprehension practice: Participants were asked to describe the picture to the experimenter, so that the experimenter could pick up the picture that participants described. When participants could not describe the picture, the experimenter described the card for them. Feedback in the form of recasting was given if participants produced an incorrect form of the verb. (Narrative task) participants performed a narrative task, describing what a person in a video was doing. Each action was performed for 10 seconds, and the participants were told to describe the action using the *-te* form while the video was played. After each video clip, the correct sentence was presented both aurally and visually on the screen for 4 seconds, and the participants automatically moved on to the next movie clip.	Test-with-feedback (for incorrect ones in the production task/ for all items in the narrative task) trials

Table S4.4.

Coding for variable Activity type

Coding	Activity type
1	**Paired associate** (*k* = 19): Pashler *et al*., 2003 (Ex 1); Bahrick & Hall, 2005 (Ex1); Pyc & Rawson, 2007 (Ex 1 & 2); Cepeda *et al*., 2009 (Ex 1); Pyc & Rawson, 2009 (Ex 1 & 2); Karpicke & Bauernschmidt, 2011; Gerbier & Koenig, 2012 (Ex 1 & 2); Kang *et al*., 2014; Gerbier *et al*., 2015; Nakata, 2015a; Kanayama & Kasahara, 2016; Lotfolahi & Salehi, 2016; Nakata & Webb, 2016 (Ex 1 & 2); Nakata & Suzuki, 2019a; Koval, 2019
2	**Comprehension activities** (*k* = 15): Bird, 2010; Carpenter & Mueller, 2013 (Ex 1 & 2); Rogers, 2015; Snoder, 2017; Cekiç & Bakla, 2019; Kasprowicz *et al*., 2019; Koval, 2019; Li & DeKeyser, 2019; Nakata & Suzuki, 2019b; Pan *et al*., 2020 (Ex 1, 2, 3, & 4); Nakata & Elgort, 2021
3	**Production activities** (*k* = 9): Carpenter & Mueller, 2013 (Ex 3 & 4); Shuetze, 2014 (Ex 1 & 2); Khoii & Abed, 2017; Suzuki, 2017; Rogers & Cheung, 2020a, 2020b; Suzuki *et al*., 2020
4	**Combined activities (both comprehension and production activities provided)** (*k* = 5): Bloom & Shuell, 1981; Lee & Choe, 2014 (Ex 1); Miles, 2014; Suzuki & DeKeyser, 2017a; Serrano & Huang, 2018

Note. k = Number of study experiments.

Description of comprehension, production, and combined activities

- Comprehension activities = input-based activities (e.g., multiple-choice, error identification, selecting a correct response to complete a sentence, matching such as sentence-picture or form-meaning matching), structured input activities (reading and listening), and meaning-focused reading or listening activities
- Production activities = output-focused activities (e.g., fill-in-the-blanks, sentence completion), controlled production activities (e.g., translating sentences), meaning-focused output activities (e.g., writing sentences following examples, producing target item based on the aural/written cues or triggers), production-based activities such as orally describing a picture using target items
- Combined activities = activities involved both comprehension and production activities. For example, if a study involved reading and listening activities for practicing target items, followed by giving opportunities to practice the items through writing and speaking, it was coded as combined activities (e.g., Lee & Choe, 2014, Ex 2). Also, a study that involves both multiple-choice questions (comprehension activity) and fill-in-the blanks (production activity) was coded as combined activities (e.g., Bloom & Shuell, 1981).

Table S4.5.

Details of Coding Variable "Activity type" for the Studies Included in the Meta-Analysis

Comprehension activities (*k* = 15):	
Bird, 2010	Reading each sentence and judging it by writing a correct response
Carpenter & Mueller, 2013 (Ex 1 & 2)	Listening and choosing a correct response (multiple-choice)
Rogers, 2015	Reading each sentence and performing a yes/no comprehension question
Snoder, 2017	Reading short texts
Cekic & Bakla, 2019	Reading short passages, followed by comprehension questions (multiple-choice)
Kasprowicz *et al*., 2019	Reading and listening activities for form-meaning mapping

Koval, 2019	Reading sentences
Li & DeKeyser, 2019	Listening and choosing a correct response
Nakata & Suzuki, 2019b	Performing a multiple-choice exercise to complete a sentence
Pan *et al*., 2019 (Ex 1, 2, 3, & 4)	Reading each sentence and judging it by writing a correct response
Nakata & Elgort, 2021	Reading each sentence and guessing the meaning of a target item

Production activities (*k* = 9):

Carpenter & Mueller, 2013 (Ex 3 & 4)	Pronouncing each target item
Schuetze, 2014 (Ex 1 & 2)	Writing each target item
Khoii & Abed, 2017	Making a sentence orally using each target item
Suzuki, 2017	Using a target item orally to describe an animation video
Rogers & Cheung, 2020a	Performing crossword puzzles or word search
Rogers & Cheung, 2020b	Performing crossword puzzles
Suzuki *et al*., 2020	Describing a picture orally using target item

Combined activities (both comprehension and production-based activities) (*k* = 5):

Bloom & Shuell, 1981	Performing exercises (multiple-choice, fill-in-the-blank, writing target item for L1 meaning given)
Lee & Choe, 2014 (Ex2)	Reading and listening the target item, followed by writing the item and speaking the item in a structured conversation
Miles, 2014	Sentence completion: completing a sentence and sharing with a partner Correcting each sentence: writing the correct one Translation (L1-L2)
Suzuki & DeKeyser, 2017a	Listening a sentence and choosing a correct picture, followed by practicing using a target item orally to describe a picture as well as a 10-second video

Serrano & Huang, 2018	Reading a passage, while listening to it at the same time, followed by comprehension questions (true/false, multiple-choice, fill-in-the-blank)

Note. k = Number of study experiments.

Appendix S5: Coding Reliability

Calculation of Cohen's (1960) kappa was performed according to the following formula: $\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$

Cohen's kappa (κ) is a useful statistic for either interrater or intrarater reliability testing. Pr(*a*) refers to the actual observed agreement and Pr(*e*) refers to expected agreement. Pr(*e*) is obtained through the following formula (see McHugh, 2012 for details):

$$\text{Expected (Chance) Agreement} = \frac{\left(\frac{cm^1 \times rm^1}{n}\right) + \left(\frac{cm^2 \times rm^2}{n}\right)}{n}$$

cm^1 refers to column 1 marginal, cm^2 represents column 2 marginal, rm^1 refers to row 1 marginal, rm^2 represents row 2 marginal, and *n* refers to the number of observations. Cohen (1960) suggested the Kappa result be interpreted as follows: value of Kappa ≤ 0 as indicating no agreement, 0.01−0.20 as none to slight, 0.21−0.39 as minimal, 0.40−0.59 as weak, 0.60−0.79 as moderate, 0.80−0.90 as strong, and 0.91−1.00 as almost perfect agreement. There are 48 experiments from 37 studies, and each experiment includes 9 coding variables (age, learning target, number of

sessions, type of practice, activity type, provision of feedback, feedback timing, frequency of practice, and retention interval). Table S5.1 presents data in table format for kappa calculation of a variable *age,* and Table S5.2 shows the results of kappa values for 9 variables. Average kappa value for the current study is .95, which indicates very strong agreement (almost perfect agreement, Cohen, 1960).

Table S5.1.

Data for kappa calculation example (Age)

		Rater 1		Row Marginals	
		young	adult		
Rater 2	young	10	0	10	rm^1
	adult	0	38	38	rm^2
Column Marginals		10	38	48	*n*
		cm^1	cm^2		

Pr(*a*): (10 + 38)/48 = 1.00

Pr(*e*): (((10 x 10)/48 + (38 x 38)/48)) / 48 = (2.08 + 30.08) / 48 = .67

Kappa ($_K$) = (1 − .67) / (1 − .67) = .33 / .33 = 1.00 (perfect agreement)

Standard error of kappa (SE_K) = 0 $SE_\kappa = \sqrt{\frac{p(1-p)}{n(1-p_e)^2}}$ p represents Pr(*a*) and p_e represents Pr(*e*).

Table S5.2.

Kappa value matrix for each variable

Variable	1	2	3	4	5	6	7	8	9
1	1.00 (0)								
2		1.00 (0)							
3			1.00 (0)						
4				1.00 (0)					
5					1.00 (0)				
6						1.00 (0)			
7							.96 (.04)		
8								1.00 (0)	
9									.96 (.04)
10									

Average Kappa ($_K$) = 1+1+1+1+.96+1+1+1+.96= 8.92 (total value of kappa) / 9 (the number of variable) = .991 = .99 (almost perfect agreement)

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37−46.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276−282.

Appendix S6: Publication Bias Analyses

In order to ascertain the impact of publication bias on our dataset, we assessed publication bias for two subsets (immediate and delayed posttests) under each of three categories (spaced vs. massed, longer vs. shorter, equal vs. expanding) using funnel plots and Egger's regression test. A funnel plot was plotted with effect size on the X axis and the standard error on the Y axis (see Figures below). The null hypothesis for Egger's test is that symmetry is present in the funnel plot, with the alternative indicating that asymmetry exists in the plot (Egger, Davey Smith, Schneider, & Minder, 1997). For example, the *p*-value for Egger's test is 0.9 (which is not significant, $p > .05$): This confirms that there is no evidence to reject the null hypothesis in favor of the alternative 5% level of significance, and it can be concluded that symmetry is present in the funnel plot and that no publication bias exists in the studies included in the meta-analysis.

Category 1: Spaced vs. Massed

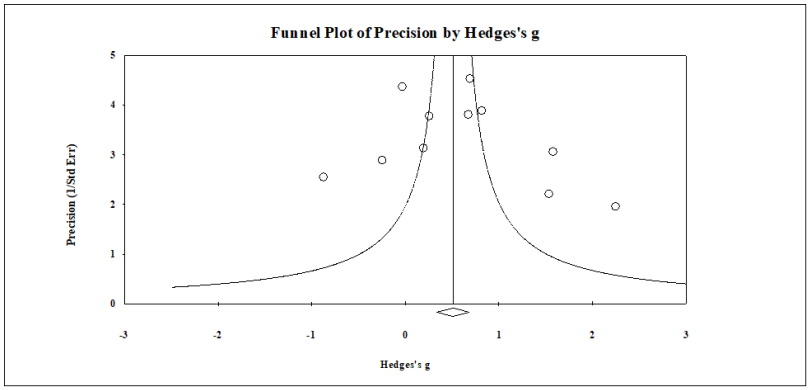


Figure S6.1. Funnel plot for subset 1 (immediate posttest, $k = 11$)

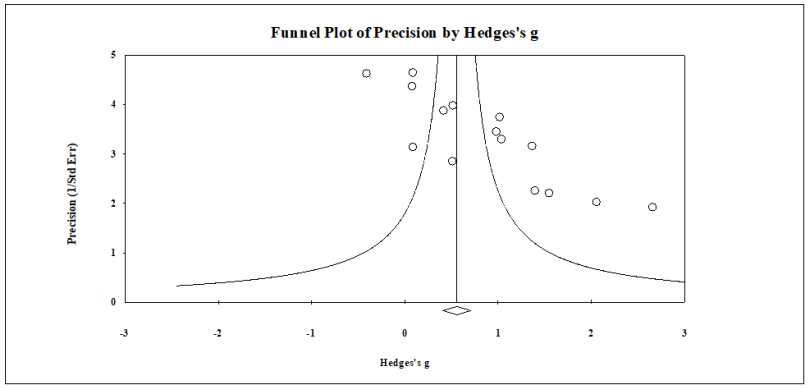


Figure S6.2. Funnel plot for subset 2 (delayed posttest, $k = 15$)

For Figure S6.1, Smaller studies appear toward the top of the graph, and (since smaller studies have more sampling error variation in effect sizes) tend to be spread across a broad range of values. Also, there is a suggestion of missing studies in the middle and right of the plot, in the area of non-significance (i.e., inside the funnel where p > 0.1), making publication bias plausible. This subjective impression may support the presence of asymmetry. Egger's regression test was not significant (t= 0.811, intercept = 2.422, 95% CI = - 4.335 to 9.178, *p* = .219), suggesting that the funnel plot was symmetrical and therefore that there was no publication bias.

For Figure S6.2, studies with small sample sizes, because of their greater sampling error and lower precision values, appear toward the top of the graph. There is a clear downward trend (with studies missing in the middle on the left), suggesting that there is a potential publication bias. Furthermore, the significant Egger's regression test (t = 5.278, intercept = 7.355, 95% CI = 4.345 to 10.366, *p* = .0001) confirmed that the funnel plot was asymmetrical and it is likely that there was an evidence for publication bias. Therefore, the mean effect size in the subset of the delayed posttest outcomes from the spaced versus massed category may be overestimated.

To estimate the mean effect size by taking into account the publication bias, we used the trim and fill method. This method aims to identify and correct for funnel plot asymmetry from publication bias (Duval & Tweedie, 2000).

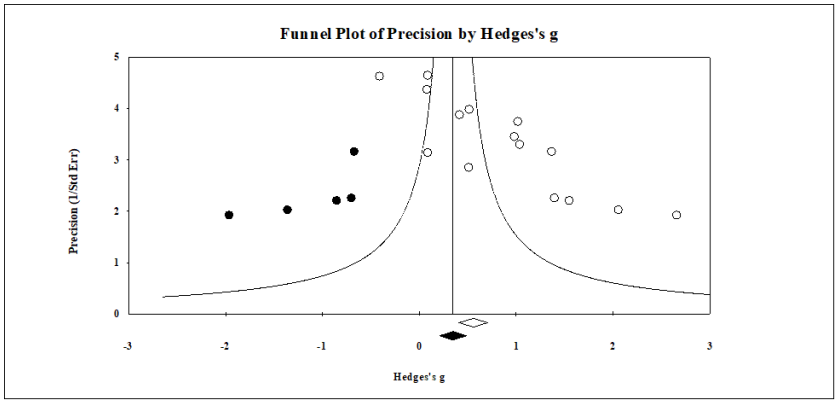


Figure S6.3. Trim-and-fill funnel plot for subset 2 (delayed posttest, spaced vs. massed category) publication bias correction

For Figure S6.3, white points are effect sizes from the included studies, while black points are those added by the trim-and-fill procedure. It was found that five values (hypothetical new outcome effect sizes) on the left side of the mean effect were missing. Imputing would then change the mean effect size from g = 0.804, 95% CI [0.440, 1.168] to g = 0.389, 95% CI [0.001, 0.776]. Consequently, we might consider the overall effects from this subset to provide an inflated estimate of the effect of spaced practice on L2 learning.

It can be argued that imputed intervention effect estimates inappropriately contribute information that reduces the uncertainty in the summary intervention effect (Higgins & Green, 2011; https://handbook-5-1.cochrane.org/chapter_10/10_4_4_2_trim_and_fill.htm).

We, therefore, selected studies with extreme effect sizes (e.g., any absolute value (regardless of whether it was positive or negative) larger than 2.0, Li, 2010) from the subset (delayed posttest in the spaced vs. massed category), followed by explicit summary judgements about the risk of bias assessment by one of authors. Furthermore, we also found one study with effect size larger than 2 from the subset of immediate posttest in the spaced vs. massed category. As a result, three studies (Bahrick & Hall, 2005, Ex 1; Lotfolahi & Salehi, 2016; Koval, 2020) were selected

from the two subsets in the spaced vs. massed category. The mean effect size (Hedges' *g*) in the subset (immediate posttest) from the spaced vs. massed category is 0.579. The mean effect size (Hedges' *g*) in the subset (delayed posttest) from the spaced vs. massed category is 0.804. Bahrick and Hall's (2005) study had effect size of 2.657, Lotfolahi and Salehi's (2016) study had effect size of 2.055, and Koval's (2020) had effect size of 2.247 (1.853 SD, 1.251 SD, and 1.668 SD from the mean effect size, respectively). We applied Cochrane's risk of bias tool to assess the publication bias (Higgins *et al.,* 2016): https://www.riskofbias.info/welcome/rob-2-0-tool. The risk of bias tool is structured into five domains (bias arising from the randomization process, bias due to deviations from intended interventions, bias due to missing outcome data, bias in measurement of the outcome, and bias in selection of the reported result). We had randomized parallel-group trial design template for addressing these domains. We then responded signaling questions (e.g., Was the allocation sequence random?, Were participants aware of their assigned intervention during the trial?, Were data for this outcome available for all, or nearly all, participants randomized?, and Could measurement of the outcome have differed between intervention groups?). There are five response options: Yes, Probably yes, Probably no, No, and No information (see the details for the bias domains, Higgins & Altman, 2008; Higgins *et al*., 2016). The tool recommended by Cochrane for assessing risk of bias produced an overall judgement of risk of bias for the results being assessed. The overall judgement for each study is derived from assessments of individual bias domains (Higgins *et al.*, 2011).

Overall risk of bias judgement:

- Low risk of bias: Bias is unlikely to alter the results seriously. The study is judged to be at low risk of bias for all domains for this result.
- Some concerns: The risk of bias that raises some doubts about the results. The study is judged to raise some concerns in at least one domain for this result, but not to be at high risk of bias for any domain
- High risk of bias: Bias may alter the results seriously. The study is judged to be at high risk of bias in at least one domain for this result / The study is judged to have some concerns for multiple domains in a way that substantially lowers confidence in the result

Table S6.1 presents the results of the risk of bias assessment.

Table S6. 1.

Risk of Bias from a Cochrane Review

Study	Bias domain	Algorithm result	Assessor's responses
Bahrick & Hall, 2005, Ex 1	Randomization process	Low	Y/PY/N
	Deviations from intended intervention	Low	Y/Y/PN/PY
	Missing outcome data	Low	PY
	Measurement of the outcome	Low	PN/N/Y/N
	Selective reporting	Some concerns	NI/N/N
	Overall bias	Some concerns	
Lotfolahi & Salehi, 2016	Randomization process	Some concerns	N/PY/N
	Deviations from intended intervention	Low	NI/Y/N/PY
	Missing outcome data	Low	PY
	Measurement of the outcome	Low	PN/N/Y/N
	Selective reporting	Some concerns	NI/N/N
	Overall bias	Some concerns	
Koval, 2020	Randomization process	Low	Y/PY/N
	Deviations from intended intervention	Low	Y/Y/PN/PY
	Missing outcome data	Low	PY
	Measurement of the outcome	Low	PN/N/Y/N

	Selective reporting	Low PY/N/N
	Overall bias	Low risk of bias

Note. Y = Yes, PY = Probably yes, N = No, PN = Probably no, NI = No information

In terms of selective reporting (bias domain), since two studies (Bahrick & Hall, 2005; Lotfolahi & Salehi, 2016) had the response "No information" about the question, "Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalized before unblended outcome data were available for analysis?", the algorithm result for this domain (Selective reporting) showed "some concerns". It would be possible that the studies are judged to be at low risk of bias for all domains for this result if the researchers' (the study investigators) pre-specified intentions are available in sufficient detail and the planned outcome measurements and analyses can be compared with those presented in the published report. To avoid the possibility of selection of the reported results, finalization of the analysis intentions must precede availability of unblinded outcome data to the study investigators (Higgins *et al.*, 2016). This may be challenging to judge published reports (without pre-registration) regarding the domain of selective reporting. Although algorithm result showed some concerns, the assessor's (one of authors who applied the risk of bias assessment tool) result would be "low risk of bias" for this domain. However, note that Koval (2020) is a Ph.D. thesis, which is carefully designed and provide detailed information on research methodology and statistical analyses. For this selective reporting domain, algorithm result showed low risk of bias, and the assessor's (one of authors who applied risk of bias assessment) result would be "low risk of bias". To conclude, two studies were judged to raise some concerns in one domain about the results, but to be at low risk of bias for the rest of the domains. Although three studies (Bahrick & Hall, 2005, Ex 1; Lotfolahi & Salehi, 2016; Koval, 2020) had large effect sizes, including them in this meta-analysis is unlikely to alter the results seriously.

