Nakata, T., Suzuki, Y., & He, X. (2022). Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing? *Language Learning*. https://doi.org/10.1111/lang.12553

This is the peer reviewed version of the following article: Costs and Benefits of Spacing for Second Language Vocabulary Learning: Does Relearning Override the Positive and Negative Effects of Spacing?, which has been published in final form at https://doi.org/10.1111/lang.12553. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley's version of record on Wiley Online Library and any embedding, framing or otherwise making available the article or pages thereof by third parties from platforms, services and websites other than Wiley Online Library must be prohibited.

Costs and Benefits of Spacing for Second Language Vocabulary Learning: Does Relearning Override the Positive and Negative Effects of Spacing?

Tatsuya Nakata,[a] Yuichi Suzuki,[b] and Xuehong (Stella) He[c]

[a]Rikkyo University [b]Kanagawa University [c]Nagoya University of Commerce and Business

CRediT author statement – **Tatsuya Nakata**: conceptualization (equal); data curation (supporting); funding acquisition (lead); methodology (equal); project administration (lead); software (supporting); supervision (lead); writing – original draft (lead); writing – review & editing (lead). **Yuichi Suzuki**: conceptualization (equal); formal analysis (lead); methodology (equal); project administration (supporting); software (supporting); visualization (lead); writing – original draft (supporting); writing – review & editing (supporting). **Xuehong (Stella) He**: data curation (lead); software (lead); validation (lead); writing – review & editing (supporting).

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

This research was supported by JSPS KAKENHI grant (grant number 22K00743) and Rikkyo University Special Fund for Research awarded to the first author. We greatly appreciate the invaluable suggestions given by five anonymous reviewers and the handling editors Dr. Emma Marsden and Dr. Pavel Trofimovich. We would also like to thank Dr. Keiko Hanzawa for her help with data collection and Dr. Phil Bennett for proofreading earlier versions of the manuscript.

Correspondence concerning this article should be addressed to Tatsuya Nakata, College of Intercultural Communication, Rikkyo University, 3-34-1 Nishi-Ikebukuro, Toshima-ku, Tokyo 171-8501, Japan. Email: nakata@rikkyo.ac.jp

The handling editors for this article were Emma Marsden.

Abstract

Research has suggested that long spacing (i.e., temporal intervals) within a training session facilitates second language vocabulary learning. Studies, however, have been limited to treatment that involved sessions for only initial learning but not subsequent relearning. Furthermore, most studies have investigated only the benefits of spacing without considering its potential costs (i.e., increased duration of the treatment). In our study, we examined the benefits and costs of within-session spacing for both initial learning and relearning. In this study, 170 Japanese-speaking university students learned 20 English–Japanese word pairs using one of the following four combinations of initial and relearning spacing: long–long, long–short, short–long, and short–short spacing. The results showed that introducing long spacing for both initial learning and relearning (long–long) led to better long-term retention and higher efficiency scores (i.e., number of words learned per trial) despite the increased duration of the treatment. These findings suggest that the benefits of long spacing outweigh its costs.

Keywords vocabulary learning; spaced learning; relearning; lag effect; distributed practice; relearning override effect

Introduction

Given the importance of lexical knowledge for second language (L2) learning and use, how vocabulary development can be facilitated is an important question for researchers, teachers, and students. One way to increase vocabulary learning would be to introduce temporal spacing between multiple encounters of a given target item. Previous studies have suggested that spaced learning in which exposure to a given item is repeated after certain intervals facilitates retention relative to massed learning in which a given item is repeated multiple times without any intervals between exposures (Karpicke & Bauernschmidt, 2011; Nakata,

2015; Nakata & Elgort, 2021). The advantage of spaced learning over massed learning is referred to as the spacing effect. A separate but related phenomenon to the spacing effect is the lag effect (Rogers, 2017, 2022). The lag effect is concerned with the question of whether varying interval lengths between exposures have differential effects on retention (e.g., whether longer spacing between repetitions facilitates learning more than does shorter spacing).

When the effects of spacing are discussed, it would be valuable to distinguish between two types of spacing: between-session spacing and within-session spacing (Kang et al., 2014; Kornell, 2009). Between-session spacing refers to the amount of spacing between separate training sessions. Within-session spacing, in contrast, refers to the amount of spacing between encounters of a given item in a single training session. Although L2 vocabulary studies manipulating between-session spacing have produced somewhat inconsistent results (Bahrick et al., 1993; Cepeda et al., 2009; Rogers & Cheung, 2020; Serrano & Huang, 2018), most studies have suggested that longer within-session spacing facilitates L2 vocabulary learning (Karpicke & Bauernschmidt, 2011; Nakata, 2015; Nakata & Suzuki, 2019; Nakata & Webb, 2016; Pashler et al., 2003; Pyc & Rawson, 2009). The findings of these studies have been valuable because they provided useful guidelines for how teachers and students can capitalize on the advantages of spacing to enhance vocabulary learning inside and outside the classroom.

At the same time, vocabulary studies on within-session spacing have had at least two limitations: (a) In most studies that have manipulated within-session spacing, the treatment involved only a single training session, and (b) most vocabulary studies on within-session spacing have investigated benefits, but not costs, of spacing. Our study aimed to assess benefits and costs of different within-session spacing schedules when the treatment involved both initial learning and subsequent relearning sessions. In our study, we defined relearning

as opportunities for learners to review previously studied items in a subsequent, separate training session (Rawson & Dunlosky, 2013; Rawson et al., 2018; Vaughn et al., 2016). We did not, therefore, consider learners' exposure to previously studied items within the same session (e.g., receiving the correct answer as feedback immediately after retrieval attempts) as a form of relearning.

Background Literature

Although the findings of previous L2 vocabulary studies on within-session spacing have been valuable, one limitation has been that most studies involved only a single training session and did not include subsequent relearning sessions (Karpicke & Bauernschmidt, 2011; Nakata, 2015; Nakata & Suzuki, 2019; Nakata & Webb, 2016; Pashler et al., 2003; Pyc & Rawson, 2009). In these studies, learners were introduced to and practiced novel lexical items in one session that lasted between approximately 17 minutes (Nakata & Webb, 2016) and 90 minutes (Pyc & Rawson, 2009), but the lexical items were never reviewed on later days. When researchers examine the effects of within-session spacing, it would be valuable for them to include relearning sessions for at least two reasons. First, cognitive psychology research has suggested that, although long spacing initially leads to better retention than does short spacing, the positive effects of spacing may attenuate significantly after relearning sessions, a phenomenon known as the relearning override effect (Rawson & Dunlosky, 2013; Rawson et al., 2018). Given this possibility, it would be useful to examine whether the benefits of within-session spacing persist after relearning. Second, vocabulary learning in real-life situations often involves not only an initial learning session but also relearning sessions on later days. As a result, examining the effects of within-session spacing for treatments that involve both initial learning and subsequent relearning reflects authentic situations and increases ecological validity.

When learners study vocabulary in multiple sessions, how much within-session spacing should be used in each session to maximize their learning? Studies on lag effects that found an advantage for longer within-session spacing over shorter spacing have suggested that long within-session spacing should be used for both initial learning and relearning sessions. The study phase retrieval theory (Toppino & Bloom, 2002) and the retrieval effort hypothesis (Pyc & Rawson, 2009), in contrast, predict that initial learning should involve relatively short within-session spacing, whereas relearning should involve relatively long within-session spacing. According to the study phase retrieval theory, successful retrievals enhance memory more than do unsuccessful retrievals. The retrieval effort hypothesis suggests that effortful but successful retrievals enhance learning more than successful but easy retrievals. Taken together, these two theories predict that, in initial learning when memory for the novel L2 words is still weak, relatively short within-session spacing should be used to avoid unsuccessful retrieval. In subsequent relearning sessions, longer within-session spacing may be used because, during relearning, retrievals are likely to be successful even after longer spacing due to initial learning strengthening learners' memory for novel L2 words and facilitating subsequent retrieval of the words. Since long spacing introduces more retrieval difficulty than does short spacing, longer within-session spacing in relearning sessions may allow learners to benefit from the positive effects of not only successful retrieval but also greater retrieval effort.[1]

L2 Vocabulary Research on the Relearning Override Effect

To our knowledge, a study conducted by Rawson et al. (2018) has been the only investigation into the relearning override effect in L2 vocabulary learning. In their Experiment 1, 88 American undergraduate students studied 48 Lithuanian words. Participants were randomly assigned to one of three groups with different spacing schedules: Lag 7, Lag 15, and Lag 47

(where each number designates the number of intervening, unrelated words). The experiment was administered on a computer, and the treatment in all three groups consisted of an initial learning session and four weekly relearning sessions. In the initial learning session, participants studied half of the target items until they correctly retrieved the items once, and they studied the other half of the items until they correctly retrieved them three times.

