System 103 (2021) 102641

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

When to conduct a vocabulary quiz, before the review or after the review?

Kiwamu Kasahara *, Kohei Kanayama

Hokkaido University of Education, Japan

ARTICLE INFO

Keywords:
Intentional vocabulary learning
Distributed learning
Testing effect

ABSTRACT

The purpose of this study is to find the optimal timing for giving learners a small quiz for L2 vocabulary learning. This study had 68 university students remember 20 English and Japanese word pairs. A week later, the S-TS group ($n = 30$) took the first L1 meaning recall test before the review, whereas the S-ST group ($n = 38$) took the review before the recall test. Both groups took Posttest 1 a week later and Posttest 2 a month later. The S-TS group showed a higher retention rate in Posttest 1 and the gap became larger in Posttest 2. The present study provided original Results in that (a) it showed the positive effects of a test session before the review on long-term retention; (b) it took into consideration the effect of participants' expectations of Posttest 1 on the test results; (c) it revealed that higher expectations of a subsequent test by the S-TS group may have led to better performance in the test.

1. Introduction

The *testing effect* is a widely known robust phenomenon of improving one's memory for material. This effect is defined as a phenomenon in which having a test on learned items can lead to longer retention of these items than additional study of them (Barcroft, 2007; Carpenter et al., 2006; Roediger & Karpicke, 2006b; Rowland, 2014). The testing effect has been confirmed in various learning materials, including L2 vocabulary learning (Barcroft, 2007; Nakata et al., 2020). Taking a test can provide learners with two types of benefits: direct and indirect effects of testing (Roediger, Putnam, & Smith, 2011b). Learners can experience the direct effects of testing by their retrieval effort in a test. The act of retrieving learned items requires great mental effort, which can cause a deeper trace of the target items to be left in the learner's memory. This reinforced memory can lead to long-term retention (Craik & Tulving, 1975). The indirect effects of testing mean feedback effects provided by tests (Roediger & Karpicke, 2006a). Tests can help learners to distinguish what items they have already acquired and what items they have not. Just having a test tells test-takers which items they can answer and which items they cannot. With this information, they can have more efficient learning in subsequent learning sessions because they can spend more time and effort on the items they have not acquired. They can reduce the number of items they have to work on in learning sessions afterward (Karpicke & Grimaldi, 2012; Soderstrom & Bjork, 2014; Son & Kornell, 2008).

However, if learners fail to retrieve any item in a test, they cannot narrow down the number of items they have to work on for the same test again. They cannot take advantage of indirect effects of testing. Kanayama and Kasahara (2018) investigated this problem and found that even retrieval failure in a test can be effective for long-term retention of learned L2 lexical items. Their Results are encouraging for L2 learners, but they confirmed their results in a relatively short period of time (one week). In addition, they admitted

* Corresponding author. Hokkaido University of Education, 071-8621, 9-chome, Hokumon Cho, Asahikawa, Hokkaido, Japan.
E-mail address: kasahara.kiwamu@a.hokkyodai.ac.jp (K. Kasahara).

https://doi.org/10.1016/j.system.2021.102641
Received 24 November 2020; Received in revised form 19 August 2021; Accepted 6 September 2021
Available online 11 September 2021
0346-251X/

some limitations in their experiment procedure. In order to examine whether their promising findings can be a robust phenomenon, we conducted a replication study of Kanayama and Kasahara over a longer time span (one month) with some modification of their experimental design. We also adopted a more reliable statistical method, a Generalized Linear Mixed Model, instead of ANOVA, which was used in Kanayama and Kasahara.

2. Literature review

2.1. Distributed learning and direct effects of testing in L2 vocabulary learning

For successful L2 vocabulary acquisition, learners need to encounter target lexical items repeatedly (Nakata, 2017; Nation, 2013, van Zeeland & Schmitt, 2013; Zhang, 2014) both in incidental and intentional learning, because it is impossible to master various aspects of vocabulary knowledge in just one encounter (Nation, 2013; Schmitt, 2010). In incidental vocabulary learning, where learners obtain some vocabulary knowledge as a byproduct of listening or reading comprehension, repeated encounters can contribute to increasing vocabulary knowledge (Brown et al., 2008; Vidal, 2011; Webb, 2007). On the other hand, in intentional vocabulary learning, where learners are engaged in acquiring vocabulary knowledge with a conscious effort, mechanical repetition or practice is not effective enough to lead to substantial learning (Baddeley, 2014).