Furthermore, van Aert, Wicherts, and van Assen (2016) recommended that in case of evidence of publication bias researchers need to report results of *p*-uniform (Recommendation 3, p. 714). We used a *p*-uniform* web application: https://rvanaert.shinyapps.io/p-uniformstar, retrieved on 21st of June, 2020 (see van Aert & van Assen, 2018; van Aert *et al*., 2016). *P*-uniform* is an improvement over *p*-uniform, because *p*-uniform* enables estimating and testing of the extent of heterogeneity and considers the significant and non-significant effect sizes (van Aert & van Assen, 2018); Therefore, we can have effect sizes not only conditional on significance but also on non-significance. Two different meta-analytic estimates (*p*-uniform and random-effects) of the mean effect size underlying the delayed posttest in the spaced vs. massed category are presented in Table S6.2.

Table S6.2.

Results of p-Uniform and Random-Effects Meta-Analysis (for Delayed Posttest in the Spaced vs. Massed Category, k = 15)*

	p-Uniform*	Random-effects
Effect-size estimate	0.843	0.804
95% CI	[0.432, 1.253]	[0.440, 1.168]
Test of H_0: $\delta = 0$	$z = 4.026$, $p < .001$	$z = 4.329$, $p < .001$
Publication bias test	$z = 0.749$, $p = .69$	

Note. H_0: $\delta = 0$ refers to the null hypothesis of no effect. CI = confidence interval.

The results showed that the estimate of *p*-uniform* is 0.843, 95%CI [0.432, 1.253] and that random-effects meta-analysis yielded effect size estimate of 0.804, 95%CI [0.440, 1.168], which is statistically significant ($z = 4.329$, $p < .001$). The *p*-uniform*'s estimate (0.843) was larger than the estimate of random-effects (0.804). The *p*-uniform* outperforms random-effects in case the majority of primary studies were statistically significant. This subset involved 9 statistically significant effect sizes (60%, out of 15 studies). The results of *p*-uniform*'s publication bias test suggested no evidence of publication bias ($z = 0.749$, $p = .69$). Consequently, random-effects meta-analysis result may be interpreted as the standard meta-analytic estimates (see Recommendation 3, van Aert *et al.*, 2016). Furthermore, effect size of 0.843 from the *p*-uniform* could be considered medium-to-large, given that the 95% CI did not include zero. This size of effect was not different from the estimate of random-effects (g = 0.804, with 95% CI far above zero) which could be also considered medium-to-large.

Category 2: Longer vs. Shorter

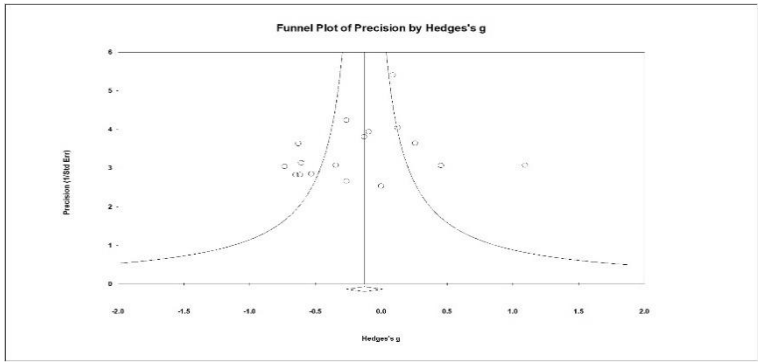


Figure S6.4. Funnel plot for subset 1 (immediate posttest, $k = 17$)

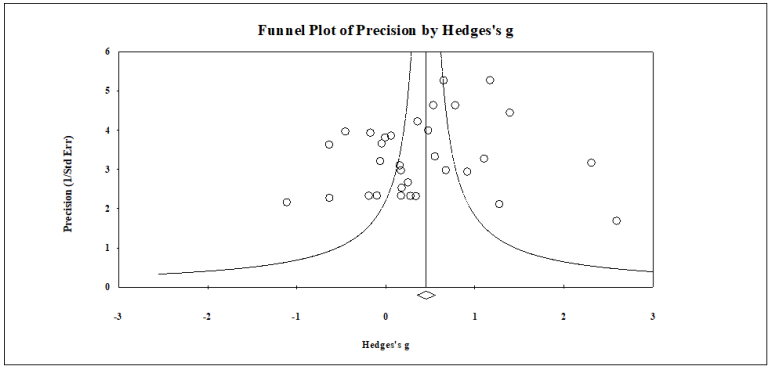


Figure S6.5. Funnel plot for subset 2 (delayed posttest, $k = 32$)

For Figure S6.4 (immediate posttest, longer vs. shorter category), Egger's regression test was not significant, indicating no evidence of publication bias (t = 0.964, intercept = - 1.755, 95% CI = - 5.634 to 2.124, p = .350). For Figure S6.5 (delayed posttest, longer vs. shorter category), Egger's regression test was not significant, suggesting that the funnel plot was symmetrical and therefore that there was no publication bias (t = 0.963, intercept = - 1.429, 95% CI = -4.461 to 1.603, p = .172).

Category 3: Equal vs. Expanding

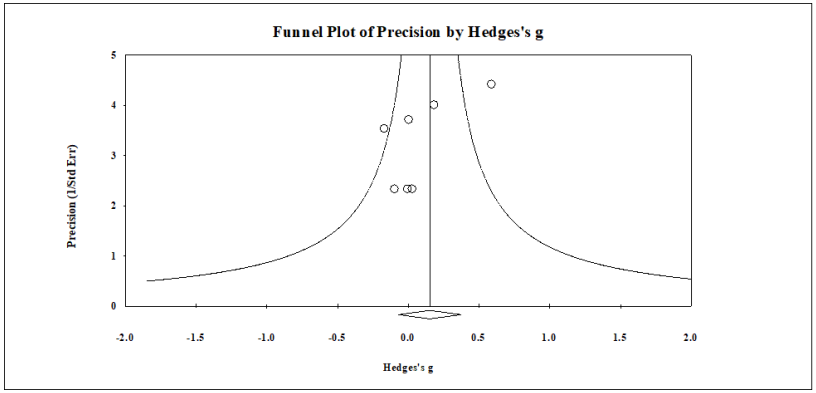


Figure S6.6. Funnel plot for subset 1 (immediate posttest, k = 7)

For Figure S6.6 (immediate posttest, equal vs. expanding category), there is a suggestion of missing studies in the middle and the right of the plot. Although funnel plot is used to detect bias in studies included in the meta-analysis, assessment of symmetry in the funnel plot is often subjective and difficult to identify publication bias, particularly if the number of studies is small (less than 10) (Terrin, Schmid, & Lau, 2005). Egger's regression test was not significant, indicating no evidence of publication bias (t = 1.521, intercept = -2.095, 95% CI = - 5.634 to 1.445, *p* = .094). However, the effect size was homogeneous (I^2 = 0, tau = 0). van Aert *et al.* (2016, recommendation 5a) recommended that researchers need to interpret the estimates of *p*-uniform as estimates of the average population effect size if the effect size is homogeneous or if the heterogeneity is small (I^2 < 0.5). We used a *p*-uniform* web application: https://rvanaert.shinyapps.io/p-uniformstar. Two different meta-analytic estimates (*p*-uniform* and random-effects) of the mean effect size underlying the immediate posttest in the equal vs. expanding category are presented in Table S6.3.

Table S6.3.

Results of p-Uniform and Random-Effects Meta-Analysis (for Immediate Posttest in the Equal vs. Expanding Category, k = 7)*

	p-Uniform*	Random-effects
Effect-size estimate	0.053	0.151
95% CI	[-0.217, 0.340]	[-0.070, 0.373]
Test of H_0: $\delta = 0$	z = 0.141, p = .71	z = 1.337, p = .18

Publication bias test	$z = 0.526$, $p = .77$

Note. H_0: $\delta = 0$ refers to the null hypothesis of no effect. CI = confidence interval.

The results of *p*-uniform*'s publication bias test showed $z = 0.526$, $p = .769$, suggesting no evidence of publication bias. In this homogeneous subset, small percentages of statistically significant primary effect sizes (14.3%, one significant effect size out of 7 studies) led to the conclusion that evidence for publication bias is weak. The estimate of *p*-uniform* is 0.053, 95%CI [-0.217, 0.340] and random-effects meta-analysis yielded effect size estimate of 0.151, 95%CI [-0.070, 0.373]. The 95% CIs from the both estimates (*p*-uniform and random-effects) crossed zero, which indicated that there was no significant difference between experimental (equal spacing) and control/baseline (expanding spacing) groups. Although we might consider the overall effect from this subset to provide a slightly inflated estimate of the effect of spaced practice on L2 learning (*p*-uniform = 0.05, random-effects = 0.15), both *p*-uniform* effect size and random effects show very small effects and their 95% CIs crossed zero. Consequently, random-effects meta-analysis result may be interpreted as the standard meta-analytic estimates.

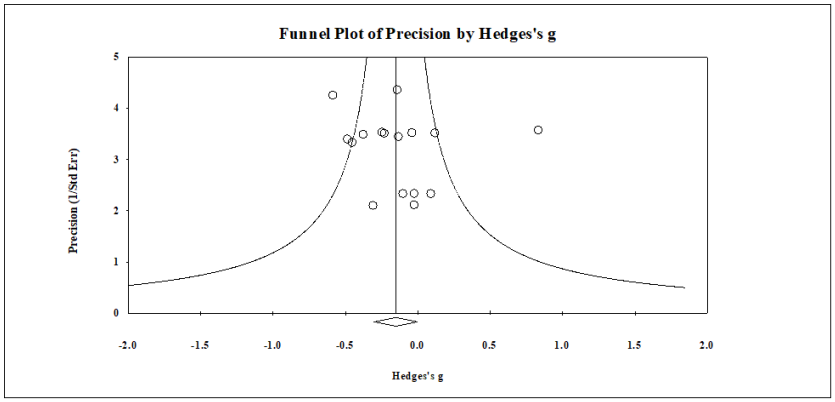


Figure S6.7. Funnel plot for subset 2 (delayed posttest, k = 16)

For Figure S6.7 (delayed posttest, equal vs. expanding category), Egger's regression test was not significant, suggesting that the funnel plot was symmetrical and therefore that there was no publication bias (t = 0.512, intercept = 0.720, 95% CI = - 2.293 to 3.732, p = .308).

Overall, publication bias analyses from Egger's test indicated that apparent bias exists in the subset of delayed posttest from the spaced vs. massed category. However, the results of *p*-uniform* showed that the bias is negligible. In the subset of immediate posttest from the equal vs. expanding category, I^2 and *tau* were zero, which was recommended interpreting the estimates of *p*-uniform*. However, the results of both p-uniform* and random-effects were similar (very small effects with 95% CIs crossed zero), it led to the conclusion that random-effects meta-analysis results may be interpreted as the standard meta-analytic estimates. Since most studies included in this meta-analysis were published

studies (published = 35, conference proceeding = 1, and PhD thesis = 1), a symmetrical distribution may not rule out publication bias. Therefore, we need to consider the overall effects from the current meta-analyses to provide a slightly inflated estimate of the effect of spaced practice on L2 learning.

References

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52,* 566−577. http://doi.org/10.1016/j.jml.2005.01.012

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ, 315, 629–634. http://doi.org/10.1136/bmj.315.7109.629

Higgins, J. P. T., & Altman, D. G. (2008). Assessing risk of bias in included studies. In J. P. T. Higgins & S. Green (Eds.),

Cochrane Handbook for Systematic Reviews of Interventions (pp. 187-241). Chichester, UK: John Wiley & Sons.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011). The Cochrane collaboration's tool for assessing risk of bias in randomized trials. *BMJ, 343:d5928* http://doi.org/10.1136/bmj.d5928

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from www.handbook.cochrane.org.

Higgins, J., Sterne, J., Savović, J., Page, M. J., Hróbjartsson, A., Boutron, I., Reeves, B., & Eldridge, S. (2016). A revised tool for assessing risk of bias in randomized trials. In J. Chandler, J. McKenzie, I. Boutron & V. Welch (Eds.), *Cochrane Methods. Cochrane Database of Systematic Reviews*, *10*. http://doi.org/10.1002/14651858.CD201601

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*(2), 309–365.

Lotfolahi, A. R., & Salehi, H. (2016). Learners' perceptions of the effectiveness of spaced learning schedule in L2 vocabulary learning. *SAGE Open, 6*(2), 1−9. http://doi.org/10.1177/2158244016646148

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology, 58*(9), 894–901. http://doi.org/10.1016/j.jclinepi.2005.01.006

van Aert, R. C. M., & van Assen, M. A. L. M. (2018, October 2). P-uniform*. http://doi.org/10.31222/osf.io/zqjr9

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*(5), 713–729. http://doi.org/10.1177/1745691116659874

Appendix S7: Overall Results Under Each Category

Table S7.1.

Overall Results of Comparative effects (Immediate and Delayed Posttests)

Category	Variables	*k*	*g*	*p*	SE	95% CI		Between-group Contrast		*Tau*	I^2	Tau^2	PI
						Lower	Upper	Q_b	*p*				
Spaced vs. Massed													
	Immediate	11	0.58	.01	0.21	0.16	1.00	54.72	.00	0.631	81.72	0.398	-0.93, 2.09
	Delayed	15	0.80	.00	0.19	0.44	1.17	79.83	.00	0.639	82.46	0.409	-0.64, 2.24
Longer vs. Shorter													
	Immediate	17	-0.15	.16	0.11	-0.37	0.06	37.07	.00	0.332	56.84	0.111	-0.90, 0.60
	Delayed	32	0.40	.00	0.12	0.16	0.64	163.63	.00	0.607	81.05	0.369	-0.87, 1.67
Equal vs. Expanding													
	Immediate	7	0.15	.18	0.11	-0.07	0.37	5.90	.43	0	0	0	-0.14, 0.44
	Delayed	16	-0.15	.11	0.09	-0.33	0.03	20.60	.15	0.188	27.19	0.035	-0.60, 0.30

Note. PI = prediction interval

*Statistically significant at $p < .05$.

Heterogeneity in effect sizes

We reported *Q*-value, *I*-squared (what proportion), *Tau* (the standard deviation of true effects), *Tau*-squared (the variance of true effects), and prediction interval (how widely the effect sizes vary across studies), which are intended to quantify heterogeneity (what the distribution of effects looks like) (Borenstein, Higgins, Hedges & Rothstein, 2017). In the current meta-analysis, we used Hedges' *g* as effect sizes. Therefore, we need to convert all the numbers to a common metric before computing the prediction interval. For this reason, we used software (from www.meta-analysis-workshop.com) to compute the interval, and then report the prediction interval for each comparison (spaced vs. massed, longer vs. shorter, and equal vs. expanding).

Spaced vs. Massed

In the immediate effects, the effect size of spaced practice on L2 learning was 0.58, and the confidence interval for the spacing effects was 0.16 to 1.00. In the delayed effects, the effect size of spaced practice on L2 learning was 0.80, and the confidence interval for the spacing effects was 0.44 to 1.17. Each of these ranges did not include an effect size of zero, which indicates that the mean effect size is probably not zero. Our finding suggests that spaced practice has medium to large effects on L2 learning. The I^2 statistics (approximately 82% in both immediate and delayed effects) indicates that 82% of the variance in observed effects reflects variance in true effects rather than sampling error. The variance of true effects (Tau^2) was 0.398 and the standard deviation of true effects (*Tau*) was 0.631. The prediction interval was -0.927 to 2.087 for the

immediate effects and -0.641 to 2.241 for the delayed effects. We would predict that the true effect sizes would fall in these wide ranges, and we would be correct 95% of the time. This indicates that effects of spaced practice on L2 learning can vary: there would be some populations where the impact of spaced practice on L2 learning is very small, some where it is very large, or some where there is no spacing effect. It makes sense to apply moderator analyses or meta-regression to explain the variance (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Longer vs. Shorter

In the immediate effects, longer spacing was as effective as shorter spacing in L2 learning (g = -0.15, the confidence interval is -0.37 to 0.06). However, in the delayed effects, longer spacing was more effective than shorter spacing in L2 learning (g = 0.40, the confidence interval is 0.16 to 0.64). The range in the delayed effects did not include an effect size of zero, which indicates that the mean effect size is probably not zero. Our finding suggests that the effect of longer spacing has small to medium effects on L2 learning. The I^2 statistics (approximately 81% in the delayed effects) indicates that 81% of the variance in observed effects reflects variance in true effects rather than sampling error. The variance of true effects (Tau^2) is 0.369 and the standard deviation of true effects (Tau) is 0.607. The prediction interval is -0.866 to 1.666 for the delayed effects. We would predict that the true effect sizes would fall in this wide range, and we would be correct 95% of the time. This indicates that effects of spaced practice on L2 learning can vary: there would be some populations where the impact of spaced practice on L2 learning is very small, some where it is very large, or some where there is no spacing effect. It makes sense to apply moderator analyses or meta-regression to explain the variance (Borenstein *et al.*, 2009).

Equal vs. Expanding

Equal spacing was as effective as expanding spacing in L2 learning (g = 0.15, CI -0.07, 0.37 for the immediate effects and g = -0.15, CI -0.33, 0.03 for the delayed effects). The I^2 statistics (approximately 27% in the delayed effects) indicates that 27% of the variance in observed effects reflects variance in true effects rather than sampling error. The variance of true effects (Tau^2) is 0.035 and the standard deviation of true effects (*Tau*) is 0.188. The prediction interval is -0.597 to 0.297 for the delayed effects. We would predict that the true effect sizes would fall in this range, and we would be correct 95% of the time. This indicates that there would be some populations where the impact of relative spacing (either equal or expanding spacing) on L2 learning is very small, some where it is large, or some where there is no effect. It makes sense to apply moderator analyses or meta-regression to explain the variance (Borenstein *et al.*, 2009).