In the Lag 47 group, the treatment consisted of multiple cycles of 48 items, and each item was presented for retrieval practice only once in each cycle. Target words that reached the criterion of one or three correct retrievals were dropped from subsequent cycles. The treatment continued until all target items reached the predetermined criterion (i.e., one or three correct retrievals) or until the time limit (58 min) expired. In the Lag 7 group, the 48 target items were divided into six blocks of eight items and repeated. When all eight items in a given block reached the predetermined criterion, the next block of eight items was practiced. In the Lag 15 group, the 48 target items were repeated in three blocks of 16 items and studied to the criterion of one or three correct retrievals. The treatment in the Lag 15 group, hence, involved shorter within-session spacing than in the Lag 47 group but longer within-session spacing than in the Lag 7 group. One week after the initial learning session, the first relearning session was conducted. Three further relearning sessions were also given, with a 1-week interval between each. In each of the relearning sessions, the same schedule that was used for the Lag 47 group in the initial learning session was used for all groups. As a result, the treatment in the Lag 7 and Lag 15 groups involved relatively short within-session spacing (Lag 7 or 15) for initial learning but relatively long within-session spacing (Lag 47) for relearning, whereas the treatment in the Lag 47 group involved relatively long within-session spacing (Lag 47) for both initial learning and relearning.

Results of Experiment 1 conducted by Rawson et al. (2018) suggested that, although long spacing initially led to better retention than did short spacing, the positive effects of

spacing attenuated after each relearning session and were lost completely on the final posttest conducted approximately 3 weeks after the last relearning session. Rawson et al.'s Experiment 2 that included three relearning sessions instead of four also suggested that relearning reduced the positive effects of spacing over time, replicating the relearning override effect. The findings reported by Rawson et al. (2018) are valuable because they suggested that it may not necessarily be desirable to use long within-session spacing for initial learning as long as such initial learning is followed by relearning, challenging the widely-held view that long within-session spacing should be used at all times. At the same time, Rawson et al.'s Experiment 2 suggested that the Lag 47 schedule that involved long within-session spacing (Lag 47) for both initial learning and relearning led to significantly higher posttest scores than did the Lag 7 schedule that involved short within-session spacing (Lag 7) for initial learning but long spacing (Lag 47) for relearning. These findings were inconsistent with the prediction (based on the study phase retrieval theory and retrieval effort hypothesis) that gradually increasing within-session spacing facilitates retention.

Although the findings reported by Rawson et al. (2018) have been valuable, within-session spacing was only manipulated in their study during the initial learning session and not during relearning sessions. As a result, their study did not allow the optimal length of within-session spacing that should be used for relearning to be identified.

Benefits and Costs of Within-Session Spacing

In addition to failing to examine the relearning override effect, another limitation of vocabulary studies on within-session spacing has been that they examined the benefits (i.e., retention) but not potential costs (i.e., amount of time needed) of spacing (Karpicke & Bauernschmidt, 2011; Pashler et al., 2003; Pyc & Rawson, 2009). Rawson and Dunlosky (2013) pointed out that, although long spacing has the potential to enhance retention because

it increases the amount of time needed, it may not necessarily be cost-effective, for instance, if learners want to learn L2 words to the criterion of one correct retrieval. When L2 words are repeated after long intervals, learners' retrieval attempts tend to be error-prone because target items are likely to be encountered for retrieval practice only after forgetting has occurred. In contrast, when the treatment involves shorter intervals between repetitions of a given item, learners' retrieval is likely to be more successful because retrieval attempts may occur before memory decays. As a result, when the treatment involves long spacing, learners may require more retrieval attempts and thus more time until all items reach the criterion of one correct retrieval.

At the same time, it is possible that, despite the additional initial cost, long within-session spacing in the initial learning session speeds up subsequent relearning. This is because using longer within-session spacing for initial learning often leads to better long-term retention than shorter spacing (lag effect). As a result, although long-spaced items may reach the criterion more slowly than short-spaced items in the initial learning session, they may reach the criterion more quickly in subsequent relearning sessions, thus compensating for the larger initial cost (Rawson & Dunlosky, 2013).

The above discussion has suggested that when investigating the effects of within-session spacing, it would be valuable for researchers to examine the extent to which relearning attenuates not only the benefits but also the costs of spacing over time. Although no L2 vocabulary study has done this to date, Rawson and Dunlosky (2013, Experiment 3) examined both benefits and costs of within-session spacing for first language (L1) vocabulary learning using the relearning paradigm. Results from their study suggested that the costs associated with spacing (i.e., increased time) may be justified given its benefits (i.e., better retention). Although the findings reported by Rawson and Dunlosky were valuable, since their investigations focused on the acquisition of L1 vocabulary, it was unclear to what extent

their results could be generalized to L2 vocabulary learning. Furthermore, Rawson and Dunlosky compared the effects of pure massing (which does not involve any spacing) and spacing. Yet, considering that learners seldom use pure massing (Koval, 2019; Rogers & Cheung, 2020), the treatment adopted in their experiment may not necessarily reflect authentic learning. Instead, it would be more informative to compare the effects of different amounts of spacing (e.g., relatively short vs. long).

The Current Study

L2 vocabulary studies on lag effects have suggested that longer within-session spacing facilitates learning. At the same time, as our review of the literature indicated, previous studies on within-session spacing have had at least two limitations that we noted: (a) In most studies manipulating within-session spacing, the treatment involved only a single training session, and (b) most vocabulary studies on within-session spacing have investigated benefits, but not costs, of spacing. With the limitations of previous research in mind, the purpose of our study was to assess the benefits and costs of different within-session spacing schedules using the relearning paradigm. Unlike Rawson et al.'s (2018) study, we manipulated within-session spacing schedules for not only initial learning but also for subsequent relearning. This allowed us to identify the within-session spacing schedules that should be used for initial learning and relearning to optimize vocabulary learning.

We asked two research questions in our study:

1. To what extent do variations in the length of within-session spacing for initial learning and relearning affect L2 vocabulary learning?
2. To what extent do variations in the length of within-session spacing for initial learning and relearning affect costs associated with L2 vocabulary learning?

On the basis of previous research, we formulated two hypotheses regarding our research

questions:

1. Using long within-session spacing for both initial learning and relearning facilitates L2 vocabulary learning.
2. The benefits of using long within-session spacing for both initial learning and relearning outweigh its costs.

Hypothesis 1 concerned how much within-session spacing should be used for the initial learning and relearning sessions. Studies on lag effects that had found an advantage for longer spacing over shorter spacing indicated that long within-session spacing should be used for both initial learning and relearning sessions. The study phase retrieval theory (Toppino & Bloom, 2002) and retrieval effort hypothesis (Pyc & Rawson, 2009), in contrast, predict that long spacing should not be introduced until relearning sessions. Hypothesis 1 predicted that using long within-session spacing for both initial learning and relearning facilitates L2 vocabulary learning. This was because the findings reported by Rawson et al. (2018, Experiment 2) suggested that the Lag 47 schedule (which involved long within-session spacing for both initial learning and relearning) led to significantly higher posttest scores than did the Lag 7 schedule (which involved short within-session spacing for initial learning but long spacing for relearning), that is, the findings failed to support the prediction that gradually increasing within-session spacing facilitates retention.

Hypothesis 2 predicted that the benefits of using long spacing for both initial learning and relearning would outweigh its costs on the basis of results reported by Rawson and Dunlosky (2013). Their Experiment 3 suggested that costs associated with spacing (i.e., increased time) may be justified given its benefits (i.e., better retention). Although their experiments focused on the acquisition of L1 vocabulary, we expected their findings to apply also to L2 vocabulary learning.

Method

Full preregistration materials for this study are available in a publicly accessible OSF profile (https://osf.io/t29ka).

Participants

The participants were 170 students at a university in Japan. We assigned them to one of the following four groups: short–short, short–long, long–short, and long–long. The short–short group used a short spacing schedule for both initial learning and relearning sessions. The short–long group used a short spacing schedule for the initial learning session and a long spacing schedule for the relearning session. The long–short group used a long spacing schedule for the initial learning session and a short spacing schedule for the relearning session. The long–long group used a long spacing schedule for both initial learning and relearning sessions (see below for details).

We determined the number of participants for each group (170/4 = 42.5) on the basis of an accuracy in effect size estimation approach (Norouzian, 2020) rather than using a traditional power analytic approach. One potential disadvantage of a traditional power analysis is that it is driven by the binary significant/nonsignificant interpretation of findings. In other words, although a power analysis may be useful for distinguishing between statistically significant and nonsignificant effects, it is not necessarily helpful for the accurate estimation of effect sizes of interest. In contrast, the effect size estimation approach, which is aimed at producing sufficiently narrow confidence intervals for effect sizes, enables researchers to accurately estimate the size of effects, which is more informative than the dichotomous statistical significance.

Norouzian (2020) has recommended using meta-analyzed effect sizes for estimating confidence intervals of effect sizes in a specific research domain. Due to the absence of a

published meta-analytic study on the effects of distributed practice on L2 vocabulary learning, we first conducted a small-scale meta-analysis.[2] Since we found no studies that had examined effects of distributed practice on L2 vocabulary learning by manipulating both initial and relearning spacing, our aim was to establish a benchmark effect size for lag effects in a single treatment session. We included empirical studies meeting the following criteria in our meta-analysis. The studies had to have:

- examined L2 vocabulary learning,
- used a paired-associate learning paradigm,
- examined lag effects by comparing shorter and longer spacing conditions within a single treatment session (excluding studies examining spacing effects by comparing massed and spaced conditions),
- administered posttests after a delay of 24 hours or greater, and
- manipulated spacing between participants, not within.