Mechanical repetition or practice in intentional paired-associate vocabulary learning was severely criticized in the heyday of Communicative Language Approach because it was not a natural way of language acquisition (Krashen, 1982, 1985). However, the concept of practice has been broadly reconceptualized from the perspectives of applied linguistics and cognitive psychology (DeKeyser, 2007; Suzuki et al., 2019a). Under desirable conditions, intentional vocabulary learning practice can play a crucial role in L2 vocabulary acquisition (Suzuki et al., 2019a, 2019b). In fact, a number of L2 vocabulary studies have shown the effectiveness of intentional paired-associate learning in terms of establishing form-and-meaning connections (Elgort, 2011; Laufer & Shmueli, 1997; Prince, 1996; Webb, 2007). Suzuki et al. (2019a) identified five crucial research areas on L2 practice, which include *distribution of practice* and *retrieval practice*. These conditions are indispensable keys to provide learners with successful intentional vocabulary learning because these two conditions can help them maintain their full attention to target items.

A great number of studies have shown that *distributed learning* is more effective for long-term retention of learned items than *massed learning* (Baddeley, 2014; Kapler et al., 2015; Kornell et al., 2009; Nakata & Suzuki, 2018; Nakata & Webb, 2016; Sobel et al., 2011). In the former learning style, learners divide their learning sessions into several occasions with a certain interval between them. In the latter learning style, learners study target items repeatedly and intensively on a single occasion over a short period of time. Students tend to believe that massed learning is more effective than distributed learning because massed learning can often yield higher scores in immediate posttests than distributed learning (Kornell et al., 2009). They seem to believe that learned items in their short-term memory can easily be transferred to their long-term memory. However, the delayed tests of target items in the aforementioned studies showed that distributed learning produced much better Results. One recent study to show the superiority of distributed learning on L2 vocabulary learning is Nakata et al. (2020). They had 72 Japanese university students learn 80 low-frequency English words over nine weekly classes. The participants were assigned to a cumulative group or a noncumulative group. Both groups studied 10 items in each class and took a vocabulary quiz in the next class. In the noncumulative group, the 10 items introduced in the previous class were tested. In the cumulative group, not only these 10 items, but items introduced in earlier classes were tested. In their delayed posttest, the cumulative test group significantly outperformed the noncumulative group, because the former obtained the benefits of distributed learning.

One possible theory for the advantage of distributed learning in long-term retention is called *deficient processing* (Carpenter, 2020). This theory argues that learners are likely to pay more attention to target items in distributed learning than in massed learning. In massed learning, learners pay less and less attention to target items and may feel bored because familiar items are repeated in a short span of time. On the other hand, distributed learning repeats target items with breaks between, which can help learners feel afresh and keep their full attention to the targets. Another theory is *consolidation* (Landauer, 1969), which means a set of neural processes that make memories stable over time. Human brains work unconsciously to maintain information obtained from an experience. Intervals in distributed learning can help learners take advantage of consolidation.

Another indispensable condition for enhancing intentional vocabulary learning is retrieval practice, or *direct effect of testing* (Arnold & McDermott, 2013; Roediger, Agarwal, et al., 2011). Introducing small vocabulary quizzes between learning sessions is effective for long-term retention because it can enhance the effect of retrieval practice. Whether a test is an L2-form recall test or an L1-meaning recall test, it asks learners for a mental effort to retrieve the target vocabulary knowledge. This mental effort can leave a deep trace of the knowledge in the mental lexicon, which can lead to long-term retention (Karpicke & Roediger, 2007; Roediger, Agarwal, et al., 2011). Multiple-choice (MC) items seem to be weak in requiring strong mental effort because they provide several possible choices. In L2 vocabulary learning, L1-L2 paired-associate learning had been widely used for novice L2 learners to establish form-and-meaning connections. This type of learning often takes the form of word cards or word lists. Kanayama and Kasahara (2015) showed that word cards were more effective for long-term retention than word lists, because the former gave learners opportunities for retrieval whereas the latter did not. Strong and Boers (2019) compared retrieval practice and trial-and-error practice in learning phrasal verbs. Retrieval groups were given target phrasal verbs and their meanings before they were asked to retrieve the particles from the memory. On the other hand, trial-and-error groups had to guess the particles before they were provided with the correct answers. The Results showed that the retrieval groups outperformed the trial-and-error groups in the immediate posttests. These studies have shown the effect of retrieval practice in L2 vocabulary learning.

In this study, we used L1 meaning-recall tests, which asked test-takers produce L1 equivalents of L2 target items, following

Kanayama and Kasahara (2018). We could have used MC L1 meaning recall tests, where test-takers select the correct L1 equivalent from three or four choices. It is true that MC items have some testing effects and that they are good to tap a partial knowledge of L1-L2 link acquired in incidental learning situations (Read, 2000). Learners can select the key answer because of the choices even if they do not have full meaning-and-form connections. However, this study focused on the effect of retrieval failure in a test. If we had used MC items, most participants could have high scores close to the full mark. This could not be desirable for our research question. Therefore, we employed L1 meaning recall tests in this study.