Table S7.2

Summary of Effects for Receptive and Productive knowledge

						95% CI		Group Contrast			
Category	Variables	*k*	*g*	*p*	SE	Lower	Upper	Q_b	*p*	*Tau*	I^2
Spaced vs. Massed	Immediate										
	Receptive	8	0.62	.05*	0.31	0.02	1.22	48.14	.00*	0.793	85.46
	Productive	5	0.35	.17	0.26	-0.15	0.85	15.47	.00*	0.491	74.15
	Delayed										

	Receptive	13	0.88	.00*	0.22	0.45	1.31	78.94	.00*	0.707	84.70
	Productive	5	0.42	.01*	0.15	0.12	0.72	5.56	.23	0.180	28.07
Longer vs. Shorter	Immediate										
	Receptive	13	-0.10	.31	0.10	-0.30	0.10	19.19	.08	0.220	37.46
	Productive	7	-0.19	.23	0.16	-0.50	0.12	12.49	.05*	0.299	51.95
	Delayed										
	Receptive	21	0.35	.02*	0.15	0.06	0.65	109.09	.00*	0.605	81.67
	Productive	14	0.33	.06	0.17	-0.01	0.66	61.63	.00*	0.559	78.91
Equal vs. Expanding	Immediate										
	Receptive	7	0.13	.27	0.11	-0.10	0.35	4.20	.65	0	0
	Productive	4	-0.05	.78	0.17	-0.37	0.28	0.12	.99	0	0
	Delayed										
	Receptive	14	-0.15	.17	0.11	-0.36	0.06	20.03	.09	0.234	35.10
	Productive	5	-0.10	.48	0.15	-0.39	0.18	2.04	.73	0	0

*Statistically significant at $p < .05$.

As in Shintani, Li, and Ellis (2013), the dependent variables were receptive and productive L2 knowledge. Receptive knowledge was measured through receptive tests such as a paired-associate receptive retrieval format (e.g., writing the L1 meaning of an L2 words), multiple-choice or grammaticality judgement tests. Productive knowledge was measured through productive tests such as a paired-associate productive retrieval format (e.g., writing an L2 word corresponding to L1 word), describing pictures by using target features, or pronouncing target items. Detailed information is presented in Table S4.1, Appendix S4 (see posttest format).

In the spaced vs. massed comparison, spacing effect was larger on the acquisition of L2 receptive knowledge (g = 0.62 for immediate effects, g = 0.88 for delayed effects) than on the acquisition of L2 productive knowledge (g = 0.35 for immediate effects, g = 0.42 for delayed effects). In the longer vs. shorter comparison, we found that longer spacing and shorter spacing were similarly effective in immediately developing both receptive and productive knowledge. In the long term, while both longer and shorter spacing were effective in developing productive knowledge, longer spacing led to more durable receptive knowledge than shorter spacing (g = 0.35, CI = 0.06, 0.65). In the equal vs. expanding comparison, however, we found that both equal and expanding spacing were effective in developing receptive and productive knowledge. Overall, our findings suggest that spaced practice benefits the developments of both receptive and productive L2 knowledge, but the benefits were larger for receptive knowledge. Furthermore, longer spacing enhances greater retention than shorter spacing for developing receptive knowledge developing. However, developing productive L2 knowledge is not sensitive to type of spacing. More L2 research examining the effects of spaced practice for developing productive knowledge is needed.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* UK: Wiley.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(5), 5-18.

Appendix S8: Moderator Analyses for Each Posttest (Immediate and Delayed) Under Each Category

Table S8.1.

Moderator Analysis for Categorical Variables (Spaced vs. Massed) with Immediate Posttests (k = 11)

	k	*g*	Variance	95% CI		*p*	*Q* tests	
				Lower	Upper		*Q*	*p*
Age							0.30	.58
Young	3	0.39	0.03	-0.44	1.22	.36		
Adult	8	0.66	0.10	0.13	1.20	.01*		
Learning target							1.71	.19
Vocabulary	8	0.76	0.08	0.26	1.25	.00*		
Grammar	3	0.14	0.08	-0.41	0.92	.72		
Number of sessions							5.86	.02*
Single session	6	1.04	0.10	0.49	1.59	.00*		
Multiple sessions	5	0.04	0.06	-0.55	0.63	.88		
Type of practice							1.34	.72
Test-restudy (all) trial	6	0.69	0.13	0.05	1.34	.04*		
Test-restudy (not recalled) trial	2	0.48	0.05	-0.06	1.55	.39		
Study trial	2	0.81	0.45	-0.34	1.97	.17		
Test trial	1	-0.25	0.12	-1.84	1.35	.76		
Activity type							1.91	.59
Paired associate	3	0.67	0.60	-0.29	1.63	.17		
Comprehension activities	3	0.97	0.37	0.04	1.91	.04*		

Production activities	2	0.68	0.03	-0.42	1.78	.22		
Combined activities	3	0.07	0.03	-0.85	1.00	.88		
Provision of feedback							1.32	.25
Absence	2	0.02	0.06	-0.99	1.02	.98		
Presence	7	0.69	0.09	0.15	1.23	.01*		
Feedback timing							6.47	.04*
Immediate	6	0.90	0.07	0.39	1.42	.00*		
Delayed	1	-0.88	0.15	-2.24	0.49	.21		

Note. Feedback timing was not reported in the main manuscript, because only one study involved delayed feedback.

*Statistically significant at $p < .05$.

Spaced vs. Massed with Immediate Posttests

Results showed that number of sessions is the factor that significantly moderates the effects of spaced practice on L2 learning ($Q = 5.86$, $p = .02$): spacing effects were more pronounced when the spaced practice was manipulated within one session ($g = 1.04$, CI = 0.49, 1.59) than when the practice was manipulated between multiple sessions ($g = 0.04$, CI = -0.55, 0.63). We also found that feedback timing significantly moderates the spacing effect ($Q = 6.47$, $p = .04$), spacing effect was larger when immediate feedback was provided ($g = 0.90$, CI = 0.39, 1.42) than when delayed feedback was provided. However, because only one study involved the delayed feedback, it should be careful to interpret the results. It is notable that the spacing effects were larger when the spaced practice was manipulated with comprehension activities than with paired associate learning task (e.g., word cards or word lists) even though activity type did not significantly moderate the effects of spaced practice. While the effect size of spaced practice with comprehension activities was 0.97 (CI = 0.04, 1.91), the spaced practice with paired

associate learning task was 0.67 (CI = -0.29, 1.63; crossed zero, indicating that the effect size could be zero). Although the sample size was small (*k* = 3, in both activities), it is worth considering the interaction between activity type and spaced practice. More research is needed to look at the role of activity types in spaced practice.

Table S8.2.

Moderator Analysis for Categorical Variables (Spaced vs. Massed) with Delayed Posttests (k = 15)

	k	*g*	Variance	95% CI		*p*	*Q* tests	
				Lower	Upper		*Q*	*p*
Age							0.16	.69
Young	3	0.97	0.25	0.11	1.82	.03		
Adult	12	0.77	0.04	0.36	1.18	.00*		
Learning target							13.78	.00*
Vocabulary	10	1.15	0.04	0.81	1.49	.00*		
Grammar	5	0.11	0.03	-0.32	0.54	.61		
Number of sessions							1.91	.17
Single session	9	0.61	0.05	0.16	1.05	.01*		
Multiple sessions	6	1.12	0.10	0.55	1.69	.00*		
Type of practice							3.35	.34
Test-restudy (all) trial	10	0.70	0.05	0.25	1.14	.00*		
Test-restudy (not recalled) trial	2	1.73	0.65	0.67	2.79	.00*		

	Study trial	2	0.69	0.43	-0.36	1.73	.20	
	Test trial	1	0.51	0.12	-0.93	1.95	.49	
Activity type							6.26	.10
	Paired associate	6	1.36	0.08	0.80	1.92	.00*	
	Comprehension activities	5	0.43	0.10	-0.15	1.00	.14	
	Production activities	1	0.42	0.07	-0.85	1.68	.52	
	Combined activities	3	0.53	0.07	-0.23	1.29	.52	
Provision of feedback							0.00	.95
	Absence	3	0.85	0.03	0.02	1.68	.05*	
	Presence	10	0.82	0.06	0.36	1.27	.00*	
Feedback timing							10.40	.00*
	Immediate	8	0.52	0.05	0.10	0.94	.02*	
	Delayed	2	2.35	0.13	1.36	3.34	.00*	

*Statistically significant at $p < .05$.

Spaced vs. Massed with Delayed Posttests

Results showed that learning target and feedback timing significantly moderate the effects of spaced practice on the retention of L2 target items. Spaced practice was more effective for L2 vocabulary learning (g = 1.15, CI = 0.81, 1.49) than for L2 grammar learning (g = 0.11, CI = -0.32, 0.54). Spacing effect was larger when delayed feedback was provided (g = 2.35, CI = 1.36, 3.34) than when immediate feedback was provided (g = 0.52, CI = 0.10, 0.94). However, it should be careful to interpret the results from the feedback timing due to small sample size (k = 2). It is also notable that the result from activity type in the delayed effect (from the delayed posttest scores) showed a different pattern form that in the

immediate effect (from the immediate posttest scores): while spaced practice with comprehension activities (g = 0.97, CI = 0.04, 1.91) was more effective for the immediate learning of L2 target items than that with paired associate learning task (g = 0.67, CI = -0.29, 1.63), paired associate task (g = 1.36, CI = 0.80, 1.92) was more effective than comprehension activities (g = 0.43, CI = -0.15, 1.00) for the retention of L2 target items. This suggests that comprehension activities in the spaced practice benefit L2 learning but the paired associate task enhances the retention of the L2 items.

Table S8.3.

Moderator Analysis for Continuous Variables (Spaced vs. Massed)

				95% CI		
	k	*Q*	*B*	Lower	Upper	*p*
Frequency of practice (Immediate posttests)	11	0.02	0.0137	-0.1713	0.1987	.88
Frequency of practice (Delayed posttests)	15	0.27	-0.0383	-0.1819	0.1054	.60
Retention interval	15	0.16	0.0069	-0.0269	0.0407	.69

Note. CI = confidence interval

Table S8.4.

Moderator Analysis for Categorical Variables (Longer vs. Shorter) with Immediate Posttests (k = 17)

	k	*g*	Variance	95% CI		*p*	*Q* tests	
				Lower	Upper		*Q*	*p*
Age							0.45	.50
Young	3	-0.03	0.04	-0.42	0.37	.89		
Adult	14	-0.19	0.02	-0.44	0.06	.14		
Learning target							15.59	.00*
Vocabulary	9	0.14	0.02	-0.11	0.38	.28		
Grammar	4	-0.41	0.02	-0.70	-0.13	.01*		
Pronunciation	4	-0.64	0.03	-0.98	-0.30	.00*		
Number of sessions							0.78	.38
Single session	10	-0.08	0.03	-0.40	0.23	.60		
Multiple sessions	7	-0.27	0.02	-0.52	-0.01	.04*		
Type of practice							11.74	.01*
Test-restudy (all) trial	6	0.22	0.02	-0.08	0.51	.16		
Test-restudy (not recalled) trial	3	-0.54	0.03	-0.89	-0.18	.00*		
Study trial	5	-0.41	0.05	-0.86	0.04	.07		
Study-test trial	3	-0.24	0.02	-0.54	0.07	.13		
Activity type							13.75	.00*
Paired associate	7	0.17	0.02	-0.11	0.45	.24		
Comprehension activities	4	-0.38	0.03	-0.69	-0.07	.02*		
Production activities	3	-0.64	0.03	-0.99	-0.28	.00*		
Combined activities	3	-0.15	0.08	-0.71	0.41	.60		
Provision of feedback							2.18	.34

	Absence	1	-0.26	0.06	-0.73	0.20	.26		
	Presence	11	-0.04	0.02	-0.30	0.21	.75		
Feedback timing								10.41	.02*
	Immediate	10	-0.03	0.02	-0.30	0.25	.85		
	Delayed	1	-0.26	0.14	-1.00	0.47	.48		

Notes. Provision of feedback in the longer vs. shorter comparison (with immediate posttest data) was not reported in the main manuscript, because there was only one study that did not involve feedback (k = 1 for absence, k = 11 for presence). Feedback timing was not reported in the main manuscript, because only one study involved delayed feedback.

*Statistically significant at p <.05.

Table S8.5.

Moderator Analysis for Categorical Variables (Longer vs. Shorter) with Delayed Posttests (k = 32)

		k	*g*	Variance	95% CI		*p*	*Q* tests	
					Lower	Upper		*Q*	*p*
Age								4.35	.04*
	Young	8	-0.04	0.03	-0.52	0.44	.86		
	Adult	24	0.54	0.02	0.27	0.81	.00*		
Learning target								0.54	.76

	Vocabulary	22	0.34	0.02	0.04	0.64	.03*	
	Grammar	8	0.56	0.07	0.06	1.06	.03*	
	Pronunciation	2	0.42	0.06	-0.57	1.42	.41	
Number of sessions							6.83	.01*
	Single session	11	0.76	0.04	0.42	1.11	.00*	
	Multiple sessions	21	0.18	0.02	-0.10	0.45	.21	
Type of practice							15.86	.00*
	Test-restudy (all) trial	16	0.38	0.02	0.10	0.67	.01*	
	Test-restudy (not recalled) trial	6	1.06	0.09	0.61	1.50	.00*	
	Study trial	6	-0.12	0.06	-0.62	0.38	.64	
	Study-test trial	3	0.40	0.06	-0.23	1.03	.22	
Activity type							10.72	.01*
	Paired associate	12	0.58	0.05	0.23	0.93	.00*	
	Comprehension activities	9	0.73	0.03	0.31	1.15	.00*	
	Production activities	8	-0.24	0.03	-0.72	0.24	.32	
	Combined activities	3	0.16	0.03	-0.55	0.86	.66	
Provision of feedback							0.71	.40

	Absence	4	0.24	0.12	-0.41	0.89	.47	
	Presence	23	0.55	0.02	0.27	0.82	.00*	
Feedback timing								2.83 .09
	Immediate	15	0.39	0.03	0.08	0.71	.01*	
	Delayed	8	0.87	0.06	0.41	1.34	.00*	

*Statistically significant at $p < .05$.

Longer vs. Shorter with Immediate and Delayed Posttests

Results showed that learning target significantly moderated the effects of absolute spacing (longer vs. shorter) on the immediate learning of L2 items. Shorter spacing was more beneficial than longer spacing in L2 pronunciation learning (g = -0.64, CI = -0.98, -0.30). However, the pattern was different in the delayed effects (from the delayed posttest scores). Longer spacing was more beneficial than shorter spacing in L2 grammar learning (g = 0.56, CI = 0.06, 1.06) than L2 vocabulary (g = 0.34, CI = 0.04, 0.64) and pronunciation (g = 0.42, CI = -0.57, 1.42) learning. Although it did not reach statistical significance, it should be notable that there was an interaction between lag (different length of spacing; shorter or longer) and learning target (vocabulary, grammar, or pronunciation). Results from the delayed posttests scores (for the delayed effects) showed that age, number of sessions, type of practice, and activity type significantly moderate the effects of absolute spacing on L2 learning. First, the effect of longer spacing was more pronounced for adult learners (g = 0.54, CI = 0.27, 0.81) than young learners. However, it should be noted that sample size for young learners was smaller (k = 8) than that for adult learners (k = 24). Second, the effect of longer spacing was larger

when the spaced practice was manipulated within a single session (g = 0.76, CI = 0.42, 1.11) than when the practice was manipulated between multiple sessions (g = 0.18, CI = -0.10, 0.45). This suggests that single session (spaced practice manipulated within one session) benefits the retention of L2 items more than multiple sessions (practice manipulated between multiple sessions). Third, spaced practice with test-restudy trials (g = 0.38~0.80, CI = 0.10, 1.49) were more beneficial than practice with the other trials (study-only trials, test-only trials, and study-test trials). Longer spacing was more beneficial than shorter spacing in the spaced practice with test-restudy trials. Lastly, the effect of longer spacing was larger than the effect of shorter spacing in the spaced practice with paired associate learning task and the practice with comprehension activities: the longer spacing effect was greater when the practice was manipulated with comprehension activities (g = 0.73, CI = 0.31, 1.15) than with paired associate learning task (g = 0.58, CI = 0.23, 0.93). This suggests that spaced practice with comprehension activities benefits the retention of L2 items.

Table S8.6.

Moderator Analysis for Continuous Variable (Longer vs. Shorter)

				95% CI		
	k	*Q*	*B*	Lower	Upper	*p*
Frequency of practice (Immediate posttests)	17	0.81	-0.0156	-0.0497	0.0185	.37
Frequency of practice (Delayed posttests)	30	1.41	-0.0293	-0.0778	0.0191	.24
Retention interval	31	0.02	-0.0015	-0.0200	0.0170	.87

Notes. CI = confidence interval. In the moderator analyses for RI, Serrano and Huang's (2018) study was excluded due to different delayed posttest time point (RI was manipulated between participants): Shorter spacing condition involved 4-day delayed posttest and longer spacing condition involved 28-day delayed posttest based on the optimal ISI/RI ratio.

Table S8.7.

Moderator Analysis for Categorical Variables (Equal vs. Expanding) with Immediate Posttests (k =7)

	k	*g*	Variance	95% CI		*p*	*Q* tests	
				Lower	Upper		*Q*	*p*
Age							2.18	.14
Young	2	0.35	0.09	0.01	0.69	.05*		
Adult	5	0.01	0.02	-0.29	0.30	.96		
Number of sessions							0.25	.62
Single session	4	0.07	0.03	-0.29	0.44	.70		
Multiple sessions	3	0.19	0.06	-0.12	0.51	.23		
Type of practice							4.95	.08
Test-restudy (all) trial	5	0.01	0.02	-0.29	0.30	.96		
Test-restudy (not recalled) trial	1	0.59	0.05	0.14	1.03	.01*		
Study-test trial	1	0.00	0.07	-0.53	0.53	.99		
Activity type							4.95	.08
Paired associate	5	0.01	0.02	-0.29	0.30	.96		
Comprehension activities	1	0.00	0.07	-0.53	0.53	.99		
Production activities	1	0.59	0.05	0.14	1.03	.01*		
Provision of feedback							1.55	.21
Absence	1	-0.17	0.08	-0.73	0.38	.54		

Presence	6	0.21	0.02	-0.03	0.45	.09

Notes. Learning target was excluded because all the studies (k = 7) involved vocabulary. Feedback timing was not reported because all the studies (k = 6) involved immediate feedback.

*Statistically significant at p <.05.

Table S8.8.

Moderator Analysis for Categorical Variables (Equal vs. Expanding) with Delayed Posttests (k =16)

	k	*g*	Variance	95% CI		*p*	*Q* tests	
				Lower	Upper		*Q*	*p*
Age							13.42	.00*
Young	1	0.83	0.08	0.28	1.38	.00*		
Adult	15	-0.23	0.01	-0.39	-0.08	.00*		
Number of sessions							0.68	.41
Single session	6	-0.04	0.02	-0.35	0.28	.81		
Multiple sessions	10	-0.20	0.02	-0.42	0.02	.08		
Type of practice							15.33	.00*
Test-restudy (all) trial	8	-0.32	0.01	-0.54	-0.10	.00*		
Test-restudy (not recalled) trial	3	-0.05	0.03	-0.37	0.27	.76		
Study trial	2	-0.17	0.11	-0.82	0.49	.62		
Test trial	2	-0.23	0.03	-0.59	0.12	.19		

Study-test trial	1	0.83	0.08	0.28	1.38	.00*		
Activity type							13.42	.00*
Paired associate	13	-0.23	0.01	-0.41	-0.06	.01*		
Comprehension activities	1	0.83	0.08	0.28	1.38	.00*		
Production activities	2	-0.23	0.03	-0.59	0.12	.19		
Provision of feedback							0.01	.93
Absence	6	-0.16	0.01	-0.45	0.14	.31		
Presence	8	-0.14	0.03	-0.42	0.15	.36		
Feedback timing							1.06	.30
Immediate	5	0.04	0.09	-0.44	0.52	.88		
Delayed	3	-0.36	0.03	-0.94	0.22	.23		

*Statistically significant at $p < .05$.