We included the last criterion because a learning condition that is manipulated within participants tends to yield larger effect sizes than does a learning condition that is manipulated between participants (Plonsky & Oswald, 2014). Because spacing was a between-participant variable in our study, we excluded studies manipulating spacing within participants. The results of the meta-analysis indicated that the overall fixed-effect of lag effects as measured by Hedges's *g* was 0.88 (95% CI [0.71, 1.06], median CI width = 1.11; for details, see Appendix S1 in the Supporting Information online). We considered this effect size to be medium according to field-specific guidelines proposed by Plonsky and Oswald (2014) for L2 research, and this effect size estimate suggested that longer within-session spacing leads to better L2 vocabulary learning than does shorter spacing. The effect size estimation approach suggested that each of the four groups in this study should consist of 34 participants for us to obtain the mean effect size (0.88) with the median confidence interval

width (1.11) with 99% assurance. Given the expected attrition of participants, we recruited 42 or 43 participants (34 + 8 or 9) for each of the four groups in our study.

To ensure that the English proficiency levels of the four groups would be roughly equivalent, we assigned the participants to the four groups so that there would be no statistically significant difference in the group mean scores on the TOEIC® Test, $F(3, 156) = 0.03$, $p = .994$, $\eta_p^2 < .01$. We excluded 10 participants because they did not have TOEIC® scores. On the basis of their TOEIC® scores ($M = 489.3$, $SD = 144.7$), we estimated most participants to fall between the A2 (elementary) and B1 (intermediate) levels in the Common European Framework of Reference for Languages benchmarks. Prior to the experiment, we administered the paired-associate section of Language Aptitude Battery for the Japanese (LABJT; Sasaki, 1993). This section of LABJT is a Japanese translation of Part V of the Modern Language Aptitude Test, and it targets test-takers' ability to learn L1 and L2 word pairs in a decontextualized format. We used the scores on this test to investigate whether short and long spacing schedules had differential effects on learners with different verbal memory capacities.

Materials

We used the same 20 low-frequency English words from Nakata and Webb's (2016) study as target items (see Appendix S2 in the Supporting Information online for a list of target items). All words were from the 10,000 word level or lower in the British National Corpus frequency lists (Nation, 2006). We chose these words because previous research suggested that they were likely to be unfamiliar to most Japanese undergraduate students (Nakata, 2015; Nakata & Webb, 2016). Out of the 20 items, 12 were nouns and eight were verbs, which approximated the 6:4 ratio of nouns to verbs in natural text (Webb, 2005). The materials and instructions that we used in this study are publicly available via the IRIS Database (Nakata et

al., 2022b).

Procedure

We conducted the experiment over three weekly sessions using Gorilla Experiment Builder (https://app.gorilla.sc; see Appendix S6 in the Supporting Information online for details). Each of the participants had access to a computer, and they studied and completed the tests individually. Figure 1 summarizes the procedure of this study.

<Place Figure 1 near here>

Session 1

At the beginning of the first session, the participants received explanations about the study (see Appendix S5 in the Supporting Information online for details) and were asked to indicate their consent by clicking an Agree button. After that, the participants practiced using the computer program with four sample words. After the practice, we administered the productive and receptive pretests in that order. In the productive pretest, Japanese (L1) translations of the target items were presented one by one (e.g., 彫る = ____?), and the participants were asked to type an English (L2) word corresponding to the translation (e.g., *gouge*). To ensure that the participants did not provide synonyms for a target word (e.g., *carve* for *gouge*), we provided the number of letters in the target word as well as one letter from the word as a hint (e.g., _ o _ _ _ for *gouge*). We used the same hints as in Nakata and Webb's (2016) study (see Appendix S2 in the Supporting Information online). In the receptive pretest, we presented the participants with target items one by one (e.g., *gouge* = ____?), and asked them to type the meaning in Japanese (e.g., 彫る). In both productive and receptive pretests, to reduce a possible order effect, we randomized the order in which the target items appeared for each participant. To familiarize the participants with the test format, we presented four sample items before the participants encountered the 20 target items in

both pretests.

After the pretests, we conducted the initial learning session. We used different spacing schedules depending on the group to which we had assigned the participants. Specifically, for the long–short and long–long groups, we used a long spacing schedule, whereas for the short–short and short–long groups, we used a short spacing schedule. Figure 2 summarizes the item order in the long and short spacing schedules. In both schedules, the first encounter with each target item, which is shown with (P) in Figure 2, was an initial presentation trial. All other trials were retrieval trials (see below for details). In the long spacing schedule (Figure 2, left), the treatment consisted of up to five cycles of 20 items, and each item was presented for practice only once in each cycle. Target words that reached the criterion of one correct retrieval were dropped from subsequent cycles. The treatment was terminated when all target items reached the criterion of one correct retrieval or the remaining target items had been presented for retrieval practice four times. We set the maximum number of retrieval trials to four per target word on the basis of results from a study conducted by Nakata and Webb (2016) as well as on computer simulations of the experiment. In the short spacing schedule (Figure 2, right column), the 20 target items were randomly divided into five blocks of four items and repeated. As in the long spacing schedule, a given target word was removed from further practice when it was correctly answered once. When all four items in a given block had reached the criterion of one correct retrieval or when remaining items in the block had been presented for retrieval practice four times, the next block of four items was practiced.

<Place Figure 2 near here>

To minimize the potential of an order effect, we randomized the item order for each participant. For instance, for one participant, *apparition* was Item 1, and *warble* was Item 20, whereas for another participant, *rue* was Item 1, and *billow* was Item 20. We also randomized

the item order for each repetition so that it would not provide inappropriate help through the participants' remembering it. For instance, in Figure 2, "1−4" does not mean that items were encountered in the fixed order Item 1, Item 2, Item 3, Item 4. In the actual experiment, items were encountered in a random order such as Item 1, Item 2, Item 4, Item 3 or Item 2, Item 1, Item 3, Item 4. We used the same randomization algorithm as Nakata and Webb (2016) used, and the same item never occurred twice in a row (unless other items had already been removed from the treatment after reaching the criterion of one correct recall had been attained). Although randomization of the item order changed spacing for individual items, it did not affect the mean spacing as a whole because possible differences from the mean spacing (i.e., a lag of three items for the short condition and 19 items for the long condition) were cancelled out across items.

As Figure 2 shows, the first encounter with each target item was an initial presentation trial where the target item and its Japanese translation were presented simultaneously for 8 s per word pair (e.g., *gouge* = 彫る). From the second encounter, the Japanese translation of one of the target items was presented for retrieval practice (e.g., 彫る = ____?), and the participants were asked to type the corresponding English word (e.g., *gouge*). We set no time limit for the retrieval trials. In each retrieval trial, after the participants had typed their response, the Japanese translation and the corresponding target English word were presented for 5 s as feedback (e.g., *gouge* = 彫る).

As we have described above, in both short and long spacing schedules, target words that had reached the criterion of one correct retrieval were dropped from practice. This enabled us to test the view that short spacing might be more efficient than long spacing because it might allow learners to reach the criterion more quickly than would long spacing (Research Question 2). We set the criterion to one correct retrieval because previous research had suggested that the benefits of higher criterion levels (e.g., three correct retrievals instead

of one) are often eliminated completely over time, and a less strict criterion may be more efficient (Rawson et al., 2018; Vaughn et al., 2016). After the initial learning session, the participants answered Questionnaire 1 (see the questionnaire and its English translation in Appendix S3 in the Supporting Information online). In the first section of this questionnaire (Section A), we asked the participants to estimate how many target words out of 20 they would expect to remember 1 week later (judgments of learning). Previous research has suggested that, although long within-session spacing often leads to superior long-term retention in comparison with short within-session spacing, learners are not necessarily aware of the benefits of long spacing (Kornell, 2009; Nakata & Suzuki, 2019). We included this question to investigate whether the findings of existing research would also be replicated in our study.

Session 2

The second session, which we conducted 1 week after the first session, consisted of a relearning session and Questionnaire 2. In the relearning session, we used the long spacing schedule (Figure 2, left column) in the short–long and long–long groups, whereas, in the short–short and long–short groups, we used the short spacing schedule (Figure 2, right column). The relearning session was exactly the same as the initial learning session except that the relearning session consisted only of retrieval trials and did not include the initial presentation trials, shown with (P) in Figure 2. After the relearning session, participants answered Questionnaire 2, which was exactly the same as Questionnaire 1 given at the end of Session 1 (see Appendix S3 in the Supporting Information online).

Session 3

We conducted the third session 1 week after the second session. At the beginning of Session 3, we administered delayed posttests without prior notice. We administered a productive posttest first, followed by a receptive posttest. Unlike in the pretest, the hints (e.g., _ o _ _ _

for *gouge*) were not given in the productive posttest. We again randomized the order in which the target items appeared for each participant. Other than these points, the delayed posttests were exactly the same as the pretests. After the delayed posttests, the participants answered Questionnaire 3 (see the full questionnaire in Appendix S4 in the Supporting Information online.