2.2. *Indirect effects of testing in L1-L2 paired-associate learning*

Testing can provide learners with another type of benefit: retrieval practice in a test can improve later encoding in subsequent learning sessions. This is called *indirect effects of testing* (Arnold & McDermott, 2013; Roediger & Karpicke, 2006a). Taking tests can show learners what items they have already acquired and what they have not. Therefore, they can pay more attention to unlearned items in subsequent learning sessions. They allot more time to learning these items that have not yet been acquired (Karpicke & Grimaldi, 2012; Soderstrom & Bjork, 2014; Son & Kornell, 2008). Indirect effects of testing include another advantage for learners: preventing them from becoming overconfident. Taking tests can give learners precise information on how well they have acquired target items. On the other hand, repeated restudy sessions tend to give learners the false impression that they have already acquired their targets (Roediger & Karpicke, 2006b). Tests can help learners to maintain their attention and concentration on unlearned items in subsequent learning sessions (Karpicke & Roediger, 2007). These indirect effects of testing are another reason why testing can work effectively in L1-L2 paired-associate learning.

However, the question arises of whether they can still obtain the benefits of the indirect effects of testing even if they have few or no correct answers in a test. In this case, learners cannot narrow down the number of items they should work on in subsequent learning sessions. It was unknown whether poor performance in a vocabulary quiz can lead to long-term retention of target lexical items. Kanayama and Kasahara (2018) investigated this problem. They had 52 university students remember 20 L1-L2 pairs in the first learning session. A week later, the participants took an L1 recall test and a restudy session in two different conditions. The S-TS group ($n = 23$) took the first test (Initial Test) before the restudy, whereas the S-ST group ($n = 29$) took the restudy before the Initial Test. S means a study session, whereas T represents a test session. The hyphen shows the one-week interval. After this session, both groups took the same L1 recall test an hour later (Posttest 1) and a week later (Posttest 2). As shown in Fig. 1, The S-ST group had significantly lower scores than the S-TS group in the Initial Test (2% vs. 55%). This result was not surprising because the S-TS group participants forgot almost all the items after the one-week interval. On the other hand, the S-ST participants were able to refresh their memory of the learned items in the restudy session before the Initial Test. However, the S-TS group outperformed the S-ST group in Posttest 1 (84.2% vs. 53.2%) and in Posttest 2 (55% vs. 43.5%). This study revealed that a restudy session after poor performance can enhance the long-term retention of learned lexical items.

2.3. *Poor performance in a test can boost long-term retention of learned items*

Kanayama and Kasahara (2018) speculated that the better Results of the S-TS group could be attributed to the assumption that the poor performance in the Initial Test may have prevented the participants from being overconfident. Their failure in the first test may have helped them to devote a substantial amount of attention and effort to the target items in the restudy session. On the other hand,

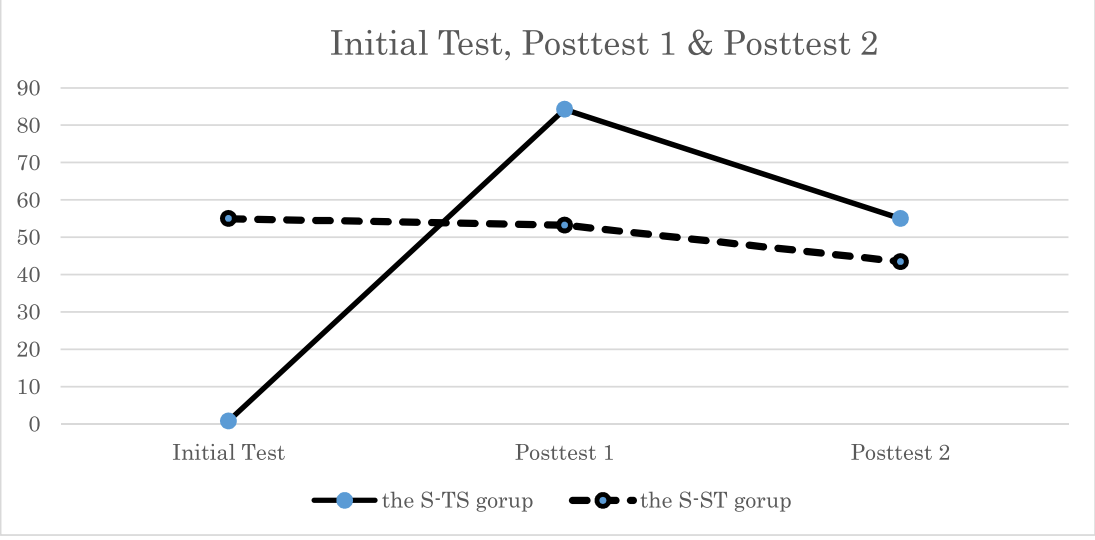

Fig. 1. The Mean Scores of Each Test (%) of the S-TS group and the S-ST group (Kanayama & Kasahara, 2018).

the participants in the S-ST group could not take advantage of the indirect effects of testing because they had the restudy session before the Initial Test. "They were not given a chance to make use of the information they had obtained from the test" (Kanayama & Kasahara, p. 9). The better learning quality of the S-TS group in the restudy session could have led to the better retention in the end.