Equal vs. Expanding with Immediate and Delayed Posttests

We found no factors that moderate the effects of spaced practice manipulated with relative spacing (equal and expanding spacing), but the results were different in the delayed effects (from the delayed posttest scores). Results from the delayed posttests showed that age, type of practice, and activity type significantly moderate the effect of relative spacing. The effect of equal spacing was more pronounced for young learners (g = 0.83, CI = 0.28, 1.38) than for adult learners. However, the effect of expanding spacing was greater for adult learners (g = -0.23, CI = -0.39, -0.08) than young learners. The effect of expanding spacing was larger when the spaced practice was manipulated with test-restudy trials (g = -0.32, CI = -0.54, -0.10) than the other trials (study-only trials and test-only trials). It should be noted that although the effect of equal spacing was very large when the practice manipulated with study-test trials (g = 0.83, CI = 0.28, 1.38), the sample size was only one. More research on spaced

practice involving other types of trials such as study-only, test-only, or study-test trials is needed. Lastly, the effect of expanding spacing was more pronounced than the effect of equal spacing in the spaced practice with paired associate learning task (g = -0.23, CI = -0.41, -0.06) than the practice with the other learning tasks. However, other learning tasks (k = 1 for comprehension activities and k = 2 for production activities) were small, it should be careful to interpret the results. More research on spaced practice involving other activity types (comprehension and production activities) is needed.

Table S8.9.

Moderator Analysis for Continuous Variables (Equal vs. Expanding)

				95% CI		
	k	*Q*	*B*	Lower	Upper	*p*
Frequency of practice (Immediate posttests)	7	0.56	-0.1232	-0.4467	0.2003	.46
Frequency of practice (Delayed posttests)	16	0.06	-0.0191	-0.1664	0.1283	.80
Retention interval	16	4.36	-0.0106	-0.0206	-0.0006	.04*

Note. CI = confidence interval

*Statistically significant at p <.05.

Appendix S9: Further Analyses for the Moderators Frequency of Practice and Retention Interval

In the current meta-analysis, following Suzuki (2017), an immediate posttest was regarded as a learning session. When coding the frequency of practice for the delayed effect (delayed posttest score was considered as a dependent variable), the immediate posttest was considered as one learning session and counted as one frequency of practice (note that this was the case only if the RI was manipulated within participants). However, one of the reviewers commented that there were some studies that involved different types of posttests on immediate posttests (e.g., receptive and productive). To make it clear whether this affects the results, we did further analyses. We recoded multiple types of posttests as two separate learning sessions to reflect the posttest. Table S9.1 describes studies that involved multiple types of posttests on immediate posttests as well as frequency of practice that we recoded. Table S9.2 presents the results from the analyses.

Table S9.1.

A Study List for Recoding the Frequency of Practice (k = 12)

Study	Frequency of Practice	Detail
Lee & Choe, 2014 (Experiment 2)	6	4 for the treatment + 2 for immediate posttest (receptive and productive)
Miles, 2014	4	2 for the treatment + 2 for immediate posttest (receptive and productive)
Nakata, 2015a	6	4 for the treatment + 2 for immediate posttest (receptive and productive)

Nakata & Webb, 2016 (Experiment 1)	7	5 for the treatment + 2 for immediate posttest (receptive and productive)
Nakata & Webb, 2016 (Experiment 2)	6	4 for the treatment + 2 for immediate posttest (receptive and productive)
Suzuki, 2017	30	27 for the treatment (4 times for vocabulary practice + 4 times for grammar practice + 1 monitoring test) + 3 for immediate posttest (productive)
Suzuki & DeKeyser, 2017a	8	6 for the treatment + 2 for immediate posttest (productive)
Kasprowicz *et al.*, 2019	5	3 for the treatment + 2 for immediate posttest (receptive)
Koval, 2019	6	4 for the treatment + 2 for immediate posttest (receptive)
Li & DeKeyser, 2019	6	3 for the treatment + 3 for immediate posttest (productive)
Koval, 2020	9	6 for the treatment + 3 for immediate posttest (receptive)
Nakata & Elgort, 2021	5	3 for the treatment + 2 for immediate posttest (receptive)

Table S9.2.

Moderator Analyses for Continuous Variable "Frequency of Practice"

					95% CI		
	Category	*k*	*Q*	*B*	Lower	Upper	*p*
Frequency of practice (Delayed posttests)	Spaced vs. Massed	15	0.02	-0.0096	-0.1431	0.1239	.89
Frequency of practice (Delayed posttests)	Longer vs. Shorter	30	1.69	-0.0294	-0.0736	0.0149	.19
Frequency of practice (Delayed posttests)	Equal vs. Expanding	16	0.01	-0.0075	-0.1429	0.1279	.91

Note. CI = confidence interval

Results showed that there was no difference between the previous analyses (immediate posttest was regarded as one learning session even though a study involved different types of posttests on the immediate posttest) and the further analyses (immediate posttest was regarded as separate learning sessions when a study involved different types of posttests on the immediate posttest). In the previous analyses, the random-effects meta regression analyses showed a negative relationship between frequency of practice and effect sizes with the delayed effects in the spaced vs. massed comparison (i.e., the more the frequency of practice, the smaller the spacing effects in the long term). The further analyses also showed a negative relationship between frequency of practice and effect sizes with the delayed effects in the comparison. In the longer vs. shorter comparison, the previous analyses showed that there was a negative relationship between frequency of practice and effect sizes, and we found a similar pattern in the further analyses (a negative relationship was also found when the different types of immediate posttests were regarded as separate learning sessions). In the equal vs. expanded comparison, we found a negative relationship between frequency of practice and effect sizes in both analyses. However, the differences in all these comparisons did not reach statistical significance.

In the current meta-analysis, following Suzuki (2017), when a study involved 7-day and 35-day delayed posttests, the calculated RI is 28 days (Note that this was the case only if the RI was manipulated within participants). One of the reviewers suggested that the first delayed posttest could be coded as a dependent variable (for delayed effect) and the interval from the last learning session to the first delayed posttest session as RI. To make it clear whether this affects the results, we did further analysis. We coded the first delayed posttest score as a dependent variable (for the delayed effect) and the interval from the last learning session to the first delayed posttest session as the RI. Table S9.3 describes studies that involved multiple numbers of delayed posttests as well as the RIs that we recounted. Table S9.4 presents the results from the analyses.

Table S9.3.

A Study List for Recoding the RI (k = 6)

Study	RI	Detail
Bird, 2010	7	When a study involved two or three delayed posttests, the interval from the last learning session to the first delayed posttest was selected as retention interval.
Schuetze, 2014 (Experiment 1)	1	
Schuetze, 2014 (Experiment 2)	1	
Lotfolahi & Salehi, 2016	7	
Suzuki, 2017	7	
Suzuki & DeKeyser, 2017a	7	

Notes. In the moderator analyses for RI, Serrano and Huang's (2018) study was excluded due to different delayed posttest time point (RI was manipulated between participants): Shorter spacing condition involved 4-day delayed posttest and longer spacing condition involved 28-day delayed posttest based on the optimal ISI/RI ratio.

Table S9.4.

Moderator Analyses for Continuous Variable "RI"

					95% CI		
	Category	*k*	*Q*	*B*	Lower	Upper	*p*
Retention interval (Delayed posttests)	Spaced vs. Massed	15	0.07	-0.0050	-0.0408	0.0308	.79
Retention interval (Delayed posttests)	Longer vs. Shorter	31	3.16	-0.0165	-0.0347	0.0017	.08
Retention interval (Delayed posttests)	Equal vs. Expanding	16	1.43	-0.0074	-0.0195	0.0047	.23

Note. CI = confidence interval

Results showed that there was no difference between the previous analyses (RI was averaged when a study involved two or three delayed posttests) and the further analyses (The interval from the last treatment to the first delayed posttest was coded as RI) in the spaced vs. massed and longer vs. shorter spacing comparisons. In the previous analyses, the random-effects meta regression analyses showed no significant relationships between RI and effect sizes in both spaced vs. massed and longer vs. shorter comparisons. The further analyses also found no significant relationships between RI and effect sizes in both comparisons. However, the results in the equal vs. expanding comparison were different. The previous analyses showed a significant negative relationship, indicating that the longer the RI the larger the expanding spacing effects. The further analyses showed a similar pattern (a negative relationship between RI and effect sizes), but this did not reach statistical significance. This is perhaps due to large differences of RI and effect sizes between the analyses: one study (Schuetze, 2014) included in the equal vs. expanding comparison changed the retention interval of 28 days (1 day, 28 days, and 56 days were averaged in the previous analyses) to the retention interval of 1 day (the first delayed posttest time point was selected as RI in the further analyses) for the further analyses; small to medium effect sizes (g =-0.14 from experiment 1 and g =-0.38 from experiment 2 in Schuetze, 2014) were changed to medium to large effect sizes (g =-0.40 from experiment 1 and g =-0.62 from experiment 2 in Schuetze, 2014).

Additional analyses for potential confounding factors with frequency of practice

One of anonymous reviewers suggested a potential confound between frequency of practice and whether the practice is within- or between-sessions. A closer inspection of the data in both spaced vs. massed and longer vs. shorter spacing comparisons showed that the effects of

distributed practice (both spacing and lag effects) diminished when studies that included larger values (e.g., 10-11 repetitions in Suzuki *et al*., 2020, 27-30 repetitions in Suzuki, 2017) were involved, regardless of whether the practice was within- or between- sessions. However, we cannot say that frequency of practice accounts for the diminished effects of distributed practice (see the description of the results from Suzuki's (2017) study above).

Since grammar studies were likely to include much larger values (more frequency of practice) (e.g., Suzuki, 2017; Suzuki *et al*., 2020), we had a look at the within-session studies on vocabulary, grammar, and pronunciation learning as well as the between-session studies on vocabulary, grammar, and pronunciation learning on the immediate and delayed posttests (see Tables S9.5 and S28).

Table S9.5.

Relationship between Learning Target (Vocabulary, Grammar, Pronunciation) and Number of Sessions (Within-and Between-Sessions) on Immediate Posttests

Nature of spacing	Comparison	Target items		
		Vocabulary	Grammar	Pronunciation
Within-session	Spaced vs. Massed (k = 6)	(k = 4) g = 1.45, CI = 0.86, 2.05, p < .001	(k = 2) g = 0.31, CI = -0.38, 1.00, p = .38	
	Longer vs. Shorter (k =10)	(k = 6) g = 0.22, CI = -0.08, 0.51, p = .16		(k = 4) g = -0.64, CI = -0.98, -0.30, p < .001
Between-session	Spaced vs. Massed (k = 5)	(k = 4) g = 0.13, CI = -0.45, 0.70, p = .67	(k = 1) g = -0.25, CI = -0.93, 0.43, p = .48	

	Longer vs. Shorter (k = 7)	(k = 3) g = -0.04, CI = -0.50, 0.42, p = .86	(k = 4) g = -0.41, CI = -0.70, -0.13, p = .01

Note. CI = confidence interval

Table S9.6.

Relationship between Learning Target (Vocabulary, Grammar, Pronunciation) and Number of Sessions (Within-and Between-Sessions) on Delayed Posttests

Nature of spacing	Comparison	Target items		
		Vocabulary	Grammar	Pronunciation
Within-session	Spaced vs. Massed (k = 9)	(k = 5) g = 1.09, CI = 0.69, 1.49, p < .001	(k = 4) g = 0.03, CI = -0.30, 0.36, p = .88	
	Longer vs. Shorter (k =11)	(k = 9) g = 0.79, CI = 0.32, 1.25, p = .001	(k = 2) g = 0.66, CI = 0.36, 0.96, p < .001	
Between-session	Spaced vs. Massed (k = 6)	(k = 5) g = 1.27, CI = 0.53, 2.02, p = .001	(k = 1) g = 0.51, CI = -0.18, 1.20, p = .15	
	Longer vs. Shorter (k = 21)	(k = 13) g = 0.02, CI = -0.22, 0.25, p = .90	(k = 6) g = 0.57, CI = -0.18, 1.32, p = .13	(k = 2) g = 0.42, CI = -0.07, 0.92, p = .09

Note. CI = confidence interval

As the tables show above, we found a pattern that when grammar studies were manipulated spacing within a session, the effects of spacing diminished on the delayed posttest (g = 0.31 on the immediate posttest and g = 0.03 on the delayed posttest in the spaced vs. massed

comparison), but the effects in grammar studies were larger when the practice was between sessions (g = -0.25 on the immediate posttest and g = 0.51 on the delayed posttest in the spaced vs. massed comparison; g = -0.41 on the immediate posttest and g = 0.57 on the delayed posttest in the longer vs. shorter comparison. For example, when we looked at two grammar studies (Suzuki *et al*., 2020 and Suzuki, 2017), Suzuki *et al*. (2020) was a within-session study (10-11 repetitions) and showed the diminished spacing effects on the delayed posttest (g = 0.67 on the immediate posttest and g = 0.41 on the delayed posttest). However, Suzuki (2017) was a between-session study (27-30 repetitions) and the effects were larger on the delayed posttest (g = -0.63 on the immediate posttest and g = -0.64 on the delayed posttest). Therefore, whether grammar studies were manipulated spacing in within a session or between sessions may account for the diminished effects of distributed practice (for both spacing effects and lag effects) on the delayed posttest. However, it should be interpreted with caution because of small sample sizes for grammar studies (k = 7). There is value in further research on grammar and pronunciation learning in this area.

Appendix S10: A Full List of All the Included Studies in the Current Meta-Analysis

1. Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52,* 566−577. http://doi.org/10.1016/j.jml.2005.01.012
2. Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics, 31,* 635−650. http://doi.org/10.1017/S0142716410000172
3. Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research 74*(4), 245−248. http://doi.org/10.1080/00220671.1981.10885317
4. Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition, 41*, 671−682. http://doi.org/10.3758/s13421-012-0291-4
5. Çekiç, A., & Bakla, A. (2019). The effects of spacing patterns on incidental L2 vocabulary learning through reading with electronic glosses. *Instructional Science, 47*(3), 353−371. https://doi.org/10.1007/s11251-019-09483-4
6. Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236−246. http://doi.org/10.1027/1618-3169.56.4.236
7. Gerbier, E., & Koenig, O. (2012). Influence of multiple-day temporal distribution of repetitions on memory: A comparison of uniform, expanding, and contracting schedules. *The Quarterly Journal of Experimental Psychology, 65*(3), 514−525. http://doi.org/10.1080/17470218.2011.600806
8. Gerbier, E., Toppino, T. C., & Koenig, O. (2015). Optimising retention through multiple study opportunities over days: The benefit of an expanding schedule of repetition. *Memory, 23*(6), 943−954. http://doi.org/10.1080/09658211.2014.944916
9. Kanayama, K., & Kasahara, K. (2016). The effects of expanding and equally-spaced retrieval practice on long-term L2 vocabulary retention. *ARELE: Annual Review of English Language Education in Japan, 27*, 217−232. http://doi.org/10.20581/arele.27.0_217
10. Kang, S., Lindsey, R., Mozer, M., & Pashler, H. (2014). Retrieval practice over the long time: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review, 21*(6), 1544−1550. http://doi.org/10.3758/s13423-014-0636-z
11. Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250−1257. http://doi.org/10.1037/a0023436

12. Kasprowicz, R., 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. http://doi.org/ 10.1111/modl.12586
13. Khoii, R., & Abed, K. F. (2017). Effects of equal spacing, expanding spacing, and massed condition on EFL learners' receptive and productive vocabulary retrieval. In Pixel, *Proceedings of ICT for language learning (19th Ed.)* (pp. 500−504). Florence, Italy: ICT for Language Learning. Retrieved from https://conference.pixel-online.net/ICT4LL/files/ict4ll/ed0010/FP/0960-SLA2580-FP-ICT4LL10.pdf
14. Koval, N. G. (2019). Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics, 40*(5), 1−37. http://doi.org/10.1017/S0142716419000158
15. Koval, N. G. (2020). *Testing the reminding account of the lag effect in L2 vocabulary acquisition from L2-L1 retrieval practice within a paired-associate learning format* (Published doctoral dissertation). Michigan State University, The United States.
16. Lee, E., & Choe, M-H. (2014). The effect of spaced repetitions on Korean elementary students' L2 English vocabulary learning. *Studies in English Education, 19*(1), 55−75.
17. Li, M., & DeKeyser, R. (2019). Distribution of practice effects in the acquisition and retention of L2 Mandarin tonal word production. *The Modern Language Journal, 103*(3), 607−628. http://doi.org/10.1111/modl.12580
18. Lotfolahi, A. R., & Salehi, H. (2016). Learners' perceptions of the effectiveness of spaced learning schedule in L2 vocabulary learning. *SAGE Open, 6*(2), 1−9. http://doi.org/10.1177/2158244016646148
19. Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System, 42*, 412−428. http://doi.org/10.1016/j.system.2014.01.014
20. Nakata, T. (2015a). 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. http://doi.org/10.1017/S0272263114000825
21. Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilities the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research, 37*(2), 233−260. http://doi.org/10.1177/0267658320927764
22. Nakata, T., & Suzuki, Y. (2019a). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition, 41*(2), 287−311. http://doi.org/10.1017/S0272263118000219
23. Nakata, T., & Suzuki, Y. (2019b). Mixing grammar exercises facilitates long-term retention: Effects of blocking interleaving and increasing practice. *The Modern Language Journal, 103*(3), 629-647. http://doi.org/10.1111/modl.12581

24. Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition, 38*(3), 523−552. https://doi.org/10.1017/S0272263115000236
25. Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., & Rickard, T. C. (2019). Does interleaved practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills. *Journal of Educational Psychology, 111,* 1172−1188. http://doi.org/10.1037/edu0000336
26. Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1051−1057. http://doi.org/10.1037/0278-7393.29.6.1051
27. Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition, 35*(8), 1917−1927. http://doi.org/10.3758/BF03192925
28. Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437−447. http://doi.org/10.1016/j.jml.2009.01.004
29. Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly, 49*(4), 857−866. http://doi.org/10.1002/tesq.252
30. Rogers, J., & Cheung, A. (2020a). Input spacing and the learning of L2 vocabulary in a classroom context. Language Teaching Research, 24, 616–641. http://doi.org/10.1177/1362168818805251
31. Rogers, J., & Cheung, A. (2020b). Does it matter when you review? Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition.* http://doi.org/10.1017/S0272263120000236
32. Schuetze, U. (2014). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research, 19*(1), 28−42. http://doi.org/10.1177/1362168814541726
33. 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. http://doi.org/10.1002/tesq.445
34. Snoder, P. (2017). Improving English learners' productive collocation knowledge: The effects of Involvement Load, spacing, and intentionality. *TESL Canada Journal*, *34*(3), 140−164. http://doi.org/10.18806/tesl.v34i3.1277

35. 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. http://doi.org/10.1111/lang.12236
36. Suzuki, Y., & DeKeyser, R. (2017a). Effects of distributed practice on the preceduralization of morphology. *Language Teaching Research, 21*(2), 166−188. http://doi.org/10.1177/1362168815617334
37. Suzuki, Y., Yokosawa, S., & Aline, D. (2020). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research*. http://doi.org/10.1177/1362168820913985

Appendices for Studies 2 and 3

Appendix A. EFL Textbook Analysis

Table 1

EFL Textbook List

1	Clandifield, L. (2011).	*Global: Intermediate coursebook: Student's book*. Oxford: MacMillan.
2	Clare, A., & Wilson, J. (2011).	*Speakout: Intermediate student's book.* Harlow: Pearson Education.
3	Cunningham, S., & Moor, P. (2005).	*New cutting edge intermediate: Student's book*. Harlow: Pearson Education.
4	Hancock, M., & McDonald, A. (2009).	*English result: Intermediate student's book*. New York: Oxford University Press.
5	Kay, S., & Jones, V. (2009).	*New inside out: Intermediate student's book*. Oxford: MacMillan.
6	Kerr, P., & Jones, C. (2005).	*Straightforward: Intermediate student's book.* Oxford: MacMillan.
7	Oxenden, C., & Latham-Koenig, C. (2006).	*New English file: Intermediate student's book*. New York: Oxford University Press.
8	Richards, J., & Bohlke, D. (2011).	*Four corners: Level 3*. Cambridge: Cambridge University Press.
9	Roberts, R., Clare, A., & Wilson, J. (2011).	*New total English: Intermediate student's book*. Harlow: Pearson Education.
10	Soars, L., & Soars, J. (2003).	*New headway: Intermediate student's book.* New York: Oxford University Press.