Changes From Preregistered (Stage 1) Manuscript

After our preregistered (Stage 1) manuscript had been accepted, we requested the following changes to the procedure:

- using Gorilla Experiment Builder instead of custom software for data collection,
- asking the participants to indicate informed consent by clicking an Agree button instead of signing a form,
- removing a question about participants' scores on English proficiency exams such as TOEIC® or TOEFL® from Questionnaire 3 (Appendix S4 in the Supporting Information online),
- adding a question about participants' L1 to Questionnaire 3 (Appendix S4 in the Supporting Information online), and
- revising instructions given to the participants before and during the experiment (Appendixes S5 and S6 in the Supporting Information online) on the basis of results of piloting.

All these changes were approved by the handling editor prior to our data collection.

Scoring

For the productive pretest and posttest, where we asked the participants to provide English (L2) target words corresponding to Japanese (L1) translations, we marked as correct only correctly spelled responses. We did not use a more lenient scoring method that would have

awarded scores to responses with minor misspellings because previous research examining the effects of within-session spacing on L2 vocabulary learning has suggested that both strict and lenient scoring methods produce similar results (Nakata, 2015; Nakata & Webb, 2016).

For the receptive pretest and posttest, where we asked the participants to provide Japanese (L1) translations corresponding to English (L2) target words, a native speaker of Japanese with a postgraduate degree in applied linguistics first created answer keys (a list of correct responses) based on English–Japanese dictionaries. Two other native speakers of Japanese with postgraduate degrees in applied linguistics independently verified the answer keys. Those creating answer keys ignored the part of speech and the transitive/intransitive distinction. In other words, for the target word *rue*, not only 後悔する [verb] but also 後悔 [noun] was regarded as correct. Similarly, both 後悔する [intransitive] and 後悔させる [transitive] were marked as correct for the target word *rue*. For consistency in scoring, responses on the receptive test were first categorized by custom software into the following four groups based on the answer keys: (a) correct (i.e., the response matched one of the answer keys for the target word), (b) blank responses, (c) interference errors (i.e., the response matched one of the answer keys for other target words), and (d) other responses. Two native speakers of Japanese with postgraduate degrees in applied linguistics independently scored the responses that had been categorized as "other responses." The interrater agreement was 100% (142/142) and 89.8% (97/108) for the scorings of category "other responses" on the receptive pretest and posttest, respectively. The two raters resolved disagreements by discussion.

Data Analysis

All analyses for this study were preregistered through the OSF (https://osf.io/t29ka). We excluded 54 participants from analysis because (a) they did not have TOEIC® scores, (b)

they missed one or more of the experimental sessions, (c) they indicated in Questionnaire 3 that Japanese was not their L1, or (d) they indicated in Questionnaire 3 that they had had exposure to one or more of the target words outside of the experiment. We assigned the remaining 116 participants to one of our four spacing groups, with 37, 32, 31, and 27 participants in the short–short, short–long, long–short, and long–long groups, respectively. We treated items answered correctly on the pretest as missing values for a given participant, and reported descriptive statistics (means, standard deviations, and 95% confidence intervals for the means) for all the dependent variables. We estimated the reliability of the delayed posttest scores using Cronbach alpha. Nakata and Webb (2016) administered the same productive and receptive posttests to participants of similar profiles to those in this study and reported Cronbach's alphas between .73 and .91. We analyzed the data obtained from our experiment using a series of analyses of covariance (ANCOVA). We conducted ANCOVAs rather than mixed-effects model analyses because the effect size estimation approach to sample size estimation that we chose to use applies only to ANCOVAs. We set the significance level at .05 for all analyses. We have reported partial eta squared as the effect size estimate for ANCOVAs and Cohen's *d* as the effect size estimate for pairwise comparisons along with the 95% confidence intervals for both estimates. We followed Richardson's (2011) suggestion for interpreting partial eta squared. We considered partial eta squared values of .01, .06, and .14 as small, medium, and large effects, respectively. For Cohen's *d*, we followed Plonsky and Oswald's (2014) guidelines. We considered *d* values of 0.40, 0.70, and 1.00 as small, medium, and large effects, respectively.

Prior to conducting the analyses, we tested the statistical assumptions of normality and homogeneity of regression slope for ANCOVA. We examined normality by inspecting histograms. If normality was violated, we first transformed all data using the square root, log, and inverse transformation methods. If more than one transformation method resulted in

normally distributed data according to visual inspections, we adopted the method that most improved the data distribution (on the basis of visual inspections and of skewness and kurtosis scores). If none of the transformation methods resulted in a normal distribution, we removed outliers ($z > 3.29$; Tabachnick & Fidell, 2013) and then transformed all data again using the square root, log, and inverse transformation methods. If none of the transformation methods resulted in normally distributed data after removing outliers according to visual inspections, then we proceeded to conduct and report rank ANCOVAs, followed by paired tests with corrected alpha values for multiple comparisons. We assessed homogeneity of regression slope by testing interaction terms of independent variables with covariates. If interaction terms were significant ($p < .05$), we intended to apply the Johnson-Neyman procedure (Huitema, 2011, Chapter 11). However, none of the interaction terms were significant in our dataset.

Effectiveness

To examine whether spacing schedules for the initial learning session affected retention 1 week after initial learning, we conducted a one-way ANCOVA. The dependent variable was the proportion of correct responses on the first retrieval attempt in the relearning session in Session 2. The independent variable was spacing schedules for initial learning (short vs. long). We added LABJT scores as covariates to control for possible differences in associative memory ability. We included LABJT scores as covariates in all subsequent ANCOVAs described below, and we will not mention this again for brevity. To examine combined effects of initial learning and relearning spacing on retention after relearning, we conducted two separate two-way ANCOVAs on delayed productive and receptive posttest scores. The independent variables were spacing for initial learning (short vs. long) and spacing for relearning (short vs. long). We also tested the interaction of initial spacing and relearning

spacing.

Efficiency

To examine costs associated with spacing (e.g., whether long spacing required more retrieval attempts than short spacing until all items reached the criterion of one correct retrieval), we conducted a mixed three-way ANCOVA. The dependent variable was the total number of trials required to complete the initial learning or relearning session. The independent variables were initial spacing (short vs. long), relearning spacing (short vs. long), and time (initial learning vs. relearning). Both initial spacing and relearning spacing were between-participant group variables, and time was a within-participant group variable. We examined all two-way interactions of the independent variables as well as the three-way interaction.

As another measure of efficiency, we compared an efficiency score using a two-way ANCOVA. We defined an efficiency score as the number of words learned per trial, and we calculated it by dividing the delayed posttest score by the total number of trials required to complete the initial learning and relearning sessions. The independent variables were spacing for initial learning (short vs. long) and relearning (short vs. long). We also tested the interaction of initial spacing and relearning spacing. We conducted two separate ANCOVAs for efficiency scores calculated using the productive and receptive delayed posttest scores.

Judgments of Learning

To examine the perceived effectiveness of different spacing schedules, at the end of Sessions 1 and 2, we asked the participants to estimate how many target words out of 20 they would expect to remember 1 week later (judgments of learning; see Section A of Appendix S3 in the Supporting Information online). We analyzed actual and predicted retention at the beginning of Sessions 2 and 3 to examine (a) whether the participants had varying estimations of the

effectiveness of different spacing schedules and (b) how accurate the participants' estimations were compared with their actual scores.

We analyzed actual and predicted retention at the beginning of Session 2 using a two-way mixed ANCOVA. We operationalized actual retention as the participants' performance on the first trial in the relearning session (Rawson et al., 2018). We operationalized predicted retention as the participants' judgments of learning given at the end of Session 1 (Questionnaire 1). Independent variables were initial spacing (short vs. long) and the score type (predicted vs. actual). Initial spacing was a between-participant group variable, whereas the score type was a within-participant group variable. We also explored the interaction of initial spacing and the score type.

We analyzed actual and predicted retention in Session 3 using a mixed three-way ANCOVA. We operationalized actual retention as scores on the delayed productive posttest. We operationalized predicted retention as judgments of learning given at the end of Session 2 (Questionnaire 2). The independent variables were initial spacing (short vs. long), relearning spacing (short vs. long), and the score type (predicted vs. actual). Initial spacing and relearning spacing were between-participant group variables, whereas the score type was a within-participant group variable. We tested all two-way interactions of independent variables as well as the three-way interaction.

Results

The raw data for this study are available through OSF (https://osf.io/t29ka) and IRIS (Nakata et al., 2022a).

Effectiveness

The mean proportion of correct responses on the first retrieval attempt in the relearning

session was 7.9% (95% CI [4.7, 11.1], *SD* = 12.1) for the long initial spacing schedule and 2.6% (95% CI [1.1, 4.2], *SD* = 6.2) for the short initial spacing schedule. Because none of the transformation methods resulted in normally distributed data, we conducted a rank ANCOVA. The ANCOVA showed that the difference between the long and short initial spacing schedules was statistically significant, $F(1, 115) = 12.16$, $p < .001$, $\eta_p^2 = .10$, producing a medium effect size. The results suggested that using long spacing in the initial learning session led to better retention 1 week later.