The assumption of Kanayama and Kasahara (2018) can be supported by a *failure-encoding-effort theory* (Yang et al., 2019). This theory suggests that retrieval failure in prior tests tells learners that there is still a gap between their leaning goal and their actual learning. Learners tend to feel dissatisfied with their poor performance, and "this dissatisfaction motivates people to commit more effort to encode new information" (Yang et al., p. 811) in subsequent learning sessions. The S-TS group participants in Kanayama and Kasahara may have had the same kind of dissatisfaction with their poor performance in the Initial Test, which made the restudy session more productive and beneficial.

2.4. The purpose of the present study

A replication study of Kanayama and Kasahara (2018) should be conducted because it can shed right on the optimal timing of a vocabulary quiz. Should it be conducted before a review or after a review? Kanayama and Kasahara also admit that it is necessary to conduct a replication study of their study because they did not check the English proficiency of the participants or their prior knowledge of the target items. These two factors may have affected the Results. Moreover, their posttests were conducted relatively soon after the last treatment (1 h and one week). Considering that their participants had only two learning sessions, the advantage of the S-TS group might not have lasted over a long period. If they had given another posttest later, there would have been no great difference in score between the groups. Therefore, it is essential to examine whether the advantage of the S-TS group could continue for a long period of time; thus, we decided to conduct a posttest 1 month as well as a week after the last treatment.

Another factor that should be considered is *the test expectancy theory* (Yang et al., 2019; Weinstein et al., 2014). The gist of the theory is that prior tests can make test-takers expect subsequent tests, which motivate them to exert more effort in encoding sessions after the prior tests. In Kanayama and Kasahara (2018), all the tests were conducted without advance notice. However, the S-TS group participants may have expected a posttest because they had the restudy session after the Initial Test. Having the Initial Test without any warning and a restudy session afterward might have led to their expectancy of another test. On the other hand, the S-ST group may not have had such an expectation because they had a restudy first and a test afterward. They might have had the impression that the study-and-test cycle ended there. Asking participants whether they would expect subsequent tests can reveal if there was a gap in expectation of subsequent tests between both groups.

The research question of this study is to examine whether a vocabulary quiz should be conducted before a review or after a review. We built two hypotheses that are based on previous studies:

(H1). Having a test before a restudy is more effective for long-term retention than having a restudy before a test.

(H2). The S-TS group had higher expectancy of a subsequent test after the Initial Test than the S-ST group.

Adding the check of participants' English proficiency, their prior knowledge of target items, a posttest after a month, and their expectation of additional tests, we conducted a replication of Kanayama and Kasahara. In addition, we employed a Generalized Linear Mixed Model (GLMM) instead of the two-way ANOVA that was used in the previous study, because a GLMM can provide us with more reliable statistical data irrespective of participants' characteristics and item characteristics.

3. Method

3.1. Participants

The participants were 68 first-year university students at a national university in Japan. They were non-English major students who had studied English for at least six years. They belonged to one of two classes in a weekly-held English course. Originally, we had 70 participants and we discarded two of them from the analyses because of their prior knowledge of one target word. The authors and their teaching assistant explained the purpose and procedure of the study and obtained informed consent from the 68 students. One class was assigned to the S-TS group ($n = 30$), and the other to the S-ST group ($n = 38$). To see their English proficiency, we obtained the data of a TOEIC IP test that they took after the English course. Table 1 shows the Results of the TOEIC IP test. Judging from the mean scores and the 95% of CI, we assumed that both groups had equivalent English proficiency. A *t*-test showed that there was no significant difference in score between the two groups, $t\ (67) = -0.51$, $p = .61$, $r = .06$, though the effect size was small. After the learning session, the S-TS group took the Initial Test before the review, whereas the S-ST group took the review before the test.

The experiment was based on a 2×3 mixed factorial design. Group (S-TS, S-ST) had a manipulated between-subjects design, and

Table 1
Descriptive statistics of the TOEIC IP test for each group.

Group	*n*	Mean	*SD*	95% CI
S-TS	30	436.6	82.10	[406.31: 466.91]
S-ST	38	424.4	110.30	[388.09: 460.60]

Note. One student in the S-ST group did not take the TOEIC IP.

Test (Initial Test, Posttest 1, Posttest 2) had a manipulated within-subjects design. In short, each group took the Initial Test, Posttest 1, and Posttest 2.