Table 2

Results of EFL Textbook Analysis

	Number of units	Matching	Glossing	Guessing meaning from context	Identifying words from text	Categorizing	Fill-in (Cloze)	Sentence production	Multiple-choice	Correct the spellings	Dictionary use and negotiation of meaning	Word parts: Classifying	Word parts: Write the correct form of the word	Association and written form	Recall	Mapping	Crossword puzzle	Word search	Total
1	10	23	1	2	4	3	21	8	10		5		11	1			1		90
2	10	24	2	3	3	4	18	2	4	1	3		2	3		5			74
3	12	20	1	3	3	6	5	2	10		15		2	4	6	5	1		83
4	10	19	1	1	5	5	14	3					8	3			7	2	68
5	12	26			2	21	32	5	10		3		1	7		1			108
6	12	18				3	16	2	4			2	3	2					50
7	7	13				2	16	2	5		3		7	5	2				55
8	12	24				4	1		2				3		11				45
9	10	23		2	5	11	17	7	9		3	1	5	2	1	2			88
10	12	11		1		5	5	1	2			2	3	4					34
		201	5	12	22	64	145	32	56	1	32	5	45	31	20	13	9	2	695

Appendix B. List of Forty-Eight Target Words

No.	Target word	Korean translation	Part of Speech	Word length	BNC/COCA25	Concreteness
1	Bicker	다투다	V	6	8K	3.15
2	Otter	수달	N	5	8K	4.86
3	Snip	자르다	V	4	8K	3.68
4	Wilt	시들다	V	4	8K	2.90
5	Squander	낭비하다	V	8	8K	2.54
6	Stammer	말을 더듬다	V	7	8K	2.93
7	Cringe	민망하다	V	6	8K	3.34
8	Gobble	게걸스럽게 먹다	V	6	8K	3.37
9	Boar	야생돼지	N	4	8K	4.80
10	Casket	장례식 관	N	6	8K	4.86
11	Serpent	큰 뱀	N	7	8K	4.97
12	Tandem	2인용자전거	N	6	8K	3.16
13	Stag	수사슴	N	4	9K	4.39
14	Antler	사슴 뿔	N	6	9K	4.86
15	Croon	노래하다	V	5	9K	3.19
16	Belch	트림하다	V	5	9K	4.14
17	Haggle	흥정하다	V	6	9K	2.93
18	Amble	느릿느릿 걷다	V	5	9K	3.17

19	Sentry	보초병	N	6	9K	4.04
20	Holler	소리 지르다	V	6	9K	3.57
21	Cackle	낄낄거리며 웃다	V	6	9K	3.63
22	Meddle	간섭하다	V	6	9K	2.43
23	Weasel	족제비	N	6	10K	4.74
24	Chastise	꾸짖다	V	8	10K	2.68
25	Mauve	연 보라색	N	5	10K	4.00
26	Bawl	울다	V	4	10K	3.52
27	Mirage	신기루	N	6	10K	3.32
28	Hatchet	손도끼	N	7	11K	4.93
29	Wick	양초심지	N	4	11K	4.69
30	Azalea	진달래꽃	N	6	11K	4.40
31	Tetanus	파상풍	N	7	11K	4.19
32	Scowl	노려보다	V	5	11K	3.85
33	Shrew	뾰족뒤 쥐	N	5	11K	4.07
34	Trowel	모종삽	N	6	11K	4.16
35	Gazebo	정원의 정자	N	6	12K	4.79
36	Icicle	고드름	N	6	12K	4.96
37	Gourd	식물 박	N	5	12K	4.86
38	Fawn	아양을 떨다	V	4	12K	4.30
39	Quail	메추라기 새	N	5	12K	4.65
40	Notary	공증인	N	6	12K	3.69

41	Carousel	수하물 벨트	N	8	12K	4.93
42	Faucet	수도꼭지	N	6	13K	4.48
43	Conch	소라 조개	N	5	13K	4.52
44	Slobber	침 흘리다	V	7	13K	4.33
45	Toboggan	썰매	N	8	14K	4.76
46	Snitch	고자질하다	V	6	15K	3.85
47	Abacus	주판	N	6	15K	4.52
48	Toupee	부분가발	N	6	16K	4.65

Notes. Concreteness refers to the degree to which a word is concrete. The index represents the average concreteness score for content words, based on the crowd-sourced norms for 40,000 words collected by Brysbaert, Warriner, and Kuperman (2014). A 5-point rating scale going from abstract (1) to concrete (5) was used. The concreteness scores for target words included in the present study ranged from 2.43 to 4.97, and the mean score was 4.02 (SD = 0.75).

Appendix C. Sentences Used in the Presentation Phase

No.	Target word	Sentences used in the presentation phase	Korean translation
1	Bicker	The husband and wife **bicker** less.	그 부부는 잘 다투지 않아.
2	Otter	I saw a sea **otter** on the island.	그 섬에서 바다 수달을 봤어.
3	Snip	We started to **snip** apart the packaging.	우리는 포장된 것들을 자르기 시작했다.
4	Wilt	Some flowers were beginning to **wilt**.	어떤 꽃들은 시들기 시작하고 있었다.
5	Squander	We watched the guy **squander** all his money on games.	그 남자가 게임에서 돈을 모두 낭비하는 걸 봤어.
6	Stammer	Many children **stammer** but grow out of it.	많은 아이들이 말을 더듬지만, 크면서 괜찮아진다.
7	Cringe	I **cringe** when I think of the poems I wrote at night.	밤에 썼던 시들을 생각하면 민망해.
8	Gobble	I am so hungry that I can **gobble** the whole thing up.	너무 배가 고파서, 전부 다 먹어 치울 수 있어.
9	Boar	I saw a **boar** run on the road last night.	지난 밤, 야생 돼지가 찻길 위를 뛰어가는 것을 봤어.
10	Casket	Six men carried the **casket** into the church.	남자 여섯 명이 관을 교회 안으로 운반했다.
11	Serpent	I was surprised to find a **serpent** in my garden.	나는 정원에서 큰 뱀을 발견하고 깜짝 놀랐어.
12	Tandem	There were two people riding a **tandem** in a park.	공원에서 2인용 자전거를 타는 사람들이 있네.
13	Stag	I saw a **stag** in the forest this morning.	오늘 아침에 수사슴 한 마리를 봤어.
14	Antler	The **antler** on the wall was beautiful.	벽에 걸린 사슴 뿔이 너무 이뻤어.
15	Croon	I am happy when I hear someone **croon**.	나는 누군가가 노래를 부르면, 기분이 좋아.
16	Belch	A mother tries to make a baby **belch** after he eats.	엄마는 아기가 밥은 먹고 나면, 트림을 시키려 한다.
17	Haggle	I often saw my friend **haggle** over the price.	나는 가끔 친구들이 가격 흥정하는 것을 봤어.
18	Amble	People **amble** along the road for miles every day.	사람들이 매일 그 길을 따라 수 마일을 걷더라구.
19	Sentry	We took a picture of a **sentry** standing at the front gate.	우리는 문 앞에 서 있는 보초병의 사진을 찍었다.
20	Holler	Don't **holler** at me!	나한테 소리 지르지 마!

21	Cackle	I was so scared when I saw the Joker **cackle**.	나는 조커가 웃는 걸 보면 무서워.
22	Meddle	She likes to **meddle** in other people's business.	그녀는 다른 사람들 일에 간섭하는 걸 좋아해.
23	Weasel	A **weasel** ran quickly towards him and climbed up his body.	족제비가 재빠르게 그에게로 달려와 올라탔다.
24	Chastise	Parents are not always right to **chastise** their children.	부모가 아이들을 꾸짖는 것이 항상 옳지만은 않다.
25	Mauve	I like **mauve**, the color of Lavender.	나는 연 보라색이 좋아. 라벤더 색깔 말이야.
26	Bawl	All of a sudden, a baby started to **bawl** so hard.	갑자기 아기가 울어대기 시작했다.
27	Mirage	His idea is like a **mirage**. It never happens.	그의 생각은 신기루 같은 거야 현실적이지 않아.
28	Hatchet	When you swing a **hatchet**, you should keep your eyes on it.	손도끼를 휘두를 때는, 그 도끼에서 눈을 떼면 안돼.
29	Wick	Father made a small lamp by putting oil and a **wick** in a glass.	아빠는 유리잔에 오일과 심지를 넣어 작은 램프를 만들었다.
30	Azalea	My mother likes flowers, and she has an **azalea** garden.	엄마는 꽃을 좋아해서, 진달래 꽃 정원을 가지고 계셔.
31	Tetanus	You may need a shot for **tetanus**.	아마도 파상풍 주사를 맞아야 할거야.
32	Scowl	She turned to **scowl** at me.	그녀가 나를 노려보기 시작했다.
33	Shrew	He saw a **shrew** was running on the floor.	그는 바닥 위를 지나가는 쥐를 보았다.
34	Trowel	She was working with her **trowel** on a new garden.	그녀는 모종삽으로 새 정원을 가꾸고 있었다.
35	Gazebo	An old lady took a seat in a **gazebo** near the lake.	한 노인이 호숫가 근처 정자에 앉았다.
36	Icicle	There is a giant **icicle** hanging over the window.	창문에 커다란 고드름이 달려있다.
37	Gourd	My grandmother uses a **gourd** as a water basket.	할머니는 박을 물통으로 사용한다.
38	Fawn	Everyone seems to **fawn** over the new boss to get a higher position.	사람들이 승진 때문에, 새 상사에게 아양을 떠는 것 같다.
39	Quail	I like a hat decorated with a **quail** feather.	나는 메추라기 새 깃털 장식이 된 모자가 마음에 들어.
40	Notary	All agreements are signed by a **notary**.	모든 합의안들이 공증인에 의해 서명된다.
41	Carousel	Waiting for my bags to come out of the **carousel** drives me crazy.	수하물 벨트로 나오는 짐들을 기다리는 건 너무 싫어.
42	Faucet	I turned on a **faucet** to prepare baths for my children.	수도꼭지를 틀어, 아이들을 위해 욕조에 물을 받았다.
43	Conch	We found a silver **conch** on the beach.	우리는 해변가에서 은빛 소라를 발견했다.
44	Slobber	Babies **slobber** everywhere.	아기들은 침을 여기저기 흘려.

45	Toboggan	The winter park is open, and we can ride a **toboggan**.	윈터 파크가 개장해서, 우리는 썰매를 탈 수 있어.
46	Snitch	I don’t **snitch** on anyone.	나는 누구에 대해서도 고자질 하지 않아.
47	Abacus	He is clever with the **abacus**.	그는 주판을 잘해.
48	Toupee	He has a **toupee**.	그는 부분 가발을 쓰고 있어.

Appendix D. Sentences Used in the Fill-in-the-blank Exercise

No.	Target word	Sentences used in the fill-in-the-blank exercise
1	Bicker	We always _____ while choosing a delivery menu.
		Sisters always _____ over what to wear.
		They always _____, and they don't seem to like each other.
		They _____ about how to decorate the room.
		You always _____ with your brother but you two are close?
2	Otter	There is a five-week-old southern _____ at the Sea World Park.
		An _____ is a great swimmer.
		An _____ swims well and eats fish.
		A sea _____ can sleep in the water.
		The boy swims like an _____.
3	Snip	I _____ off a piece of gray hair when it grows.
		You need to _____ young leaves to grow them well.
		I asked a hair designer just to _____ the ends of my hair.
		Can you _____ the corner off the package?
		I _____ out the photos of BTS in magazines and give them to my daughter.
4	Wilt	If plants dry out, leaves will _____ and drop.
		If plants are well watered, they won't _____.
		The leaves _____ quickly when they are not watered.
		Vegetables _____ quickly if they are washed and left for long.
		She uses only dried flowers because real ones _____ too quickly.
5	Squander	We saw the team _____ several good scoring chances.
		No one wants to _____ the time or money.
		I don't want to _____ an opportunity for my career.
		Parents _____ their energy playing with their kids.
		I don't want to _____ my summer break at home.

6	Stammer	People sometimes _____ when they lie. I am not a good presenter because I often _____ during my presentation. I used to _____ so badly, and I didn't like reading out loud. I _____ when I give a speech in public. I'm worried about the presentation because I _____ when I'm nervous.
7	Cringe	I _____ when my jokes don't work. Hearing my voice recorded makes me _____. People _____ at the sound of their own voice recorded. I still _____ when I watch myself act or dance. Every time I watch a film with family, I _____ at the love scene.
8	Gobble	The football players _____ their food after a match. After a long walk, I _____ a big sandwich. I am so hungry that I can _____ the entire basket of bread. I am so hungry that I _____ down my brother's chips. I saw lions _____ weak or small animals on TV.
9	Boar	We eat pigs, but do we eat _____? What is the difference between a pig and a _____? A _____ is a wild-born pig. A _____ can be more dangerous to hunt than a bear. You cannot have a _____ as a pet, because it is different from a pig.
10	Casket	We bought a _____ when her grandmother died. There was a tiny _____ for the body of a 6-year-old girl. He was not ready to let his grandmother go, so he couldn't close her _____. We had an open _____ for our uncle, and viewing is tomorrow. Grandfather didn't want people to look at his dead body, so we had a closed _____.
11	Serpent	A _____ is a snake, but it is much bigger. A _____ is more scary than a small snake. A _____ is a poisonous snake? You can be bitten by a _____ in the woods. A _____ is much bigger than a snake.

12	Tandem	A father and his daughter ride on a _____, because she cannot ride a bike. Riding a _____ appeals to people who prefer riding together. A _____ is a bike designed for two people. A _____ is on sale, and it is cheaper than a bike. I was riding down the coast with my girlfriend on a _____.
13	Stag	A _____'s head is a long-time Christmas decoration favorite in Germany. A _____ has long legs, but I didn't expect it to be that fast. A _____ with a red nose is a famous Christmas character. My children like decorating a _____ with a red nose and a Santa hat for Christmas. I saw a _____ standing and feeding on grass in the park.
14	Antler	The price of an _____ depends on its size and shape. A hunter likes the decoration of an _____ on the wall. An _____ is used as a coat hanger. A large _____ is expensive but a good decoration on the wall. An _____ is made up of bone, skin, and blood.
15	Croon	My mother used to _____ when she was happy. They _____ a song together at their wedding. I like to _____ love songs. I like to hear birds _____ early in the morning. Mother used to _____ songs when I was sleeping.
16	Belch	He covered his hand across his mouth, then began to _____. I _____ after I eat fast. When you _____, you should hide it with your hand. You _____ after you drink a spring water. I can hear a man _____ after he eats a big bucket of chicken at KFC.
17	Haggle	My friends _____ well over prices. You can _____ over prices of used clothes. It is not easy to _____ with a salesperson. I don't always _____ for a better deal. We can _____ over the prices on the market.

18	Amble	I often _____ down toward the river when alone.
		He likes to _____ from his apartment down to his restaurant.
		I _____ along the sea with my dog.
		We _____ through the park after lunch.
		I saw elephants _____, and they never run.
19	Sentry	A _____ standing at the gate raises his hand when the Queen comes.
		A _____ in England is a guard for the Queen.
		My dog is like a trained _____, and I feel safe.
		A _____ stands at each corner of the castle.
		The _____ standing at the door is a doll, not a real person.
20	Holler	My parents never _____ at me even though they are angry.
		I just get too angry and _____ at the TV.
		My sisters always _____ whenever they are angry.
		A couple of fans _____ at the actor when he gets out of the car.
		Some people _____ and cry when they are angry.
21	Cackle	I heard women _____ last night, and it was noisy.
		The girls ____, watching a funny scene on TV.
		The girls started to _____ at his joke.
		The ladies always sit and _____ loudly at the party.
		Women _____ from the backseat, but the presenter ignores their laughter.
22	Meddle	Parents always _____ in their children's lives.
		Parents should not _____ too much unless their children are in danger.
		The president should not _____ in the election.
		Although I am a boss, I don't want to _____ in their project.
		I don't want my parents to _____ in my life.
23	Weasel	A _____ has a long body with brown fur.
		The _____ and red fox fur are used for coats and jackets.
		A long-tailed _____ runs through the grass.
		A _____ eats vegetable, fruit, and small animals such as birds and rats.
		A _____ looks like a wild cat but it has a thin body with a small head.

24	Chastise	The boss would _____ building managers who left the lights on. Teachers _____ their students when they did something wrong. Parents _____ me for not doing the right thing. I _____ my son for his bad eating habits. Grandchildren are so lovely that grandparents can't even _____ them.
25	Mauve	She looks good with _____, not yellow or red. I prefer the color _____ to purple. What colors make _____? They are purple and white. Do you prefer to dress in red or _____? She dressed in black and _____ and wore too much makeup.
26	Bawl	She began to _____ like a baby in front of everyone. Babies _____ when they're hungry. Kids started to _____ when their parents left the daycare center. I always ____ at sad movies. She started to _____ like a child who has lost her ice cream.
27	Mirage	A _____ sometimes appears, but it is not real. People often see a _____ in the desert. A _____ naturally occurs in the desert or at sea, but it is not real. You can see a _____ on a hot day in the desert. Our dreams never come true, like a _____ in the desert.
28	Hatchet	A woman was killed with a _____ last night. The man is cutting wood with a _____. The man was working with a _____ in a wood yard. A _____ is for single-handed use to cut wood. Father used a _____ to cut down a tree.
29	Wick	I turned up the _____ and looked outside. I don't like the smell of a burning _____. Does a wooden _____ burn faster than a cotton one? A _____ inside an oil lamp burns quickly. I cut the _____ before lighting the lamp.