Table 1 presents the delayed posttest scores for the four spacing schedules. Cronbach alpha coefficient was .83 for both the productive and receptive tests. This was within the range of .73 to .91 reported by Nakata and Webb (2016), and we deemed it satisfactory. Using a two-way between-participant group design, we conducted a 2 (initial spacing: short vs. long) × 2 (relearning spacing: short vs. long) ANCOVA for the productive posttest scores (corrected with the square root transformation for normality). The ANCOVA showed a significant and large main effect of initial spacing, $F(1, 113) = 44.75$, $p < .001$, $\eta_p^2 = .28$. The main effect of relearning spacing was also significant with a small effect size, $F(1, 113) = 5.66$, $p = .019$, $\eta_p^2 = .05$. The interaction of initial and relearning spacing, however, was not statistically significant and the effect size was small, $F(1, 113) = 0.90$, $p = .346$, $\eta_p^2 = .01$. Post hoc comparisons with Holm correction (see Table 2) revealed that all groups significantly differed each from the other except for the comparison of the short–short and short–long groups.

<Place Table 1 near here>

<Place Table 2 near here>

Results on the receptive test were similar to those on the productive test. A two-way ANCOVA on the receptive test revealed a significant and large main effect of initial spacing, $F(1, 113) = 32.96$, $p < .001$, $\eta_p^2 = .23$. The main effect of relearning spacing was also

significant but the effect size was small, $F(1, 113) = 4.67$, $p = .033$, $\eta_p^2 = .04$. The interaction of initial and relearning spacing, however, was not statistically significant and the effect size was negligible, $F(1, 113) = 0.25$, $p = .616$, $\eta_p^2 < .01$. Post hoc comparisons with Holm correction (see Table 2) showed that the long–long group significantly outperformed, with a large effect size, the short–short group ($d = 1.46$) and the short–long group ($d = 1.16$). We found no statistically significant difference, however, between the long–long and long–short groups nor between the short–short and short–long groups, with respectively negligible ($d = 0.31$) and small ($d = 0.49$) effect sizes (the 95% CIs crossing 0).

The findings regarding effectiveness can be summarized as:

- Retention 1 week after initial learning

  long (long–long; long–short) > short (short–long; short–short)

- Retention 1 week after relearning (productive posttest)

  long–long > long–short > short–long = short–short

- Retention 1 week after relearning (receptive posttest)

  long–long = long–short > short–long = short–short

Efficiency

We operationalized the costs associated with spacing as (a) the number of trials required to complete the initial and relearning sessions (see Table 3) and (b) efficiency scores (i.e., number of words learned per trial) for the productive and receptive tests (Table 4). For the number of trials required to complete the sessions, we used a mixed three-way design and conducted a 2 (initial spacing: short vs. long) × 2 (relearning spacing: short vs. long) × 2 (time: initial learning vs. relearning) ANCOVA. The main effect of initial spacing was significant with a medium effect size, $F(1, 113) = 12.90$, $p < .001$, $\eta_p^2 = .10$. The main effect of relearning spacing was not significant and the size effect was small, $F(1, 113) = 3.02$, p

= .085, η_p^2 = .03. The interaction of initial spacing and relearning spacing was not significant and the effect size was negligible, $F(1, 113) = 0.34$, $p = .561$, $\eta_p^2 < .01$. The interaction of time and initial spacing was statistically significant with a large effect size, $F(1, 113) = 294.14$, $p < .001$, $\eta_p^2 = .72$, and so was the interaction of time and relearning spacing, $F(1, 113) = 46.24$, $p < .001$, $\eta_p^2 = .29$. The three-way interaction was not significant and the effect size was small, $F(1, 113) = 2.60$, $p = .110$, $\eta_p^2 = .02$.

<Place Table 3 near here>

<Place Table 4 near here>

Post hoc comparisons with Holm correction (see Table 5) indicated that for initial learning, the short–short and short–long groups required fewer trials than did the long–short and long–long groups. The findings indicated that during initial learning, short spacing enabled the participants to reach the criterion of one correct retrieval more quickly than did long spacing. During relearning, however, (a) the long–short group required fewer trials than did the short–short group, and (b) the long–long group required fewer trials than did the short–long group. In other words, when the relearning schedule was held constant, the participants who had studied with long spacing initially required fewer trials than did those who had studied with short spacing initially. The findings suggested that long within-session spacing in the initial learning session probably speeded up subsequent relearning, compensating for the larger initial cost. When collapsed across initial learning and relearning, however, the short–short group required fewer trials than did the long–short and long–long groups, although no significant difference existed between the short–short and short–long groups. The results suggested that, while relearning attenuated the initial costs associated with long spacing to some extent, savings in relearning were perhaps not large enough to fully override the initial costs of long spacing.

<Place Table 5 near here>

We used a two-way between-participant group design and conducted a 2 (initial spacing: short vs. long) × 2 (relearning spacing: short vs. long) ANCOVA for the efficiency scores on the productive posttest (corrected with the inverse transformation for normality). The analysis showed a significant main effect of initial spacing with a large effect size, $F(1, 113) = 23.35$, $p < .001$, $\eta_p^2 = .17$. Neither the main effect of relearning spacing, $F(1, 113) = 1.53$, $p = .219$, $\eta_p^2 = .01$, nor the interaction of initial and relearning spacing was statistically significant, $F(1, 113) = 0.91$, $p = .341$, $\eta_p^2 = .01$, and the effect sizes for both the main effect and the interaction were small. Post hoc analysis (see Table 6) suggested that the long–short and long–long groups significantly outperformed the short–short and short–long groups. The results indicated that using long spacing in the initial learning session led to higher productive efficiency scores regardless of the relearning schedule.

<Place Table 6 near here>

The results for the receptive posttest were similar to those for the productive test. Although the main effect of initial spacing was statistically significant with a medium effect size, $F(1, 113) = 8.73$, $p = .004$, $\eta_p^2 = .07$, the main effect of relearning spacing was not significant and the effect size was negligible, $F(1, 113) = 0.61$, $p = .438$, $\eta_p^2 < .01$. The interaction of initial and relearning spacing was not significant either and the effect size was negligible, $F(1, 113) < 0.01$, $p = .967$, $\eta_p^2 < .01$. The significant main effect of initial spacing suggested that using long spacing in the initial learning session led to higher efficiency scores, regardless of the relearning schedule, not only for the productive but also for the receptive posttest.

The findings regarding efficiency can be summarized as:

- Number of trials to complete initial learning

  short–long = short–short < long–long = long–short

- Number of trials to complete relearning

long–short = long–long < short–long; long–short < short–short < short–long; short–short = long–long

- Number of trials to complete initial learning and relearning

  short–short < long–short = long–long; short–long = short–short, = long–short, = long–long

- Efficiency scores (productive posttest)

  long–long = long–short > short–long = short–short

- Efficiency scores (receptive posttest)

  long (long–long; long–short) > short (short–long; short–short)

Judgments of Learning

After the initial learning session, we asked the participants to estimate how many target words out of 20 they would expect to remember 1 week later. After the initial learning session, the average predicted score was 23.9% (95% CI [17.0, 25.7], *SD* = 16.5) for the short initial spacing schedule and 21.4% (95% CI [18.7, 29.1], *SD* = 21.1) for the long initial spacing schedule. Because the scores on the initial retrieval in the relearning session were severely positively skewed (see above), we conducted a ranked ANCOVA. We used a two-way mixed design and conducted a 2 (initial spacing: short vs. long) × 2 (score type: predicted vs. actual) ANCOVA. The ANCOVA showed a significant main effect of score type with a large effect size, $F(1, 113) = 82.35$, $p < .001$, $\eta_p^2 = .42$. However, neither the main effect of initial spacing, $F(1, 113) = 0.08$, $p = .779$, $\eta_p^2 < .01$, nor the interaction of initial spacing and score type, $F(1, 113) = 0.38$, $p = .539$, $\eta_p^2 < .01$, was statistically significant, and both effect sizes were negligible. The results indicated that the participants' predicted scores (short initial: 23.9%, long initial: 21.4%) were higher than were their actual scores (short initial: 2.6%, long initial: 7.9%) for both initial spacing schedules, suggesting overestimation.

After the relearning session, the mean predicted score was 14.6% (95% CI [9.6, 19.5], *SD* = 14.4) for the short–short, 13.0% (95% CI [6.2, 19.9], *SD* = 19.0) for the short–long, 25.3% (95% CI [17.4, 33.3], *SD* = 22.1) for the long–short, and 22.2% (95% CI [15.4, 29.0], *SD* = 17.4) for the long–long group. We used a three-way mixed design and conducted a 2 (initial spacing: short vs. long) × 2 (relearning spacing: short vs. long) × 2 (score type: predicted vs. actual) ANCOVA. The analysis showed a significant main effect of score type with a small effect size, $F(1, 113) = 4.46$, $p = .037$, $\eta_p^2 = .04$. This indicated that the participants' actual scores (range: 19.6–48.2%) were higher than the predicted scores (range: 13.0–25.4%) for all four groups, suggesting underestimation after relearning, as opposed to the overestimation that we had observed after initial learning. The main effect of initial spacing was also significant with a medium effect size, $F(1, 113) = 15.37$, $p < .001$, $\eta_p^2 = .12$. This significant main effect indicated that using long spacing in the initial learning session resulted in higher predicted and actual scores (predicted: 23.8%; actual: 42.4%) than did using short spacing (predicted: 13.8%; actual: 20.6%), regardless of the relearning schedule. None of the other effects were statistically significant, producing no more than small effects ($\eta_p^2 \leq .03$): main effect of relearning spacing, $F(1, 113) = 0.08$, $p = .783$, $\eta_p^2 < .01$; Initial Spacing × Relearning Spacing interaction, $F(1, 113) < 0.01$, $p = .983$, $\eta_p^2 < .01$; Score Type × Initial Spacing interaction, $F(1, 113) = 2.89$, $p = .092$, $\eta_p^2 = .03$; Score Type × Relearning Spacing interaction, $F(1, 113) = 1.49$, $p = .225$, $\eta_p^2 = .01$; and Score Type × Initial Spacing × Relearning Spacing interaction, $F(1, 113) = 0.15$, $p = .698$, $\eta_p^2 < .01$.