3.2. Materials

This study adopted 20 low-frequency English words and their Japanese equivalent pairs that were used in Kanayama and Kasahara (2018). One of the authors conducted a pretest to examine whether the participants had any prior knowledge of the 20 target words. Given a list of the target words, they were asked to put a checkmark next to a word if they knew the word and write down the Japanese meaning of the word. They were given three to 5 min, which was enough for them to finish the pretest. The Results showed that they had little knowledge of the target words: only two of them knew one word, *ointment,* and they were discarded from the analyses. The authors judged that all the words could be used in the experiment. All the target words and their Japanese translations are shown in Table 2.

In accordance with Kanayama and Kasahara (2018), the authors made an L1 meaning recall test, which asked the participants to write down the Japanese equivalents of the target words. The same test was conducted three times as the Initial Test, Posttest 1, and Posttest 2, though the order of the items was randomized to avoid a learning-order effect. Every test was conducted without advance notice. The Initial Test was conducted one week after the learning session, but the S-TS group took it before the first review session; the S-ST group took it after the first review session. Both groups took Posttest 1 a week after the review and Posttest 2 a month after the review. Each test included a question asking the participants whether they had expected to be given the test.

3.3. Procedure

The experiment consisted of four stages carried out during part of the time allocated to four weekly English lessons. One of the authors conducted the experiment in the S-TS group class, and his teaching assistant conducted it in the same way in the S-ST group. In the first week, both groups took the pretest for three to 5 min to check whether they already knew the meanings of the 20 target words. As shown above, they had little prior knowledge of the target words. Then they had the first learning session. Both groups were asked to remember the 20 L2-L1 word pairs, which were displayed on PowerPoint slides. Whenever a word pair appeared on the screen, the first author pronounced the target English word, and had the S-TS group participants repeat it. His teaching assistant gave the same instructions on pronunciation to the S-ST group participants. Each word pair was shown on a screen in front of them. All the participants had three cycles in which they encountered each word pair. In the first round, they looked at a target word and its Japanese translation side by side at the same time for 6 s (e.g., mutineer: 反逆者). In the following two rounds, each word pair was shown to the participants in the same way as the first round, but for 4 s. The total time of learning the word pairs was the same for both groups: 4 min and 40 s, which was the same amount of time as in Kanayama and Kasahara (2018).

In the second week, the procedure was different for each group. The S-TS group took the Initial Test, which asked them to write down the Japanese meanings of the English target words in 3 min. Then they had a review session, which was identical to the learning session in the first week. On the other hand, the S-ST group worked in the opposite order: they took the review session first, and then the Initial Test later. In the third week, both groups took Posttest 1, and a month after Posttest 1, they had Posttest 2. All the tests (the Initial Test, Posttest 1, and Posttest 2) were the same L1 meaning recall tests except for the order of the items. At the bottom of the test sheet was a question asking the participants whether they had expected this recall test. They had to circle *yes* or *no* written on the sheet. Table 3 also shows the four-stage procedure of the experiment.

3.4. Scoring

This study changed a scoring system from the one Kanayama and Kasahara (2018) employed, because we would like to reexamine their Results more strictly. Kanayama and Kasahara gave two points if the student gave the same Japanese equivalent as on the PowerPoint slide, and one point for an answer that was almost the same but slightly different from the Japanese equivalent on the slide, such as an error in a part of speech (e.g., *sakkin* for *sakkin-suru*). They gave no point for a different Japanese word even if its part of speech was the same as the target. However, the cutting point between one point or two points was ambiguous. In the present study, we

Table 2
Target nouns and verbs.

English	Japanese	English	Japanese
Mutineer	*hangyakusha*	gnaw	~ *wo kajiru*
Lemur	*kitsunezaru*	smuggle	~ *wo mitsuyusuru*
Ligament	*Jintai*	sterilize	~ *wo shodokusuru*
encroachment	*shinryalu*	sham	(*byoki-nadono*) *furiwosuru*
Adhesive	*secchakuzai*	impute	*Hitonoseinisuru*
Ointment	*keshouyou-kuriimu*	belittle	~ *wo kenasu*
Deceit	*sagi*	assail	*hageshiku hinansuru*
Palliative	*kanwazai*	contort	~ *wo nejiru*
Janitor	*youmuin*	foray	*Shugekisuru*
Knack	*kotsu*	immerse	(*ekitai-nadoni*) *hitasu*

Table 3
The four stages of the experiment.

	Stage 1(Week 1)	Stage 2(Week 2)	Stage 3(Week 3)	Stage 4(Week 7)
the S-TS group	Pretest + Study Session	Initial Test + Review	Posttest 1	Posttest 2
the S-ST group	Pretest + Study Session	Review + Initial Test	Posttest 1	Posttest 2

gave one point for the same Japanese equivalent on the slide, and translations which were slightly different from the one on the slide but had the same part of speech as the target. No point was given to a different Japanese word whose part of speech was the same as the Japanese word on the slide, or a similar translation whose parts of speech was different from the target. We hoped that this strict scoring would allow us to make a strong reconfirmation of the previous study. The full score was 20 points. The first author and the second author did the scoring separately, and calculated the inter-rater reliability of the Initial test, Posttest 1 and Posttest 2. The inter-rater reliability between them were 0.97 in the Initial test, 0.98 in Posttest 1 and 0.99 in Posttest 2. If there was any discrepancy between us, it was solved through discussion. In the end we agreed on all the scores of the participants.