30	Azalea	We can see pink _____ in the spring. An _____ plant has pink flowers. We can see _____ plants with different colors in the flower farm. I remember my grandmother liked pink _____ the most. My favorite flower is _____, and I prefer white one to pink one.
31	Tetanus	_____ is a serious disease. _____ is the disease caused by dust or animals. What are the first signs of _____? You can get _____ through a cut or other wound. _____ is a serious disease, and you cannot breathe.
32	Scowl	People _____ at the man who is rude to an old woman. People on the bus _____ at a person who does not wear a face mask. I saw Jack _____ a lot today. Why was he so angry? People in the theatre _____ at me when I walk in late. The girls didn't raise their voice but they _____ at me.
33	Shrew	A _____ is much smaller than a mouse. A ____ is a small mouse with a long nose. A _____ is a very small mouse, and it eats food every few hours. A _____ is a small rat and makes 12 body movements per second. A _____ is a type of mouse, and it is about finger size or smaller.
34	Trowel	You need a _____ when you plant flowers. He used a _____ to change plant pots. Where can I buy a _____ for gardening? I bought a _____ for my grandmother's gardening. You can use a _____ to move small plants.
35	Gazebo	I enjoy taking my happy meal to the _____ in the backyard. The garden has a _____, where people can sit and take a rest. The _____ in the park is for people to sit and relax. We sometimes enjoy a picnic in a _____ if it rains. I feel more relaxed in a wood _____ while walking around the park.

36	Icicle	Children like picking off an _____ on the roof in winter. You should be careful of the _____ above the door. The _____ looks like a Christmas decoration on the roof. The _____ tastes like water. Children are looking for the longest _____ on the roof in the winter.
37	Gourd	He gives her some water from the _____. A _____ was a Chinese water bottle. A _____ is a large green vegetable with a hard skin. Is a _____ fruit or vegetable? I cut a _____ and dried it to make a bottle.
38	Fawn	All the men _____ to get a woman's attention. I don't want to _____ over him just because he is rich. People _____ over the boss. All actors _____ over the director to play in his movies. People _____ over him, because his opinion has a powerful influence.
39	Quail	The old man shoots a _____ and wild turkey for food. The _____ is the California state bird. A _____ flies on short, very broad wings. The eggs of the _____ are different in size. Our family loves eating _____ rather than chicken.
40	Notary	I need a _____ for some contracts for those paintings. This contract doesn't work without the _____ present. A _____ is a person who helps to carry out the process legally. The contract should be signed by two people and a _____. If we have a _____, it costs about 30 dollars for each contract.
41	Carousel	A _____ with luggage goes round and round again. Jack stopped in front of the moving baggage _____. He took the baggage from the _____. I was lucky, because my bag was the first off the _____. I was waiting for my bags at the _____.

42	Faucet	Students can drink water from a _____ in the playground. I heard the _____ run in the bathroom. Water is running from the _____. I turned the _____ on and began washing plates. I forgot to turn off the _____, and the floor was wet.
43	Conch	This table looks like a giant _____ shell. Tom picks up a _____ on the beach. A _____ is a large sea shell. My friend found a _____ shell and taught me how to blow it. A _____ tastes like shellfish.
44	Slobber	A dog can be trained not to _____. The dog began to jump up and _____ all over his face. A dog will _____ if you tease him with food. My dogs _____ all over the floor. Babies can _____ when they eat.
45	Toboggan	I like riding a _____ on an icy hill. My sister and I can ride a _____ together on the snow. A light wooden _____ is expensive in the winter season. My father and I used to ride a wooden _____ together in the winter. He enjoys riding a _____ to travel down hills in the winter.
46	Snitch	Don't _____ on things that are happening in your neighborhood. I don't tell him anything because he likes to _____. Don’t _____! I will never tell you anything again. I was not surprised when I watched her _____ on me to people. I saw Tom _____ to the teacher.
47	Abacus	Children like playing with the _____ to count numbers. The old man in the shop still calculates the bill on an _____. The _____ was a traditional calculator. You can learn how to add up numbers with the _____. I learned how to use the _____ in the math class.

48	Toupee	My teacher loses his hair, and he needs a _____. A _____ is a hair piece that is worn on the head. A _____ looks just like real hair. I didn't know he has a _____, and it looks like real hair. A _____ is a hair piece, and it looks real.

Appendix E. Target Items Assigned to Different Feedback Timing (immediate and delayed) for Each Posttest (Immediate and Delayed Posttests)

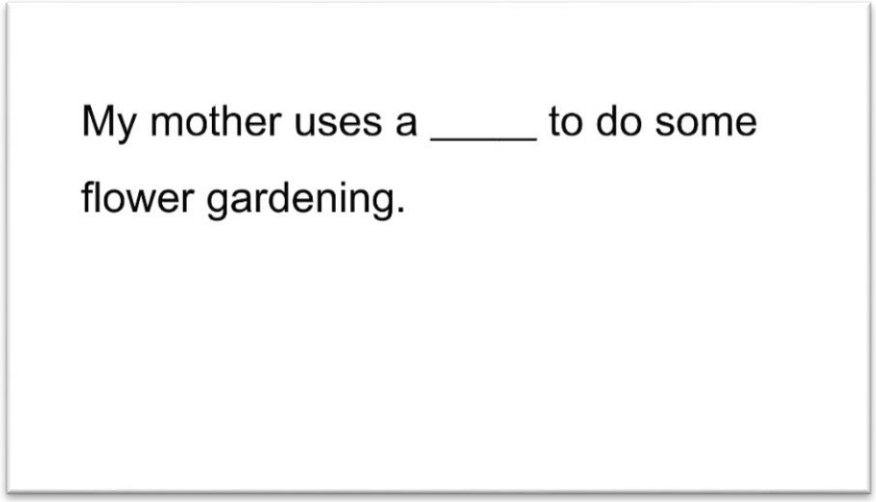
		Immediate feedback				Delayed feedback			
	SET A-1	bicker	snip	otter	boar	belch	haggle	weasel	mauve
	SET B-1	stammer	cringe	tandem	carousel	cackle	meddle	wick	azalea
Immediate posttest	SET C-1	scowl	fawn	trowel	gazebo	quail	notary	stag	faucet
	SET D-1	wilt	squander	casket	serpent	amble	holler	mirage	hatchet
	SET E-1	gobble	croon	antler	sentry	chastise	bawl	tetanus	shrew
Delayed posttest	SET F-1	slobber	snitch	icicle	gourd	conch	toboggan	abacus	toupee
		Immediate feedback				Delayed feedback			
	SET A-2	belch	haggle	weasel	mauve	bicker	snip	otter	boar
	SET B-2	cackle	meddle	wick	azalea	stammer	cringe	tandem	carousel
Immediate posttest	SET C-2	quail	notary	stag	faucet	scowl	fawn	trowel	gazebo
	SET D-2	amble	holler	mirage	hatchet	wilt	squander	casket	serpent
	SET E-2	chastise	bawl	tetanus	shrew	gobble	croon	antler	sentry
Delayed posttest	SET F-2	conch	toboggan	abacus	toupee	slobber	snitch	icicle	gourd

Appendix F. Test Items Used in the Contextualized Form Recall Test (Pretest, Immediate and Delayed Posttests)

No.	Target word	Sentences used in the contextualized form recall test
1	Bicker	My sister and I _____ more often than other sisters do.
2	Otter	An _____ sleeps, holding hands with another one, so that they don't float away.
3	Snip	I _____ a loose button and sew it.
4	Wilt	I do not like a bunch of flowers as a gift, because it's going to _____ quickly.
5	Squander	Don't _____ your money on useless things.
6	Stammer	Children sometimes _____ because they are still learning how to speak.
7	Cringe	My phone rang and made me _____ in the class.
8	Gobble	Take your time, why do you always _____ food?
9	Boar	A _____ is a wild male pig with two long sharp teeth.
10	Casket	A _____ is a box in which the body of a dead person is buried.
11	Serpent	I dreamed of a _____ last night, and I heard a big snake is a lucky dream.
12	Tandem	The man at a bike shop suggested us riding a _____ together.
13	Stag	The _____ with the red nose looks like Rudolph.
14	Antler	An _____ is an animal's head bone, and when it breaks off, it grows back.
15	Croon	I _____ to my baby every night.
16	Belch	Beer can make you _____, and it can be rude if you don't cover your mouth.
17	Haggle	People always _____ for lower prices while buying cars.
18	Amble	Some people _____ even though they are in a hurry, and they never run.
19	Sentry	There is always a _____ standing at the castle.
20	Holler	There's no need to _____, I can hear you!
21	Cackle	There was no funny scene in the movie, but he started to _____ like a crazy person.
22	Meddle	Teachers do not want parents to _____ in the school system.
23	Weasel	A _____ is a small animal with a long body.

24	Chastise	My parents always _____ me for my lack of manners.
25	Mauve	When you mix blue, red, and white, you get the color of _____.
26	Bawl	She starts to _____ like a baby when she sees people crying.
27	Mirage	When a _____ occurs in the desert, you can see a pool of water but it is not really there.
28	Hatchet	He cut down trees with his _____.
29	Wick	A _____ burns down and gives off light.
30	Azalea	An _____ is a flowering plant with different colors, but I like pink ones the most.
31	Tetanus	The disease called _____ causes painful muscle stiffness all over the body.
32	Scowl	We started to _____ at the man as he was late in the meeting.
33	Shrew	A _____ is a small rat, but it is dangerous because it can bite you.
34	Trowel	My mother uses a _____ to do some flower gardening.
35	Gazebo	My family like lying down and relaxing in the _____ in the garden.
36	Icicle	In winter, an _____ on the roof often comes crashing down on top of your car.
37	Gourd	A _____ is a vegetable with a hard skin, and it is used as a bottle to drink.
38	Fawn	The women working at the shop always _____ over customers.
39	Quail	A pen decorated with a _____ feather is expensive.
40	Notary	I need a _____ who is a public officer to sign some contracts.
41	Carousel	He gets his luggage off the _____ and opens it up.
42	Faucet	I turned on the _____ to wash my hands.
43	Conch	I can hear the sound of the sea in a _____.
44	Slobber	Dogs _____ all over my hands when I give them food.
45	Toboggan	Children like riding a _____ on the snow in the winter.
46	Snitch	I don't tell anything to my brothers, because they _____ on me to mother.
47	Abacus	Chinese children learned how to do math with an _____.
48	Toupee	He was embarrassed when the _____ on his head blew off in the wind.

Example of the Contextualized Form Recall Test for the Target Item *Trowel* (Study 3)



In the contextualized form recall test, participants were asked to type the appropriate target word at the bottom of the screen to complete the gap in the provided sentence.

Appendix G. Randomization of Posttest Order

Immediate posttest	Type 1		Type 2		Type 3	
	Form Recall	SET A-1	Production	SET B-1	Contextualized	SET C-1
	Production	SET B-1	Contextualized	SET C-1	Form Recall	SET A-1
	Contextualized	SET C-1	Form Recall	SET A-1	Production	SET B-1
	Type 4		Type 5		Type 6	
	Form Recall	SET B-1	Production	SET C-1	Contextualized	SET A-1
	Production	SET C-1	Contextualized	SET A-1	Form Recall	SET B-1
	Contextualized	SET A-1	Form Recall	SET B-1	Production	SET C-1
	Type 7		Type 8		Type 9	
	Form Recall	SET C-1	Production	SET A-1	Contextualized	SET B-1
	Production	SET A-1	Contextualized	SET B-1	Form Recall	SET C-1
	Contextualized	SET B-1	Form Recall	SET C-1	Production	SET A-1
Delayed posttest	Type 10		Type 11		Type 12	
	Form Recall	SET D-1	Production	SET E-1	Contextualized	SET F-1
	Production	SET E-1	Contextualized	SET F-1	Form Recall	SET D-1
	Contextualized	SET F-1	Form Recall	SET D-1	Production	SET E-1
	Type 13		Type 14		Type 15	
	Form Recall	SET E-1	Production	SET F-1	Contextualized	SET D-1
	Production	SET F-1	Contextualized	SET D-1	Form Recall	SET E-1
	Contextualized	SET D-1	Form Recall	SET E-1	Production	SET F-1
	Type 16		Type 17		Type 18	
	Form Recall	SET F-1	Production	SET D-1	Contextualized	SET E-1
	Production	SET D-1	Contextualized	SET E-1	Form Recall	SET F-1
	Contextualized	SET E-1	Form Recall	SET F-1	Production	SET D-1

Immediate posttest	Type 19		Type 20		Type 21	
	Form Recall	SET A-2	Production	SET B-2	Contextualized	SET C-2
	Production	SET B-2	Contextualized	SET C-2	Form Recall	SET A-2
	Contextualized	SET C-2	Form Recall	SET A-2	Production	SET B-2
	Type 22		Type 23		Type 24	
	Form Recall	SET B-2	Production	SET C-2	Contextualized	SET A-2
	Production	SET C-2	Contextualized	SET A-2	Form Recall	SET B-2
	Contextualized	SET A-2	Form Recall	SET B-2	Production	SET C-2
	Type 25		Type 26		Type 27	
	Form Recall	SET C-2	Production	SET A-2	Contextualized	SET B-2
	Production	SET A-2	Contextualized	SET B-2	Form Recall	SET C-2
	Contextualized	SET B-2	Form Recall	SET C-2	Production	SET A-2
Delayed posttest	Type 28		Type 29		Type 30	
	Form Recall	SET D-2	Production	SET E-2	Contextualized	SET F-2
	Production	SET E-2	Contextualized	SET F-2	Form Recall	SET D-2
	Contextualized	SET F-2	Form Recall	SET D-2	Production	SET E-2
	Type 31		Type 32		Type 33	
	Form Recall	SET E-2	Production	SET F-2	Contextualized	SET D-2
	Production	SET F-2	Contextualized	SET D-2	Form Recall	SET E-2
	Contextualized	SET D-2	Form Recall	SET E-2	Production	SET F-2
	Type 34		Type 35		Type 36	
	Form Recall	SET F-2	Production	SET D-2	Contextualized	SET E-2
	Production	SET D-2	Contextualized	SET E-2	Form Recall	SET F-2
	Contextualized	SET E-2	Form Recall	SET F-2	Production	SET D-2

Appendices for Study 2

Appendix 2H. Results of Logistic Mixed-Effects Models Including Time on Task as a Covariate (Immediate and Delayed Posttests)

	Immediate posttest				Delayed posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	2.50	0.56	4.43	.00	-2.42	0.60	-4.02	.00
Learning condition	-1.01	0.36	-2.82	.00	-0.14	0.37	-0.36	.72
Spacing type	-1.27	0.36	-3.53	.00	0.83	0.36	2.27	.02
Time on task	-0.06	0.02	-2.45	.01	-0.02	0.02	-0.71	.48
Learning condition x Spacing type	0.99	0.23	4.24	.00	0.45	0.23	1.96	.05
Learning condition x Time on task	0.03	0.01	1.98	.05	0.02	0.01	1.33	.18
Spacing type x Time on task	0.03	0.02	1.64	.10	0.01	0.02	0.68	.50
Learning condition x Spacing type x Time on task	-0.03	0.01	-2.60	.01	-0.02	0.01	-2.07	.04

Note. Statistically significant at $p < .05$, A model formula for the immediate posttest: Scores ~ LearningCondition * SpacingType * Timeontask + LearningCondition + SpacingType + Timeontask + (SpacingType + LearningCondition | Subject) + (Timeontask | Item), A model formula for the delayed posttest: Scores ~ LearningCondition * SpacingType * Timeontask + LearningCondition + SpacingType + Timeontask + (SpacingType + LearningCondition | Subject) + (Timeontask | Item)

Appendix 2I. Results of the Mean Gains From the Pretest to the Posttest (Immediate and Delayed Posttests)

Table 1

Results of the Mean Gains From the Pretest to the Posttest (Three Test Formats Combined)

	Immediate posttest					Delayed posttest				
			95% CI					95% CI		
	d	*SE*	Lower	Upper	*p*	*d*	*SE*	Lower	Upper	*p*
Control	1.35	0.29	0.78	1.90	.00	0.53	0.26	0.18	1.22	.04
FIB massed	5.22	0.54	4.10	6.20	.00	1.43	0.29	0.84	1.98	.00
FIB spaced	9.01	0.85	7.23	10.55	.00	2.61	0.35	1.90	3.27	.00
FC massed	5.74	0.58	4.53	6.80	.00	1.48	0.29	0.89	2.03	.00
FC spaced	3.91	0.44	3.00	4.72	.00	2.21	0.33	1.55	2.83	.00

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Confidence intervals for Cohen's d were calculated using an effect size calculator: http://www.cem.org/effect-size-calculator (accessed November 2021). The results showed that the mean percentage learning gains for the control group were 11% on the immediate posttest and 2% on the delayed posttest. The mean gains for the fill-in-the-blanks massed and spaced conditions were 69% and 73% on the immediate posttest and 25% and 53% on the delayed posttest. The mean gains for the flashcard massed and spaced conditions were 67% and 64% on the immediate posttest and 20% and 43% on the delayed posttest.