Explanatory Analysis

Results of this study suggested that contrary to the relearning override effect where the benefits of long initial spacing attenuate after relearning, the positive effects of long initial spacing might persist even after relearning. To examine whether the effects of initial spacing

changed over time, we calculated Cohen's *d* effect sizes for the comparison between short and long initial spacing 1 week after the initial learning and relearning sessions. First, for the proportion of correct responses on the first retrieval attempt in the relearning session (which was conducted 1 week after initial learning), we observed a small effect size (d = 0.64, 95% CI [0.27, 1.02]) for the comparison of short and long initial spacing. Second, on the delayed productive posttest (which was conducted 1 week after relearning), we found a large effect size (d = 1.24, 95% CI [0.84, 1.64]) for short and long initial spacing. The results suggested that the benefits of long initial spacing increased after relearning.

Discussion

Within-Session Spacing and L2 Vocabulary Learning

The first research question of this study concerned the effectiveness of initial and relearning spacing schedules. Performance on the first retrieval attempt in the relearning session suggested that using long spacing in the initial learning session led to better retention 1 week later, supporting the findings reported by Nakata and Webb (2016). The results indicate that long within-session spacing during initial learning facilitates retention, regardless of whether target items are practiced a fixed number of times (four times in Nakata & Webb's, 2016, study) or to a predetermined criterion (one correct retrieval in our study). The productive and receptive posttest scores suggested the following order 1 week after relearning: long–long $\geq$ long–short > short–long = short–short. The advantage of the long–long schedule over the other three suggests that introducing long spacing for both initial learning and relearning facilitates L2 vocabulary learning, which supports Hypothesis 1. The results are consistent with the findings reported by Rawson et al. (2018, Experiment 2), as well as studies on within-session lag effects (Karpicke & Bauernschmidt, 2011; Nakata, 2015; Nakata & Suzuki, 2019; Nakata & Webb, 2016; Pashler et al., 2003; Pyc & Rawson, 2009).

Contrary to the relearning override effect, according to which the benefits of long initial spacing attenuate after relearning, our study showed that the positive effects of long initial spacing might increase after relearning. Whereas we observed only a small effect size (d = 0.64) for the comparison of short and long initial spacing at the outset of the relearning session, we found a large effect size (d = 1.24) for the same comparison on the productive posttest conducted 1 week after relearning, suggesting that the positive effects of initial spacing nearly doubled after relearning. The effect size for the main effect of relearning spacing (productive: η_p^2 = .05; receptive: η_p^2 = .04), in contrast, was much weaker on the posttests compared to that of initial spacing (productive: η_p^2 = .28; receptive: η_p^2 = .23). The results suggest that initial spacing may have larger effects on long-term retention than relearning spacing, and that long relearning spacing may not compensate for the negative effects of short initial spacing (thus, long–short > short–long). These findings are inconsistent with the view that long spacing should not be introduced until relearning sessions, and spacing should be gradually increased as learning progresses (Ellis, 1995; Hulstijn, 2001; Schmitt, 2007). One caveat, though, is that our study focused on within-session spacing rather than between-session spacing. As a result, although the findings of this study indicate the benefits of distributing practice opportunities of a given L2 word within a single session, they cannot be taken as evidence for the advantages of long between-session spacing (e.g., lag of 10 days vs. 1 day between training sessions).

The advantage of long initial spacing on the posttests (i.e., long–long and long–short > short–long and short–short) was caused potentially because the long–long and long–short groups benefitted from positive effects of both retrieval success and retrieval effort during relearning. As indicated by (a) the higher proportion of correct responses on the first retrieval attempt and (b) the fewer number of trials during relearning, the long–short and long–long schedules resulted in more correct retrievals during relearning than for the other two

schedules. Furthermore, because relearning took place 1 week after initial learning, the long–short and long–long schedules perhaps led to retrieval that was not only successful but also effortful during relearning. Since (a) successful retrievals enhance memory more than unsuccessful retrievals, which is consistent with the study phase retrieval theory (Toppino et al., 2018), and (b) retrievals that are both effortful and successful enhance learning more than those that are successful yet less effortful, which is predicted by the retrieval effort hypothesis (Pyc & Rawson, 2009), the advantage of long initial spacing might have persisted even after relearning.

During initial learning, short spacing did lead to more successful retrievals than did long spacing. However, since target items were repeated after relatively short lags in the short spacing condition (i.e., after three intervening items on average), the retrievals, although more likely to be successful, were perhaps not as effortful as those in the long spacing condition. As a result, short initial spacing probably did not lead to better retention despite the higher level of retrieval success during initial learning. These interpretations, however, are only speculative. Although our study provided data regarding the level of retrieval success during training (i.e., number of trials required to complete the treatment), we did not collect any data that could be considered as an index of retrieval effort (for discussion, see Rogers & Leow, 2020; Suzuki et al., 2020). In future research, it may be useful for researchers to include measures such as first key response latency that has sometimes been used as an index of retrieval effort (Karpicke & Bauernschmidt, 2011; Pyc & Rawson, 2009) to examine potential relationships among spacing, retrieval success, retrieval effort, and retention.

The results of this study differed from those reported by Rawson et al. (2018), where the benefits of long initial spacing attenuated or disappeared completely after relearning (indicative of the relearning override effect). The inconsistent findings may be in part due to

the number of relearning sessions. While this study involved only one relearning session, the study conducted by Rawson et al. included three (Experiment 2) or four (Experiment 1) relearning sessions. As a case in point, Rawson et al. found that the benefits of long initial spacing decreased as a function of the number of relearning sessions, and the benefits disappeared completely after four relearning sessions. Our study, therefore, might have produced a larger relearning override effect if the treatment had included more relearning sessions.

Costs and Benefits of Within-Session Spacing

The second research question concerned the costs associated with initial and relearning spacing. The analysis of the number of trials required to complete the treatment showed that (a) during initial learning, short spacing enabled the participants to reach the criterion of one correct retrieval more quickly than did long spacing, and (b) during relearning, the participants who had studied with long spacing initially required fewer trials than did those who had studied with short spacing initially. The proportion of correct responses on the first retrieval attempt in the relearning session suggests that longer initial spacing led to better long-term retention (7.9%) than did shorter initial spacing (2.6%) 1 week after initial learning (lag effect). Because the participants who had studied with long spacing initially were able to recall more target words at the outset of the relearning session, they perhaps required fewer trials to reach the criterion during relearning. In other words, efficient relearning compensated in part for the larger cost of long initial spacing. At the same time, when collapsed across initial learning and relearning, the short–short group required fewer trials than did the long–short and long–long groups, although no significant difference existed between the short–short and short–long groups. The results suggest that, while relearning attenuated the initial costs associated with long spacing to some extent, savings in relearning were perhaps not large enough to fully override the initial costs of long spacing.

This study, at the same time, showed that long spacing during initial learning led to higher efficiency scores (i.e., number of words learned per trial) on both productive and receptive posttests, regardless of the relearning schedule. The findings indicate that the benefits of using long within-session spacing for both initial learning and relearning may outweigh its costs, which supports Hypothesis 2. Contrary to Hypothesis 2, however, no significant difference existed between the long–long and long–short groups in the number of trials or efficiency scores. The results may be in part due to the finding that initial spacing has larger effects on retention than relearning spacing, mirroring the results regarding effectiveness (under the first research question).

Our study also examined the effects of initial and relearning spacing on judgments of learning. Judgments of learning collected immediately after initial learning showed no significant differences between short and long initial spacing (short: 21.4%; long: 23.9%), although long initial spacing led to better actual retention 1 week later (short: 2.6%; long: 7.9%). These findings are consistent with vocabulary studies suggesting that learners are not necessarily aware of the benefits of long within-session spacing (Kornell, 2009; Nakata & Suzuki, 2019). One potential explanation for the findings is the higher level of retrieval success during initial learning in the short spacing condition. As indicated by the smaller number of trials required to complete the initial learning session, short initial spacing led to more successful retrievals during initial learning than did long spacing. As a result, the participants in the short initial group were perhaps under the illusion that they were learning very effectively, resulting in judgments of learning that were similar to those reported by the long initial group.