3.5. Data analysis

In order to examine Hypothesis 1, or whether there is a significant difference in each test between two groups, the authors used the GLMM instead of a two-way ANOVA which was employed in Kanayama and Kasahara (2018). Although means-based statistical procedures such as ANOVAs and *t* tests have been widely used in L2 studies (Lindstromberg, 2016; Plonsky, 2013), these statistical tests could overlook important individual differences between participants and target words since they depend on the mean of each group (Linck & Cunnings, 2015; Norris, 2015). Therefore, Results obtained from a certain group of participants might not be applied to other groups of participants. The GLMM could alleviate the defect by dealing with individual differences as a random effect as well as group differences as a fixed effect (Linck & Cunnings, 2015). In the analysis, we consider a dependent variable of binominal data (whether each participant answered each question or not), independent variables of Group (Group A, Group B) and of Test (Initial Test, Posttest 1, Posttest 2) as fixed effects. Their TOEIC scores are considered to be a covariate. The participants and target words (items) are regarded as random effects. The analyses were carried out using the *lme4* package (Bates et al., 2015) in R (version 3.5.1; R Core Team, 2018). The fitted model was: Correct ~ Group * Test + TOEIC + (1 | participant) + (1 | items). Another merit of the GLMM is that it does not require the normal distribution of the data obtained, or homogeneity of different groups. Lastly, the GLMM can deal with participants with missing data (Shimizu, 2014), while an ANOVA has to exclude such participants. We used odds ratio (OR) as index of effect size, following Chen et al. (2010). OR shows whether the probability of an event is the same or different between two groups. According to Chen et al., "OR = 1.68, 3.47, and 6.71 are equivalent to Cohen's $d = 0.2$ (small), 0.5 (medium), and 0.8 (large)" (p. 860).

To examine Hypothesis 2, or the participants' test expectancy, the authors calculated the percentage of the participants who answered "yes" and those who answered "no" to the question on Posttest 1. Since the outcome of this question is binominal data, the Logistic Regression Analysis was employed. This analysis considered whether participants answered either ''yes'' or ''no'' as the dependent variable, and the Group factor (Group A or Group B) as the independent variable.

4. Results

Table 4 presents the mean scores, standard deviations and 95% CI of each test for Groups A and B. The reliability of each test (Cronbach's α) was 0.93 for Initial Test, 0.80 for Posttest 1 and 0.81 for Posttest 2. We judged that these tests were reliable as vocabulary tests. Table 5 shows estimated coefficients (β), odds ratios, standard errors (*SE*), *z*-value, *p*-value, and 95% CI gained by the GLMM. Fig. 2 shows the mean scores in each test for the S-TS and S-ST groups in a line graph. The mean scores are much lower (except for the Initial Test of the S-TS group) than those of Kanayama and Kasahara (2018) because we adopted a strict scoring system and did not give partial points for any answers.

Table 5 revealed that there was a significant main negative effect of the S-TS group ($\beta = -3.18$, $SE = 0.38$, $z = -8.35$, $p < .001$), of Posttest 1 ($\beta = -1.46$, $SE = 0.15$, $z = -9.54$, $p < .001$), and of Posttest 2 ($\beta = -2.11$, $SE = 0.17$, $z = -12.53$, $p < .001$). These Results mean that the scores of the S-TS group were significantly lower than those of the S-ST Group in total. In addition, the scores of Posttest 1 by both groups were significantly lower than those of the Initial Test, and both groups retained less target items in Posttest 2 than in the Initial Test. Moreover, there was a significant main effect of TOEIC ($\beta = 0.006$, $SE = 0.002$, $z = 3.36$, $p < .001$). This means that the higher their TOEIC scores were, the better their test scores (Initial Test, Posttest 1, & Posttest 2) were. For example, a participant who got 400 points in the TOEIC would be 1.82 odds (= exp (0.006 × 100)) higher than another participant with 300 points.

Table 4
Means, SDs and 95% CI of each test for the S-TS group and the S-ST group (full = 20).

	the S-TS Group ($n = 30$)			the S-ST Group ($n = 38$)		
	Mean	*SD*	95%CI	Mean	*SD*	95%CI
Initial Test	1.23	1.33	[0.74: 1.72]	12.03	6.51	[9.76: 14.30]
Posttest 1	7.00	3.70	[5.62: 8.38]	5.09	3.76	[3.78: 6.40]
Posttest 2	5.71	3.69	[4.28: 7.15]	3.23	3.14	[2.15; 4.31]

Table 5
Fixed effect from GLMM.