Table 2

Results of the Mean Gains From the Pretest to the Posttest **(***Form recall Test)*

	Immediate posttest					Delayed posttest				
			95% CI					95% CI		
	d	*SE*	Lower	Upper	*p*	*d*	*SE*	Lower	Upper	*p*
Control	1.83	0.31	1.20	2.40	.00	0.53	0.26	0.00	1.03	.05
FIB massed	7.81	0.75	6.24	9.18	.00	2.01	0.32	1.37	2.60	.00
FIB spaced	14.93	1.37	12.21	17.65	.00	3.29	0.39	2.48	4.02	.00
FC massed	14.49	1.33	11.85	17.13	.00	1.67	0.30	1.06	2.23	.00
FC spaced	6.34	0.63	5.03	7.48	.00	3.48	0.41	2.64	4.24	.00

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Table 3

Results of the Mean Gains From the Pretest to the Posttest **(***Contextualized Form Recall Test)*

	Immediate posttest					Delayed posttest				
			95% CI					95% CI		
	d	*SE*	Lower	Upper	*p*	*d*	*SE*	Lower	Upper	*p*
Control	1.33	0.28	0.76	1.88	.00	0.45	0.26	-0.06	0.96	.08
FIB massed	6.71	0.66	5.33	7.91	.00	1.38	0.29	0.80	1.93	.00
FIB spaced	8.62	0.82	6.91	10.25	.00	2.68	0.35	1.95	3.34	.00
FC massed	5.64	0.57	4.45	6.68	.00	1.66	0.30	1.06	2.23	.00
FC spaced	3.57	0.41	2.71	4.33	.00	1.61	0.30	1.01	2.17	.00

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Table 4

Results of the Mean Gains From the Pretest to the Posttest (*Sentence Production Test)*

	Immediate posttest					Delayed posttest				
			95% CI					95% CI		
	d	*SE*	Lower	Upper	*p*	*d*	*SE*	Lower	Upper	*p*
Control	0.91	0.27	0.37	1.43	.00	0.59	0.26	0.06	1.10	.03
FIB massed	2.40	0.34	1.70	3.02	.00	0.87	0.27	0.33	1.39	.00
FIB spaced	4.72	0.50	3.68	5.63	.00	1.83	0.31	1.20	2.40	.00
FC massed	2.62	0.35	1.90	3.27	.00	1.02	0.27	0.47	1.54	.00
FC spaced	2.19	0.32	1.52	2.80	.00	1.34	0.29	0.76	1.88	.00

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Appendix 2J. Comparisons in the Gains Between 5 Groups (Control; Fill-in-the-blanks With Massed and Spaced; Flashcards With Massed and Spaced) From Pretest to Posttest (Three Test Formats Combined)

Comparison	Immediate posttest							Delayed posttest						
	d	variance	*SE*	95% CI		*z*	*p*	*d*	variance	*SE*	95% CI		*z*	*p*
				Lowe	Upper						Lower	Upper		
FIB massed vs. Control	4.66	0.25	0.50	3.68	5.64	9.36	.00	1.40	0.08	0.29	0.84	1.97	4.86	.00
FIB spaced vs. Control	7.59	0.55	0.74	6.14	9.04	10.27	.00	3.05	0.14	0.38	2.31	3.79	8.03	.00
FC massed vs. Control	5.37	0.31	0.55	4.28	6.45	9.69	.00	1.63	0.09	0.30	1.05	2.22	5.47	.00
FC spaced vs. Control	3.54	0.17	0.41	2.73	4.35	8.56	.00	2.72	0.13	0.36	2.02	3.42	7.59	.00
FIB spaced vs. FIB massed	0.31	0.07	0.26	-0.20	0.82	1.19	.24	1.24	0.08	0.28	0.69	1.79	4.40	.00
FIB spaced vs. FC massed	0.55	0.07	0.26	0.03	1.06	2.07	.04	1.74	0.09	0.30	1.15	2.34	5.74	.00
FIB spaced vs. FC spaced	0.61	0.07	0.26	0.09	1.12	2.29	.02	0.47	0.07	0.26	-0.04	0.99	1.80	.07
FIB massed vs. FC massed	0.13	0.07	0.26	-0.38	0.64	0.50	.62	0.28	0.07	0.26	-0.23	0.79	1.07	.28
FIB massed vs. FC spaced	0.30	0.07	0.26	-0.21	0.80	1.14	.26	-0.83	0.07	0.27	-1.36	-0.30	-3.08	.00
FC spaced vs. FC massed	0.21	0.07	0.26	-0.30	0.72	0.80	.42	1.30	0.08	0.28	0.74	1.85	4.57	.00

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Appendix 2K. Comparisons in the Gains Between 5 Groups (Control; Fill-in-the-blanks With Massed and Spaced; Flashcards With Massed and Spaced) From Pretest to Immediate Posttest (Individual Test Format)

Table 1

Results of the Mean Gains From the Pretest to the Posttest **(***Form recall Test)*

Comparison	Form recall					
	d	*SE*	95% CI		*z*	*p*
			Lower	Upper		
Control vs. FIB massed	-4.71	0.50	-3.73	-5.69	-9.39	.00
Control vs. FIB spaced	-7.06	0.69	-5.70	-8.42	-10.17	.00
Control vs. FC massed	-6.55	0.65	-5.27	-7.82	-10.06	.00
Control vs. FC spaced	-4.34	0.47	-3.41	-5.27	-9.18	.00
FIB spaced vs. FIB massed	1.05	0.28	0.51	1.59	3.80	.00
FIB spaced vs. FC massed	0.92	0.27	0.23	1.28	3.37	.00
FIB spaced vs. FC spaced	0.35	0.26	-0.16	0.86	1.33	.18
FIB massed vs. FC massed	-0.51	0.26	-1.03	0.00	-1.96	.05
FIB massed vs. FC spaced	-0.43	0.26	-0.94	0.09	-1.63	.10
FC massed vs. FC spaced	-0.08	0.26	-0.59	0.43	-0.31	.75

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Table 2

Results of the Mean Gains From the Pretest to the Posttest **(***Contextualized Form Recall Test)*

Comparison	Contextualized form recall					
	d	*SE*	95% CI		z	*p*
			Lower	Upper		
Control vs. FIB massed	-5.43	0.56	-6.52	-4.33	-9.71	.00
Control vs. FIB spaced	-6.57	0.65	-7.85	-5.29	-10.06	.00
Control vs. FC massed	-4.40	0.48	-5.34	-3.47	-9.22	.00
Control vs. FC spaced	-2.95	0.37	-3.69	-2.22	-7.91	.00
FIB spaced vs. FIB massed	.00	0.26	-0.51	0.51	.00	1.00
FIB spaced vs. FC massed	1.06	0.28	0.52	1.60	3.84	.00
FIB spaced vs. FC spaced	0.91	0.27	0.38	1.44	3.35	.00
FIB massed vs. FC massed	0.94	0.27	0.41	1.48	3.47	.00
FIB massed vs. FC spaced	0.85	0.27	0.32	1.38	3.15	.00
FC massed vs. FC spaced	0.11	0.26	-0.39	0.62	0.44	.66

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Table 3

Results of the Mean Gains From the Pretest to the Posttest (*Sentence Production Test)*

Comparison	Sentence production					
	d	*SE*	95% CI		*z*	*p*
			Lower	Upper		
Control vs. FIB massed	-1.59	0.30	-2.17	-1.01	-5.38	.00
Control vs. FIB spaced	-2.28	0.33	-2.93	-1.63	-6.88	.00
Control vs. FC massed	-1.82	0.31	-2.42	-1.21	-5.92	.00
Control vs. FC spaced	-1.41	0.29	-1.97	-0.84	-4.88	.00
FIB spaced vs. FIB massed	-0.09	0.26	-0.60	0.42	-0.35	.73
FIB spaced vs. FC massed	-0.36	0.26	-0.87	0.15	-1.38	.17
FIB spaced vs. FC spaced	0.12	0.25	-0.39	0.63	0.46	.65
FIB massed vs. FC massed	-0.22	0.25	-0.72	0.29	-0.83	.41
FIB massed vs. FC spaced	0.17	0.25	-0.34	0.67	0.64	.52
FC massed vs. FC spaced	0.38	0.26	-0.13	0.89	1.45	.15

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Appendix 2L. Comparisons in the Gains Between 5 Groups (Control; Fill-in-the-blanks With Massed and Spaced; Flashcards With Massed and Spaced) From Pretest to Delayed Posttest (Individual Test Format)

Table 1

Results of the Mean Gains From the Pretest to the Posttest **(***Form recall Test)*

Comparison	Form recall					
			95% CI			
	d	*SE*	Lower	Upper	*z*	*p*
Control vs. FIB massed	-1.88	0.31	-2.49	-1.27	-6.06	.00
Control vs. FIB spaced	-3.17	0.39	-3.93	-2.41	-8.17	.00
Control vs. FC massed	-1.52	0.29	-2.09	-0.95	-5.18	.00
Control vs. FC spaced	-3.36	0.40	-4.15	-2.57	-8.38	.00
FIB spaced vs. FIB massed	1.22	0.28	0.67	1.77	4.33	.00
FIB spaced vs. FC massed	1.67	0.30	1.08	2.25	5.56	.00
FIB spaced vs. FC spaced	-0.17	0.26	-0.68	0.34	-0.65	.52
FIB massed vs. FC massed	0.42	0.26	-0.09	0.93	1.61	.11
FIB massed vs. FC spaced	-1.39	0.29	-1.95	-0.83	-4.83	.00
FC massed vs. FC spaced	-1.84	0.31	-2.45	-1.24	-5.98	.00

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Table 2

Results of the Mean Gains From the Pretest to the Posttest **(***Contextualized Form Recall Test)*

Comparison	Contextualized form recall					
	d	*SE*	95% CI		z	*p*
			Lower	Upper		
Control vs. FIB massed	-1.25	0.28	-1.80	-0.70	-4.42	.00
Control vs. FIB spaced	-2.54	0.35	-3.22	-1.86	-7.31	.00
Control vs. FC massed	-1.43	0.29	-2.00	-0.87	-4.96	.00
Control vs. FC spaced	-1.62	0.30	-2.21	-1.04	-5.45	.00
FIB spaced vs. FIB massed	1.10	0.28	0.56	1.64	3.97	.00
FIB spaced vs. FC massed	1.44	0.29	0.87	2.00	4.96	.00
FIB spaced vs. FC spaced	0.79	0.27	0.27	1.32	2.95	.00
FIB massed vs. FC massed	0.20	0.26	-0.31	0.71	0.45	.44
FIB massed vs. FC spaced	-0.31	0.26	-0.82	0.20	-1.19	.24
FC massed vs. FC spaced	-0.55	0.26	-1.07	-0.04	-2.10	.04

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Table 3

Results of the Mean Gains From the Pretest to the Posttest (*Sentence Production Test)*

Comparison	Sentence production					
	d	*SE*	95% CI		*z*	*p*
			Lower	Upper		
Control vs. FIB massed	-0.68	0.27	-1.20	-0.16	-2.57	.01
Control vs. FIB spaced	-1.65	0.30	-2.23	-1.06	-5.52	.00
Control vs. FC massed	-0.75	0.27	-1.27	-0.23	-2.81	.01
Control vs. FC spaced	-1.16	0.28	-1.71	-0.61	-4.15	.00

FIB spaced vs. FIB massed	0.85	0.27	0.32	1.38	3.15	.00
FIB spaced vs. FC massed	1.02	0.27	0.48	1.56	3.71	.00
FIB spaced vs. FC spaced	0.43	0.26	-0.08	0.94	1.66	.10
FIB massed vs. FC massed	0.09	0.25	-0.42	0.60	0.35	.72
FIB massed vs. FC spaced	-0.41	0.26	-0.93	0.10	-1.59	.11
FC massed vs. FC spaced	-0.55	0.26	-1.06	-0.03	-2.09	.04

Note. Statistically significant at $p < .05$, FIB = fill-in-the-blanks group, FC = flashcards group

Appendix 2M. Results of Logistic Mixed-Effects Models for Learning Condition in Each Test Format (Immediate and Delayed Posttests)

	Form recall				Contextualized form recall				Sentence production			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Immediate posttest												
Intercept	-0.37	0.22	-1.70	.09	-0.28	0.30	-0.93	.35	-0.47	0.33	-1.43	.15
LC	-0.07	0.13	-0.56	.58	1.18	0.21	5.60	.00	-0.01	0.21	-0.06	.95
Time on task	0.00	0.01	0.29	.78	-0.00	0.01	-0.08	.94	0.00	0.01	0.46	.64
LC x Time on task	-0.00	0.01	-0.20	.86	-0.01	0.01	-1.22	.22	-0.00	0.01	-0.40	.69
Delayed posttest												

Intercept	-0.03	0.29	-0.10	.92	-1.46	0.36	-4.08	.00	-2.01	0.33	-6.00	.00
LC	0.18	0.18	0.96	.34	0.97	0.23	4.29	.00	0.41	0.22	1.85	.06
Time on task	-0.03	0.02	-1.07	.28	0.00	0.02	0.03	.96	0.00	0.01	0.25	.80
LC x Time on task	-0.01	0.02	-0.61	.54	-0.02	0.01	-1.78	.07	-0.00	0.00	-0.10	.92

Note. Statistically significant at $p < .05$, LC = learning condition, A model formula for each test format in the immediate posttest: Scores ~ LearningCondition * Timeontask + LearningCondition + Timeontask + (LearningCondition | Subject) + (Timeontask | Item), A model formula for each test format in the delayed posttest: Scores ~ LearningCondition * Timeontask + LearningCondition + Timeontask + (LearningCondition | Subject) + (Timeontask | Item)

Table 1

Comparisons in the Gains Between Fill-in-the-blanks and Flashcards (Individual Test Format)

	Immediate posttest						Delayed posttest					
			95% CI						95% CI			
	d	*SE*	Lower	Upper	*z*	*p*	*d*	*SE*	Lower	Upper	*z*	*p*
Form recall	0.03	0.18	-0.32	0.39	0.19	.85	-0.08	0.18	-0.43	0.28	-0.42	.68
Contextualized form recall	-0.94	0.19	-1.32	-0.56	-4.88	.00	-0.48	0.19	-0.85	-0.12	-2.61	.01
Sentence production	-0.30	0.18	-0.66	0.06	-1.65	.10	-0.09	0.18	-0.45	0.27	-0.51	.61

Note. Statistically significant at $p < .05$

Appendix 2N. Results of Logistic Mixed-Effects Models for Feedback Timing (Immediate and Delayed Posttests)

Table 1

Results of Logistic Mixed-Effects Models for Feedback Timing (Both Activities)

	Immediate posttest				Delayed posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	0.91	0.43	0.10	.04	-2.85	0.44	-6.45	.00
Feedback timing	-0.02	0.26	-0.07	.95	0.09	0.26	0.34	.73
Learning condition	0.51	0.27	1.90	.05	0.52	0.26	1.99	.05
Spacing type	-0.04	0.08	-0.42	.67	1.18	0.08	14.14	.00
Time on task	-0.03	0.00	-13.29	.00	-0.02	0.00	-7.00	.00
Feedback timing x Learning condition	-0.14	0.17	-0.83	.41	-0.09	0.16	-0.54	.59

Note. Statistically significant at $p < .05$, A model formula for the immediate posttest: Scores ~ FeedbackTiming * LearningCondition + LearningCondition + SpacingType + Timeontask + (SpacingType + LearningCondition | Subject) + (FeedbackTiming + Timeontask | Item), A model formula for the delayed posttest: Scores ~ FeedbackTiming * LearningCondition + SpacingType + Timeontask + (SpacingType + LearningCondition | Subject) + (FeedbackTiming + Timeontask | Item)

Table 2

Results of Logistic Mixed-Effects Models for Feedback Timing (Fill-in-the-blanks)

	Fill-in-the-blanks	
	Immediate posttest	Delayed posttest

	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	1.77	0.62	2.84	.00	-1.67	0.60	-2.78	.01
Feedback timing	-0.43	0.39	-1.11	.27	-0.20	0.38	-0.51	.61
Spacing type	0.15	0.40	0.37	.71	1.15	0.37	3.15	.00
Time on task	-0.04	0.00	-10.82	.00	-0.02	0.00	-6.32	.00
Feedback timing x Spacing type	0.08	0.25	0.33	.74	0.07	0.23	0.28	.78

Note. Statistically significant at $p < .05$, A model formula for the immediate posttest: Scores ~ FeedbackTiming * SpacingType + Timeontask + (SpacingType | Subject) + (FeedbackTiming + Timeontask | Item), A model formula for the delayed posttest: Scores ~ FeedbackTiming * SpacingType + Timeontask + (SpacingType | Subject) + (FeedbackTiming + Timeontask | Item)

Table 3

Results of Logistic Mixed-Effects Models for Feedback Timing (Flashcards)

	Flashcards							
	Immediate posttest				Delayed posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	1.50	0.59	2.56	.01	-2.32	0.28	-8.39	.00
Feedback timing	-0.01	0.37	-0.04	.97	0.01	0.12	0.08	.94
Spacing type	-0.14	0.36	-0.38	.70	1.12	0.12	9.27	.00
Time on task	-0.03	0.00	-8.05	.00	-0.01	0.00	-3.76	.00
Feedback timing x Spacing type	0.09	0.23	-0.40	.69	-0.17	0.18	-0.91	.36

Note. Statistically significant at $p < .05$, A model formula for the immediate posttest: Scores ~ FeedbackTiming * SpacingType + Timeontask + (SpacingType | Subject) + (FeedbackTiming | Item), A model formula for the delayed posttest: Scores ~ FeedbackTiming * SpacingType + Timeontask + (SpacingType | Subject) + (FeedbackTiming | Item)

Appendices for Study 3

APPENDIX 3H. Results of the Mean Gains from the Pretest to the Posttest

Results of the Mean Gains from the Pretest to the Posttest (Three Test Formats Combined)

	Immediate posttest						Delayed posttest					
			95% CI						95% CI			
	d	*SE*	Lower	Upper	*z*	*p*	*d*	*SE*	Lower	Upper	*z*	*p*
Control	1.91	0.31	0.30	2.52	6.12	.00	0.70	0.27	0.18	1.22	2.62	.01
SP massed	5.64	0.58	4.51	6.76	9.79	.00	1.41	0.29	0.84	1.97	4.88	.00
SP spaced	5.47	0.56	4.37	6.57	9.73	.00	3.05	0.38	2.31	3.79	8.03	.00
FC massed	7.83	0.73	6.34	9.32	10.30	.00	1.90	0.31	1.29	2.51	6.11	.00
FC spaced	4.68	0.50	3.70	5.66	9.38	.00	2.91	0.37	2.19	3.64	7.86	.00

Note. Statistically significant at $p < .05$, SP = sentence production group, FC = flashcards group

Results of the Mean Gains from the Pretest to the Posttest (Form Recall)

	Immediate posttest						Delayed posttest					
			95% CI						95% CI			
	d	*SE*	Lower	Upper	*z*	*p*	*d*	*SE*	Lower	Upper	*z*	*p*
Control	1.83	0.31	1.22	2.43	5.94	.00	0.53	0.26	0.01	1.04	2.00	.05

SP massed	7.86	0.76	6.36	9.35	10.31	.00	1.12	0.28	0.58	1.67	4.04	.00
SP spaced	6.95	0.69	5.61	8.29	10.15	.00	3.19	0.39	2.43	3.95	8.20	.00
FC massed	14.49	1.35	11.85	17.13	10.75	.00	1.67	0.30	1.08	2.26	5.57	.00
FC spaced	6.34	0.63	5.10	7.58	10.00	.00	3.49	0.41	2.68	4.29	8.51	.00

Results of the Mean Gains from the Pretest to the Posttest (Sentence Production)

	Immediate posttest						Delayed posttest					
			95% CI						95% CI			
	d	*SE*	Lower	Upper	*z*	*p*	*d*	*SE*	Lower	Upper	*z*	*p*
Control	0.91	0.27	0.38	1.44	3.35	.00	0.59	0.26	0.07	1.11	2.23	.03
SP massed	2.36	0.34	1.70	3.02	7.01	.00	1.15	0.28	0.60	1.70	4.13	.00
SP spaced	2.67	0.36	1.98	3.37	7.52	.00	2.15	0.32	1.52	2.79	6.63	.00
FC massed	2.62	0.35	1.93	3.31	7.44	.00	1.02	0.27	0.48	1.56	3.72	.00
FC spaced	2.19	0.33	1.55	2.83	6.70	.00	1.34	0.29	0.78	1.90	4.69	.00

Results of the Mean Gains from the Pretest to the Posttest (Contextualized Form Recall)