Judgments of learning collected immediately after relearning, however, suggested that long initial spacing resulted in higher predicted and actual posttest scores (predicted: 23.8%, actual: 42.4%) than did short initial spacing (predicted: 13.8%, actual: 20.6%), regardless of

the relearning schedule. The findings suggest that although the participants were not aware of the benefits of long initial spacing before relearning, they were cognizant of its value after relearning. The participants recognized the value of long initial spacing after relearning possibly due to their more successful relearning performance. As indicated by (a) the higher proportion of correct responses on the first retrieval attempt and (b) the fewer number of trials during relearning, long initial spacing resulted in more correct retrievals during relearning than did short initial spacing. This might have increased the participants' confidence, resulting in higher predicted, as well as actual, scores on the posttest. Our study also showed that whereas the participants in all groups tended to overestimate their retention after initial learning, they tended to underestimate their retention after relearning. During the first retrieval attempts in the relearning session, the participants perhaps noticed a wide gap between their predicted (short initial: 23.9%; long initial: 21.4%) and actual (short initial: 2.6%; long initial: 7.9%) retention, leading to their conservative judgments of learning after relearning.

Pedagogical Implications

Because we conducted this study in a tightly controlled laboratory setting, its findings may not necessarily be applicable to classroom settings. Results of this study, however, may help provide guidelines regarding how computer-based flashcard software should be designed. Specifically, the advantage of the long–long schedule over the other three suggests that it may be useful for flashcard software to use long within-session spacing for both initial learning and relearning. Our study also showed that initial spacing may have larger effects on retention than relearning spacing has. Pedagogically, the findings suggest that in flashcard software, initial learning should involve long within-session spacing, whereas short within-session spacing may be used in relearning without affecting retention substantially. The

findings are inconsistent with the view that spacing should be gradually increased as learning proceeds (Ellis, 1995; Hulstijn, 2001; Schmitt, 2007), which may be counterintuitive for some learners and instructors. As such, it may be useful to raise awareness about the importance of introducing long within-session spacing during initial learning. Our study also showed that long within-session spacing during initial learning may increase not only actual but predicted retention. The results suggest that in flashcard learning, long initial spacing, which has often been associated with more successful performance during relearning, may have the added benefit of boosting learners' confidence. One caveat, however, is that our study focused on within-session spacing rather than between-session spacing. As a result, although the findings of this study indicate the value of distributing practice opportunities of a given L2 word within a single session, they may not necessarily suggest that longer gaps between multiple training sessions (e.g., 10 days) facilitate vocabulary learning more than do shorter gaps (e.g., 1 day).

Another pedagogical implication of our study concerns the benefits of relearning for flashcard learning. Our study showed that 1 week after initial learning, the participants retained only 2.6% (short initial spacing) to 7.9% (long initial spacing) of the target words as measured by the proportion of correct responses on their first retrieval attempt in the relearning session. Productive posttest scores, however, showed that 1 week after relearning, the participants were able to recall 19.6% (short–short) to 48.2% (long–long) of the target words. The findings underscore the value of relearning for facilitating retention (Rawson & Dunlosky, 2022). At the same time, analysis of predicted and actual scores showed that, whereas the participants tended to overestimate their retention after initial learning, they tended to underestimate their retention after relearning. The findings indicate that the participants were not necessarily aware of the benefits of relearning. Pedagogically, the results suggest that it may be useful to raise awareness among learners that relearning has the

potential to significantly increase retention of initially learned information.

Limitations and Future Directions

While the findings of this study are useful, our study also has several limitations. First, we conducted this study in a tightly controlled laboratory setting. As such, we may need to be cautious about making recommendations regarding vocabulary learning in classroom settings. Classroom research examining the effects of initial and relearning spacing would be a useful follow-up to our work. Second, the treatment in this study included only one relearning session. Considering that authentic vocabulary learning typically involves multiple relearning sessions, this design feature limits the generalizability of the findings to real-life learning environments. In future research, it may be useful to examine the effects of initial and relearning spacing with more relearning sessions. The inclusion of multiple relearning sessions would also be valuable because the benefits of long within-session spacing may decrease as a function of the number of relearning sessions (Rawson et al., 2018).

Third, the final sample size for all the groups except the short–short group was smaller (27 to 32) than the target sample size (34) derived from the effect size estimation approach (as described under data analysis) due to relatively high attrition. In future research, it would be advisable to plan a sample size based on a higher attrition rate. Fourth, studies have suggested that the effects of spacing may be affected by a number of variables such as the type of posttest or learning activity. Several studies, for instance, have suggested that, whereas long spacing may be effective for the acquisition of declarative knowledge, short spacing is sometimes beneficial for proceduralization (Li & DeKeyser, 2019; Suzuki, 2017). Furthermore, although most studies involving paired-associate learning have found benefits of spacing, several studies investigating contextual vocabulary learning failed to do so (Elgort & Warren, 2014; Serrano & Huang, 2018; Webb & Chang, 2015). As most studies

showing these interactions investigated the effects of between-session spacing, the findings of these studies may not necessarily apply to those of within-session spacing. At the same time, in future research, it may be useful to investigate whether the effects of initial and relearning within-session spacing are also affected by these potential moderators.

Conclusion

Although many studies have suggested that long within-session spacing facilitates L2 vocabulary learning, the studies have been limited in that they examined the effects of only initial spacing, and not relearning spacing. Furthermore, most vocabulary studies on within-session spacing have investigated benefits, but not costs, of spacing. Our study was the first investigation to examine both benefits and costs of initial and relearning spacing. The findings of this study are valuable because they suggest that introducing long spacing for both initial learning and relearning may increase not only effectiveness but also efficiency of vocabulary learning. Contrary to the view that spacing should be gradually increased as learning proceeds, this study also suggests that long spacing during initial learning may facilitate long-term retention and that long relearning spacing might not compensate for the negative effects of short initial spacing. Our study also suggests that long spacing during initial learning may have the added benefit of increasing learners' confidence, as indicated by the participants' superior predicted, as well as actual, retention after relearning. Considering that relearning has the potential to significantly increase retention, future research examining the effects of not only initial but also relearning spacing would be valuable. As Rawson et al. (2018) observed, examining the effects of relearning may be "an important next frontier" (p. 57) for L2 vocabulary learning.

Final revised version accepted 18 October 2022

Notes

1 The use of short within-session spacing for initial learning and long within-session spacing for relearning is similar to an expanding spacing schedule that involves gradually increasing spacing as learning progresses (e.g., 1 min, 5 min, 9 min). Using long spacing for both initial learning and relearning sessions, in contrast, corresponds to an equal spacing schedule where the lags between encounters of a given item are fixed (e.g., 5 min, 5 min, 5 min). Although a number of studies have compared the effects of expanding and equal spacing on L2 vocabulary learning, the literature comparing these two schedules has not necessarily been informative regarding what kinds of within-session spacing schedules should be used for initial learning and relearning for two reasons. First, all L2 vocabulary studies that have examined the effects of expanding within-session spacing have involved only a single training session and did not include subsequent relearning sessions (Karpicke & Bauernschmidt, 2011; Nakata, 2015; Pyc & Rawson, 2007). Second, although several studies investigated the effects of expanding and equal spacing over multiple training sessions (Kang et al., 2014; Schuetze, 2015; Schuetze & Weimer-Stuckmann, 2011), these studies manipulated between-session spacing, not within-session spacing. As a result, studies on expanding and equal spacing have not necessarily helped researchers identify how much within-session spacing should be used for initial learning and relearning to optimize vocabulary learning.

2 Kim and Webb (2022) conducted a meta-analysis on the effects of distributed practice on L2 learning. It was not possible, however, for us to determine the sample size on the basis of their study because our study was preregistered prior to the publication of their meta-analysis.

References

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

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. https://doi.org/10.1027/1618-3169.56.4.236

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, *64*(2), 365–414. https://doi.org/10.1111/lang.12052

Ellis, N. C. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, *8*(2), 103–128. https://doi.org/10.1080/0958822940080202

Huitema, B. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. John Wiley.

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal, and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge University Press.

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, *21*(6), 1544–1550. https://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. https://doi.org/10.1037/a0023436

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, *72*(1), 269–319. https://doi.org/10.1111/lang.12479

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

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), 1103–1139. https://doi.org/10.1017/S0142716419000158

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. https://doi.org/10.1111/modl.12580

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

Nakata, T., & 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. https://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. https://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

Nakata, T., Suzuki, Y., & He, X. (2022a). *Data. Dataset from "Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing?"* [Dataset]. IRIS Database, University of York, UK. https://doi.org/10.48316/vr93-0e77

Nakata, T., Suzuki, Y., & He, X. (2022b). *Test materials. Materials from "Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing?"* [Language test]. IRIS Database, University of York, UK. https://doi.org/10.48316/vk6f-rf22

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Norouzian, R. (2020). Sample size planning in quantitative L2 research: A pragmatic approach. *Studies in Second Language Acquisition*, *42*(4), 849–870. https://doi.org/10.1017/S0272263120000017

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. https://doi.org/10.1037/0278-7393.29.6.1051

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. https://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. https://doi.org/10.1016/j.jml.2009.01.004

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*, *142*(4), 1113–1129. https://doi.org/10.1037/a0030498

Rawson, K. A., & Dunlosky, J. (2022). Successive relearning: An underexplored but potent technique for obtaining and maintaining knowledge. *Current Directions in Psychological Science*, *31*(4), 362–368. https://doi.org/10.1177/09637214221100484

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

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, *38*(6), 906–911. https://doi.org/10.1093/applin/amw052

Rogers, J. (2022). Spacing effects in task repetition research. *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12526

Rogers, J., & Cheung, A. (2020). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*, *24*(5), 616–641. https://doi.org/10.1177/1362168818805251

Rogers, J., & Leow, R. P. (2020). Toward greater empirical feasibility of the theoretical framework for systematic and deliberate L2 practice: Comments on Suzuki, Nakata, & DeKeyser (2019). *The Modern Language Journal*, *104*(1), 309–312. https://doi.org/10.1111/modl.12621

Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language Learning*, *43*(3), 313–344. https://doi.org/10.1111/j.1467-1770.1993.tb00617.x

Schmitt, N. (2007). Current trends in vocabulary learning and teaching. In J. Cummins & C. Davison (Eds.), *The international handbook of English language teaching* (pp. 827–842). Springer.