Fixed Effect	*B*	*SE*	*z*	*p*	*95% CI*
Intercept	−2.62 (0.07)	0.80	−3.30	<.001	[-4.18; −1.07]
S-TS group	−3.18 (0.04)	0.38	−8.35	<.001	[-3.93; −2.44]
Posttest 1	−1.46 (0.23)	0.15	−9.54	<.001	[-1.76; −1.16]
Posttest 2	−2.11 (0.12)	0.17	−12.53	<.001	[-2.44; −1.78]
TOEIC	0.006 (1.01)	0.002	3.36	<.001	[0.002; 0.009]
S-TS × Posttest 1	3.88 (48.48)	0.26	14.73	<.001	[3.36; 4.40]
S-TS × Posttest 2	4.18 (65.20)	0.28	15.02	<.001	[3.63; 4.72]

Note. Parentheses after estimated coefficients show odds ratios.
Note. $N = 4080$; $n_{subjects} = 68$; $n_{items} = 20$.

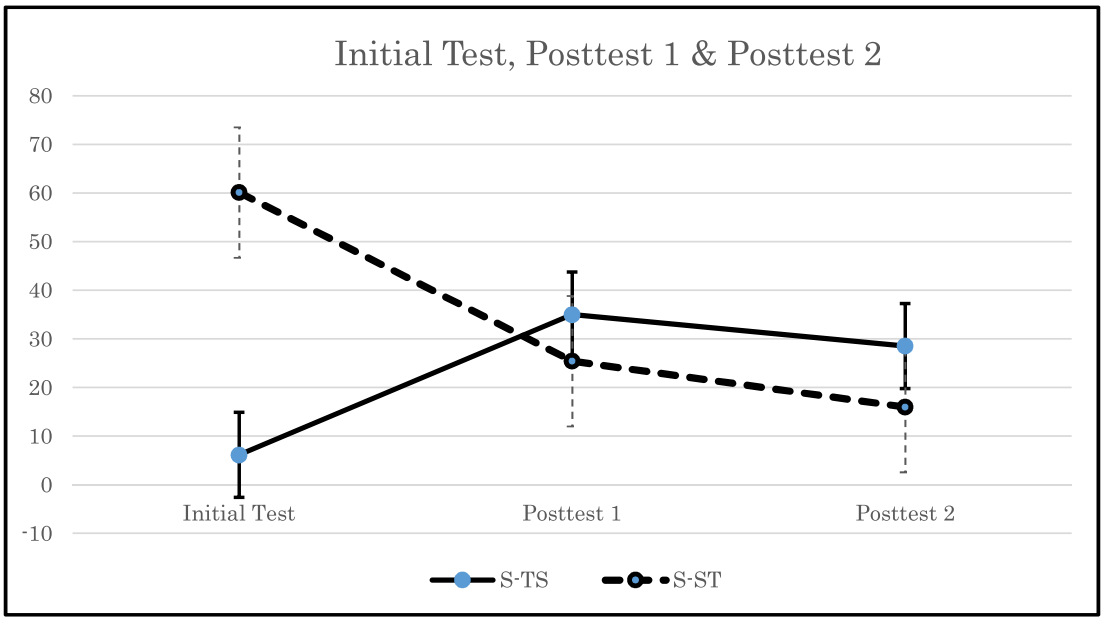

Fig. 2. The mean scores of each test in group a and group B (%).

However, the fixed effects from the GLMM also revealed that there was a significant interaction between the S-TS group and Posttest 1 ($\beta = 3.88$, $SE = 0.26$, $z = -14.73$, $p < .001$), and the S-TS Group and Posttest 2 ($\beta = 4.18$, $SE = 0.28$, $z = 15.02$, $p < .001$). This positive interaction means that the S-TS group had a higher score both in Posttest 1 and Posttest 2 than the S-ST Group. The OR of the S-TS group × Posttest 1 is 48.48, and the OR of the S-TS Group × Posttest 2 is 65.20. Both figures indicate large effect size (Chen et al., 2010) as Kanayama and Kasahara (2018) also showed the large effect size ($\eta^2 = 0.84$). Therefore, the main findings of the GLMM proved that having the Initial Test before the restudy was more effective for long-term retention than having the restudy before the Initial Test: the S-TS group showed higher retention rates than the S-ST group both in Posttest 1 (31% vs. 20%) and Posttest 2 (20% vs. 11%), which is consistent with Kanayama and Kasahara (2018).

Table 6 displays cross tabulations of the answers to the question on test expectancy in Posttest 1. It shows how many participants in each group selected *yes* or *no* to the question on whether they had expected to take Posttest 1. Table 6 also shows the total number of participants who did not answer each question and those who did not take the test.

We also compared the scores of Posttest 1 between the participants who had expected a subsequent test and those who had not. As Table 7 shows, the participants who expected a posttest outperformed those who had not in the S-TS group (35% vs. 19%), in the S-ST Group (25% vs. 15%), and in Total (32% vs. 16%).