	Immediate posttest						Delayed posttest					
			95% CI						95% CI			
	d	*SE*	Lower	Upper	*z*	*p*	*d*	*SE*	Lower	Upper	*z*	*p*
Control	1.34	0.29	0.78	1.89	4.67	.00	0.45	0.26	-0.06	0.97	1.74	.08
SP massed	4.77	0.51	3.78	5.76	9.42	.00	1.14	0.28	0.59	1.68	4.08	.00

SP spaced	3.77	0.43	2.93	4.62	8.77	.00	2.28	0.33	1.63	2.93	6.87	.00
FC massed	5.64	0.58	4.51	6.77	9.79	.00	1.66	0.30	1.08	2.25	5.55	.00
FC spaced	3.57	0.42	2.75	4.38	8.59	.00	1.76	0.30	1.16	2.36	5.79	.00

APPENDIX 3I. Comparisons in the Gains Between 5 Groups From Pretest to Posttest

Comparisons in the Gains Between 5 Groups From Pretest to Posttest (Three Test Formats Combined)

Comparison	Immediate posttest							Delayed posttest						
	d	variance	*SE*	95% CI		*z*	*p*	*d*	variance	*SE*	95% CI		*z*	*p*
				Lowe	Upper						Lower	Upper		
SP massed vs. Control	4.18	0.21	0.46	3.27	5.08	9.07	.00	1.14	0.08	0.28	0.59	1.68	4.09	.00
SP spaced vs. Control	4.20	0.21	0.46	3.29	5.11	9.09	.00	2.85	0.13	0.37	2.13	3.56	7.77	.00
FC massed vs. Control	5.37	0.31	0.55	4.28	6.45	9.69	.00	1.63	0.09	0.30	1.05	2.22	5.47	.00
FC spaced vs. Control	3.54	0.17	0.41	2.73	4.35	8.56	.00	2.72	0.13	0.36	2.02	3.42	7.59	.00
SP spaced vs. SP massed	0.14	0.07	0.26	-0.37	0.65	0.53	.59	1.61	0.08	0.30	1.03	2.19	5.42	.00
SP spaced vs. FC massed	0.11	0.07	0.26	-0.40	0.62	0.43	.67	1.41	0.08	0.29	0.85	1.98	4.89	.00
SP spaced vs. FC spaced	0.27	0.07	0.26	-0.24	0.78	1.04	.30	0.10	0.07	0.26	-0.41	0.61	-0.39	.70
SP massed vs. FC massed	-0.05	0.07	0.26	-0.56	0.46	-0.19	.85	-0.28	0.07	0.26	-0.79	0.23	-1.07	.29
SP massed vs. FC spaced	0.15	0.07	0.26	-0.36	0.65	0.57	.57	-1.50	0.09	0.29	-2.07	-0.93	-5.13	.00
FC spaced vs. FC massed	-0.21	0.07	0.26	-0.72	0.30	-0.80	.42	1.30	0.08	0.28	0.74	1.85	4.57	.00

Note. Statistically significant at $p < .05$, SP = sentence production group, FC = flashcards group

APPENDIX 3J. Comparisons in the Gains Between 5 Groups From Pretest to Posttest (Individual Test Format)

Comparisons in the Gains Between 5 Groups From Pretest to Posttest (Form Recall)

Comparison	Immediate posttest						Delayed posttest					
	d	*SE*	95% CI		*z*	*p*	*d*	*SE*	95% CI		*z*	*p*
			Lower	Upper					Lower	Upper		
Control vs. SP massed	-4.70	0.50	-5.67	-3.71	-9.38	.00	-1.01	0.27	-1.55	-0.47	-3.68	.00
Control vs. SP spaced	-4.66	0.50	-5.63	-3.68	-9.36	.00	-3.06	0.38	-3.80	-2.31	-8.04	.00
Control vs. FC massed	-6.55	0.65	-5.27	-7.82	-10.06	.00	-1.52	0.29	-2.09	-0.95	-5.18	.00
Control vs. FC spaced	-4.34	0.47	-3.41	-5.27	-9.18	.00	-3.36	0.40	-4.15	-2.57	-8.38	.00
SP spaced vs. SP massed	0.58	0.26	0.06	1.09	2.19	.03	1.57	0.30	0.99	2.15	5.31	.00
SP spaced vs. FC massed	0.17	0.26	-0.33	0.68	0.67	.50	1.49	0.29	0.92	2.07	5.12	.00
SP spaced vs. FC spaced	0.07	0.26	-0.44	0.57	0.25	.80	-0.37	0.26	-0.88	0.14	-1.43	.15
SP massed vs. FC massed	-0.60	0.26	-1.11	-0.08	-2.25	.03	-0.21	0.26	-0.72	0.30	-0.80	.42
SP massed vs. FC spaced	-0.48	0.26	-0.99	0.04	-1.83	.07	-1.89	0.31	-2.50	-1.29	-6.10	.00
FC massed vs. FC spaced	-0.08	0.26	-0.59	0.43	-0.31	.75	-1.84	0.31	-2.45	-1.24	-5.98	.00

Form Recall Immediate and Delayed Posttests

When examining the results of the form recall test format (scores out of 8) in the immediate posttest, the results showed that the four

experimental (sentence production massed and spaced; flashcards massed and spaced) groups contributed to significantly greater gains than the control group (*ps* < .001). The comparisons between the four experimental groups showed that the sentence production spaced condition had statistically greater gains than the sentence production massed condition ($z = 2.19$, $p = .03$), but the sentence production spaced condition was as effective as the flashcard massed ($z = 0.67$, $p = .50$) and spaced conditions ($z = 0.25$, $p = .80$). The sentence production massed condition had statistically greater gains than the flashcard massed condition ($z = -2.25$, $p = .03$), but the sentence production massed condition was as effective as the flashcard spaced condition ($z = -1.83$, $p = .07$). There was no significant difference between the flashcard massed and spaced conditions ($z = -0.31$, $p = .75$).

When examining the results of the form recall test format in the delayed posttest, the results showed that the four experimental groups contributed to significantly greater gains than the control group (*ps* < .001). The sentence production spaced condition had statistically greater gains than the sentence production massed ($z = 5.31$, $p < .001$) and flashcard massed conditions ($z = 5.12$, $p < .001$), but the sentence production spaced condition was as effective as the flashcard spaced condition ($z = -1.43$, $p = .15$). The flashcard spaced condition had statistically greater gains that the sentence production massed ($z = -6.10$, $p < .001$) and flashcard massed conditions ($z = -5.98$, $p < .001$), but the sentence production massed condition was as effective as the flashcard massed condition ($z = -0.80$, $p = .42$).

Comparisons in the Gains Between 5 Groups From Pretest to Posttest (Sentence Production)

Comparison	Immediate posttest						Delayed posttest					
	d	*SE*	95% CI		*z*	*p*	*d*	*SE*	95% CI		*z*	*p*
			Lower	Upper					Lower	Upper		
Control vs. SP massed	-1.72	0.30	-2.31	-1.12	-5.68	.00	-0.81	0.27	-1.33	-0.28	-3.01	.00
Control vs. SP spaced	-1.86	0.31	-2.47	-1.26	-6.02	.00	-1.91	0.31	-2.52	-1.30	-6.13	.00
Control vs. FC massed	-1.82	0.31	-2.42	-1.21	-5.92	.00	-0.75	0.27	-1.27	-0.23	-2.81	.01
Control vs. FC spaced	-1.41	0.29	-1.97	-0.84	-4.88	.00	-1.16	0.28	-1.71	-0.61	-4.15	.00
SP spaced vs. SP massed	-0.06	0.26	-0.57	0.44	-0.24	.81	1.21	0.28	0.66	1.76	4.32	.00
SP spaced vs. FC massed	0.04	0.26	-0.47	0.54	0.14	.89	1.10	0.28	0.56	1.65	3.98	.00
SP spaced vs. FC spaced	0.42	0.26	-0.10	0.93	1.59	.11	0.42	0.26	-0.09	0.93	1.62	.11
SP massed vs. FC massed	0.10	0.26	-0.41	0.60	0.38	.71	-0.05	0.26	-0.56	0.46	-0.20	.85
SP massed vs. FC spaced	0.44	0.26	-0.07	0.96	1.70	.09	-0.62	0.26	-1.13	-0.10	-2.33	.02
FC massed vs. FC spaced	0.38	0.26	-0.13	0.89	1.45	.15	-0.55	0.26	-1.06	-0.03	-2.09	.04

Sentence Production Immediate and Delayed Posttests

When examining the results of the sentence production test format (scores out of 8) in the immediate posttest, the results showed that the four experimental (sentence production massed and spaced; flashcards massed and spaced) groups contributed to significantly greater gains than the control group (*ps* < .001). The comparisons between the four experimental groups showed that there were no significant differences across the groups (*ps* ≥ .09).

When examining the results of the sentence production test format in the delayed posttest, the results showed that the four experimental groups

contributed to significantly greater gains than the control group (*ps* ≤ .01). The comparisons between the four experimental groups showed that the sentence production spaced condition had statistically greater gains than the sentence production massed ($z = 4.32$, $p < .001$) and flashcard massed conditions ($z = 3.98$, $p < .001$), but the sentence production spaced condition was as effective as the flashcard spaced condition ($z = 1.59$, $p = .11$). The flashcard spaced condition had statistically greater gains than the sentence production massed ($z = -2.33$, $p = .02$) and flashcard massed conditions ($z = -2.09$, $p = .04$), but the sentence production massed condition was as effective as the flashcard massed condition ($z = -0.20$, $p = .85$).

Comparisons in the Gains Between 5 Groups From Pretest to Posttest (Contextualized Form Recall)

Comparison	Immediate posttest						Delayed posttest					
	d	*SE*	95% CI		z	*p*	*d*	*SE*	95% CI		*z*	*p*
			Lower	Upper					Lower	Upper		
Control vs. SP massed	-3.88	0.44	-4.74	-3.02	-8.85	.00	-0.94	0.27	-1.47	-0.40	-3.45	.00
Control vs. SP spaced	-3.15	0.39	-3.91	-2.39	-8.15	.00	-2.11	0.32	-2.74	-1.48	-6.55	.00
Control vs. FC massed	-4.40	0.48	-5.34	-3.47	-9.22	.00	-1.43	0.29	-2.00	-0.87	-4.96	.00
Control vs. FC spaced	-2.95	0.37	-3.69	-2.22	-7.91	.00	-1.62	0.30	-2.21	-1.04	-5.45	.00
SP spaced vs. SP massed	-0.04	0.26	-0.54	0.47	-0.14	.89	1.16	0.28	0.61	1.71	4.15	.00
SP spaced vs. FC massed	0.07	0.26	-0.43	0.58	0.28	.78	0.87	0.27	0.34	1.40	3.21	.00
SP spaced vs. FC spaced	0.16	0.26	-0.35	0.66	0.61	.54	0.23	0.26	-0.28	0.74	0.89	.37
SP massed vs. FC massed	0.13	0.26	-0.38	0.64	0.50	.62	-0.36	0.26	-0.87	0.16	-1.27	.17
SP massed vs. FC spaced	0.21	0.26	-0.30	0.72	0.81	.42	-0.83	0.27	-1.36	-0.30	-3.08	.00
FC massed vs. FC spaced	0.11	0.26	-0.39	0.62	0.44	.66	-0.55	0.26	-1.07	-0.04	-2.10	.04

Contextualized Form Recall Immediate and Delayed Posttests

When examining the results of the contextualized form recall test format (scores out of 8) in the immediate posttest, the results showed that the four experimental (sentence production massed and spaced; flashcards massed and spaced) groups contributed to significantly greater gains than the control group (*ps* < .001). The comparisons between the four experimental groups showed that there were no significant differences across the groups (*ps* ≥ .42).

When examining the results of the contextualized form recall test format in the delayed posttest, the results showed that the four experimental groups contributed to significantly greater gains than the control group (*ps* < .001). The comparisons between the four experimental groups showed that the sentence production spaced condition had statistically greater gains than the sentence production massed ($z = 4.15$, $p < .001$) and flashcard massed conditions ($z = 3.21$, $p < .001$), but the sentence production spaced condition was as effective as the flashcard spaced condition ($z = 0.89$, $p = .37$). The flashcard spaced condition had statistically greater gains than the sentence production massed ($z = -3.08$, $p < .001$) and flashcard massed conditions ($z = -2.10$, $p = .04$), but the sentence production massed condition was as effective as the flashcard massed condition ($z = -1.27$, $p = .17$).

APPENDIX 3K. Comparisons in the Gains Between Sentence Production and Flashcards

Comparisons in the Gains Between Sentence production and Flashcards (Individual Test Format)

	Immediate posttest						Delayed posttest					
			95% CI						95% CI			
	d	*SE*	Lower	Upper	*z*	*p*	*d*	*SE*	Lower	Upper	*z*	*p*
Form recall	0.18	0.18	-0.17	0.54	1.01	.31	0.22	0.18	-0.14	0.58	1.23	.22
Sentence production	-0.25	0.18	-0.60	0.11	-1.34	.18	-0.21	0.18	-0.57	0.15	-1.14	.26
Contextualized form recall	-0.15	0.18	-0.51	0.21	-0.81	.42	0.01	0.18	-0.35	0.36	0.03	.98

Note. Statistically significant at $p < .05$**,** we combined massed and spaced learning for each activity (sentence production and flashcards).

APPENDIX 3L. Results of Logistic Mixed-Effects Models for Feedback Timing Including Time on Task as a Covariate (Both Activities)

	Immediate posttest				Delayed posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	1.56	0.24	6.62	.00	-2.58	0.25	-10.52	.00
Feedback timing	-0.27	0.11	-2.40	.02	-0.03	0.11	-0.29	.77
Learning condition	0.03	0.01	0.82	.42	-0.03	0.04	-0.71	.48
Spacing type	-0.05	0.08	-0.60	.55	1.31	0.09	14.94	.00
Time on task	-0.04	0.01	-4.40	.00	-0.02	0.01	-2.13	.03
Feedback timing x Time on task	0.01	0..01	0.84	.40	0.00	0.00	0.86	.39

Note. Statistically significant at $p < .05$, A model formula for the immediate posttest: Scores ~ FeedbackTiming * Timeontask + LearningCondition + SpacingType + (SpacingType + LearningCondition | Subject) + (FeedbackTiming + Timeontask | Item), A model formula for the delayed posttest: Scores ~ FeedbackTiming * Timeontask + LearningCondition + SpacingType + (SpacingType + LearningCondition | Subject) + (FeedbackTiming + Timeontask | Item)

APPENDIX 3M. Results of Logistic Mixed-Effects Models for Feedback Timing (Each Activity)

Results of Logistic Mixed-Effects Models for Feedback Timing (Sentence Production)

	Sentence production							
	Immediate posttest				Delayed posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	2.09	0.59	3.52	.00	-4.03	0.72	-5.61	.00
Feedback timing	-0.75	0.37	-2.04	.05	0.62	0.44	1.41	.16
Spacing type	-0.29	0.38	-0.77	.44	2.03	0.41	4.91	.00
Time on task	-0.04	0.00	-8.48	.00	-0.01	0.00	-2.04	.04
Feedback timing x Spacing type	0.33	0.23	1.40	.16	-0.35	0.26	-1.35	.18

Note. Statistically significant at $p < .05$, A model formula for the immediate posttest: Scores ~ FeedbackTiming * SpacingType + Timeontask + (SpacingType | Subject) + (FeedbackTiming + Timeontask | Item), A model formula for the delayed posttest: Scores ~ FeedbackTiming * SpacingType + Timeontask + (SpacingType | Subject) + (FeedbackTiming + Timeontask | Item)

Results of Logistic Mixed-Effects Models for Feedback Timing (Flashcards)

	Flashcards							
	Immediate posttest				Delayed posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	1.50	0.59	2.56	.01	-2.32	0.28	-8.39	.00
Feedback timing	-0.01	0.37	-0.04	.97	0.01	0.12	0.08	.94
Spacing type	-0.14	0.36	-0.38	.70	1.12	0.12	9.27	.00
Time on task	-0.03	0.00	-8.05	.00	-0.01	0.00	-3.76	.00
Feedback timing x Spacing type	0.09	0.23	-0.40	.69	-0.17	0.18	-0.91	.36

Note. Statistically significant at $p < .05$, A model formula for the immediate posttest: Scores ~ FeedbackTiming * SpacingType + Timeontask + (SpacingType

| Subject) + (FeedbackTiming | Item), A model formula for the delayed posttest: Scores ~ FeedbackTiming * SpacingType + Timeontask + (SpacingType | Subject) + (FeedbackTiming | Item)

Appendices for Ethic Approval

NMREB Initial Application (May 13, 2020)

Western Research

Date: 13 May 2020

To: Dr. Stuart Webb

Project ID: 115616

Study Title: The Effects of Spaced Practice on Second Language Vocabulary Learning

Short Title: Second Language Vocabulary Learning

Application Type: NMREB Initial Application

Review Type: Delegated

Full Board Reporting Date: June 5 2020

Date Approval Issued: 13/May/2020

REB Approval Expiry Date: 13/May/2021

Dear Dr. Stuart Webb

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the WREM application form for the above mentioned study, as of the date noted above. NMREB approval for this study remains valid until the expiry date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

This research study is to be conducted by the investigator noted above. All other required institutional approvals must also be obtained prior to the conduct of the study.

Documents Approved:

Document Name	Document Type	Document Date	Document Version
Debriefing_Form (Korean learners of English)	Debriefing Letter		
Flyer for Korean Participants	Recruitment Materials		
In Class Recruitment	Oral Script	07/May/2020	CLEAN
Language_Background_Questionnaire__Korean_version	Online Survey	07/May/2020	clean
Language_Background_Questionnaire_English_version	Online Survey	07/May/2020	clean
Letter of Information (Korean Learners of English)_Translated in Korean	Translated Documents	07/May/2020	CLEAN
Letter-of-Information-and-Consent (Korean Learners of English) -	Written Consent/Assent		

Documents Acknowledged:

Document Name	Document Type	Document Date	Document Version
Research Plan	Supplementary Tables/Figures		
Translation-certificate-YXTHY_1	Translation Certificate	07/May/2020	for all materials

No deviations from, or changes to the protocol should be initiated without prior written approval from the NMREB, except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Kelly Patterson, Research Ethics Officer on behalf of Dr. Randal Graham, NMREB Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Continuing Ethics (extended to May 13, 2022)

Western Research

Date: 12 May 2021

To: Dr. Stuart Webb

Project ID: 115616

Study Title: The Effects of Spaced Practice on Second Language Vocabulary Learning

Application Type: Continuing Ethics Review (CER) Form

Review Type: Delegated

Date Approval Issued: 12/May/2021

REB Approval Expiry Date: 13/May/2022

Dear Dr. Stuart Webb,

The Western University Non-Medical Research Ethics Board has reviewed this application. This study, including all currently approved documents, has been re-approved until the expiry date noted above.

REB members involved in the research project do not participate in the review, discussion or decision.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Please do not hesitate to contact us if you have any questions.

Sincerely,

The Office of Human Research Ethics

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).