Schuetze, U. (2015). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research*, *19*(1), 28–42. https://doi.org/10.1177/1362168814541726

Schuetze, U., & Weimer-Stuckmann, G. (2011). Retention in SLA lexical processing. *CALICO Journal*, *28*(2), 460–472. https://doi.org/10.11139/cj.28.2.460-472

Serrano, R., & Huang, H.-Y. (2018). Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? *TESOL Quarterly*, *52*(4), 971–994. https://doi.org/10.1002/tesq.445

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

Suzuki, Y., Nakata, T., & DeKeyser, R. (2020). Empirical feasibility of the desirable difficulty framework: Toward more systematic research on L2 practice for broader pedagogical implications. *The Modern Language Journal*, *104*(1), 313–319. https://doi.org/10.1111/modl.12625

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson Education.

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., Phelan, H.-A., & Gerbier, E. (2018). Level of initial training moderates the effects of distributing practice over multiple days with expanding, contracting, and uniform schedules: Evidence for study-phase retrieval. *Memory & Cognition*, *46*(6), 969–978. https://doi.org/10.3758/s13421-018-0815-7

Vaughn, K. E., Dunlosky, J., & Rawson, K. A. (2016). Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion? *Memory & Cognition*, *44*(6), 897–909. https://doi.org/10.3758/s13421-016-0606-y

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. https://doi.org/10.1017/S0272263105050023

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*(6), 667–686. https://doi.org/10.1177/1362168814559800

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. Meta-Analysis of L2 Vocabulary Studies on the Effects of Within-Session

Spacing.

Appendix S2. Target Items Used in the Experiment.

Appendix S3. Questionnaires 1 and 2 Given at the End of Sessions 1 and 2.

Appendix S4. Questionnaire 3 Given at the End of Session 3.

Appendix S5. Instructions Given to the Participants Prior to the Experiment.

Appendix S6. Instructions Given to the Participants During the Experiment.

Table 1 Proportion of correct responses for the delayed posttests

	Productive test			Receptive test		
Spacing group	*M*	95% CI	*SD*	*M*	95% CI	*SD*
Short–short	19.6%	[13.7, 25.4]	17.2	44.9%	[37.3, 52.4]	22.1
Short–long	21.7%	[15.9, 27.4]	16.0	52.2%	[45.1, 59.2]	19.8
Long–short	36.7%	[28.8, 44.5]	21.9	65.2%	[59.0, 71.4]	17.3
Long–long	48.2%	[42.0, 55.4]	18.5	70.4%	[62.3, 78.5]	20.6

Table 2 Post hoc comparisons for the delayed posttests

Group comparison		Productive test			Receptive test		
		p	*d*	95% CI	*p*	*d*	95% CI
Short–short	Short–long	.296	0.27	[−0.42, 0.95]	.107	0.49	[−0.19, 1.18]
	Long–short	< .001	1.06	[0.36, 1.77]	< .001	1.16	[0.45, 1.87]
	Long–long	< .001	1.68	[0.91, 2.46]	< .001	1.46	[0.70, 2.22]
Short–long	Long–short	.008	0.80	[0.09, 1.51]	.035	0.66	[−0.04, 1.37]
	Long–long	< .001	1.41	[0.64, 2.18]	.002	0.97	[0.22, 1.72]
Long–short	Long–long	.049	0.62	[−0.12, 1.35]	.258	0.31	[−0.42, 1.04]

Note. The Holm correction for multiple comparisons was applied to the *p* values.

Table 3 Descriptive statistics for the number of trials to complete initial learning and relearning sessions

	Initial learning			Relearning			Total		
Group	*M*	95% CI	*SD*	*M*	95% CI	*SD*	*M*	95% CI	*SD*
Short–short	36.8	[33.6, 39.9]	9.2	50.3	[47.9, 52.7]	7.1	87.1	[81.9, 92.2]	15.1
Short–long	34.5	[31.6, 37.5]	8.2	60.4	[58.1, 62.8]	6.6	95.0	[90.4, 99.6]	12.8
Long–short	54.2	[50.1, 58.3]	11.6	44.8	[41.8, 47.8]	8.4	99.0	[92.6, 105.4]	17.9
Long–long	52.1	[47.7, 56.5]	11.2	50.4	[47.2, 53.6]	8.1	102.5	[95.5, 109.5]	17.9

Table 4 Descriptive statistics for efficiency scores

	Productive test			Receptive test		
Group	*M*	95% CI	*SD*	*M*	95% CI	*SD*
Short–short	0.05	[0.03, 0.07]	0.05	0.11	[0.09, 0.14]	0.07
Short–long	0.05	[0.04, 0.06]	0.04	0.12	[0.10, 0.13]	0.05
Long–short	0.08	[0.06, 0.10]	0.06	0.14	[0.12, 0.16]	0.06
Long–long	0.10	[0.08, 0.12]	0.05	0.15	[0.12, 0.17]	0.06

Table 5 Post hoc comparisons for the number of trials by learning condition and by total

Group comparison		*p*	*d*	95% CI
Initial learning				
Short–short	Short–long	.351	−0.35	[−1.03, 0.34]
	Long–short	< .001	1.81	[1.06, 2.57]
	Long–long	< .001	1.56	[0.80, 2.33]
Short–long	Long–short	< .001	−2.16	[−2.95, −1.37]
	Long–long	< .001	1.91	[1.10, 2.71]
Long–short	Long–long	.355	−0.25	[−0.98, 0.48]
Relearning				
Short–short	Short–long	< .001	1.15	[0.30, 1.99]
	Long–short	.033	−0.75	[−1.58, 0.07]
	Long–long	1.000	−0.09	[−0.94, 0.76]
Short–long	Long–short	< .001	−1.90	[−2.82, −0.98]
	Long–long	< .001	−1.24	[−2.14, −0.33]
Long–short	Long–long	.122	0.66	[−0.21, 1.54]
Total				
Short–short	Short–long	.276	0.37	[−0.22, 0.97]
	Long–short	.014	0.67	[0.07, 1.27]
	Long–long	.002	0.86	[0.22, 1.50]
Short–long	Long–short	.375	0.30	[−0.31, 0.90]
	Long–long	.168	0.48	[−0.15, 1.12]
Long–short	Long–long	.430	0.19	[−0.45, 0.82]

Note. The Holm correction for multiple comparisons was applied to the *p* values.

Table 6 Post hoc comparisons for efficiency scores

Group comparison		Productive test			Receptive test		
		p	*d*	95% CI	*p*	*d*	95% CI
Short–short	Short–long	.837	0.05	[−0.63, 0.73]	1.000	0.15	[−0.53, 0.83]
	Long–short	.021	0.72	[0.03, 1.41]	.151	0.56	[−0.13, 1.24]
	Long–long	< .001	1.12	[0.38, 1.87]	.063	0.69	[−0.03, 1.42]
Short–long	Long–short	.034	0.67	[−0.04, 1.37]	.366	0.40	[−0.30, 1.10]
	Long–long	< .001	1.07	[0.32, 1.82]	.196	0.54	[−0.19, 1.27]
Long–short	Long–long	.273	0.41	[−0.33, 1.14]	1.000	0.14	[−0.59, 0.86]

Note. The Holm correction for multiple comparisons was applied to the *p* values. Although the efficiency score on the productive test was inverse transformed for analyses, in the table, the direction of effect size was aligned to that of effect sizes derived from the raw scores.

Session 1	Pretest
	Initial learning
	Questionnaire 1
Session 2	Relearning
	Questionnaire 2
Session 3	Delayed posttest
	Questionnaire 3

Figure 1 Overview of the experiment.

Long spacing	Short spacing
1−20 (P)	1−4 (P); 1−4; 1−4; 1−4
1−20	5−8 (P); 5−8; 5−8; 5−8; 5−8
1−20	9−12 (P); 9−12; 9−12; 9−12; 9−12
1−20	13−16 (P); 13−16; 13−16; 13−16; 13−16
1−20	17−20 (P); 17−20; 17−20; 17−20; 17−20

Figure 2 Item order in the long and short spacing schedules. "1−20" indicates that Items 1 to 20 were studied once. (P) indicates initial presentation trials, and all other trials were retrieval trials. Initial presentation trials were included only in the initial learning session in Session 1 and not in the relearning session in Session 2.