Table 6
Cross tabulation of posttest 1.

the S-TS group ($n = 31$)			The S-ST group ($n = 39$)		
Yes	No	N.A.	Yes	No	N.A.
22	8	1	13	14	12

Note. N.A. = no answer.

Table 7
Means, SDs and 95% CI of posttest 1 in "yes" participants and "no" participants.

	N	Mean	*SD*	95%CI
S-TS				
Yes	22	7.14	3.51	[5.58; 8.69]
No	8	3.75	3.28	[1.00; 6.50]
S-ST				
Yes	13	5.08	3.69	[2.85; 7.30]
No	14	3.00	3.51	[0.97; 5.03]
Total				
Yes	35	6.37	3.66	[5.11; 7.63]
No	22	3.27	3.37	[1.78; 4.77]

Table 8 also presents the estimated coefficients (β), odds ratios, standard errors (*SE*), *z*-value, *p*-value, and 95% confidential interval (*95% CI*) gained by the Logistic Regression Analysis. There was a significant negative main effect of the S-ST group ($\beta = -1.55$, $SE = 0.57$, $z = -2.71$, $p < .01$), which means that the participants in the S-ST Group had a lower test expectancy of Posttest 1 than the S-TS Group. This indicates that most of the participants in the S-TS group took the review session with a higher expectation of taking a subsequent test.

5. Discussion

Hypothesis 1 posits that having a test before a restudy is more effective for long-term retention than having a restudy before a test. This was supported because the GLMM revealed that the S-TS group had significantly higher scores both in Posttests 1 and 2 than the S-ST group. A significant interaction was confirmed between the S-TS Group and the one-week delayed posttest (Posttest 1). Moreover, this was also the case with the S-TS group and the one-month delayed posttest (Posttest 2). These interactions mean that the S-TS group showed significantly better performance in both Posttests 1 (31% vs. 20%) and 2 (25% vs. 11%) than the S-ST group. It is true that the S-ST group got a significantly higher score in total for the three tests than the S-TS group. However, this was caused by an overwhelming score gap in the Initial Test between the S-TS Group (5%) and the S-ST Group (47%). It was natural that the S-ST Group outperformed the S-TS Group in the Initial Test because the former was given two learning sessions before the Initial Test. The total amount of study time of the S-ST group was double that of the S-TS Group. In addition, the S-ST group took the Initial Test just after the restudy session, whereas the S-TS Group took the same test one week after the first study session. The Initial Test worked as a short-term memory test for the S-ST group and a long-term memory test for the S-TS Group. Aside from the Results of the Initial Test, the present study showed that having a test before a restudy can be more effective for long-term retention than having a restudy before a test.

One remarkable finding of this study is that it showed that the advantage of the S-TS group for long-term retention could last longer (one month) than the period (one week) in Kanayama and Kasahara (2018). The order of a test before a restudy can be more effective for long-term retention than the order of a restudy before a test. The present study confirmed that this order effect is a robust phenomenon that can last a long period. This finding can be explained by the desirable difficulty framework (Bjork, 1994, 2018; Suzuki at al. 2019b). This idea conceives that optimal levels of difficulty during practice can lead to post-practice retention. Creating difficulty can slow down the rate of learning in the initial stage, but it can promote long-term retention in the end (Bjork, 2018). Some studies in L2 vocabulary learning have confirmed the phenomenon that heavier burden in learning sessions can lead to better performance in delayed recall tests (Kasahara & Kabara, 2018; Nakata et al., 2020). In this study, the S-TS group and the S-ST group took the same L1 meaning recall in the Initial test. However, the burden that the test gave to the S-TS group was more challenging than the burden to the S-ST group: the former had to recall L1 meanings a week after the learning session whereas the latter recalled the meanings just after the review session. The S-TS group participants took much lower scores than the S-ST group. However, this greater demand the S-TS group experienced in the Initial Test may have created larger retrieval effort, which could have led to better long-term retention.

Another crucial reason may be the fact that the S-TS group was given a chance to restudy the learned items after the test whereas the S-ST group was not. The retrieval failures in the Initial Test prevented the former from being overconfident about their learning outcomes. Therefore, they made the most of the restudy session: they continued paying close attention to each item until the end of the session. On the other hand, the S-ST group was not given a chance to discover their learning outcomes before the study session. They could not reflect on their learning outcomes in the restudy session. In sum, it could be said that the S-TS group had a great benefit from the indirect effects of testing, but the S-ST group did not.

Hypothesis 2 assumes that the S-TS group had higher expectancy of a subsequent test after the Initial Test than the S-ST group. This hypothesis was supported because the Logistic Regression Analysis showed that the expectancy of the S-ST group was lower than that of the S-TS group. In fact, 71% of the participants in the S-TS group expected Posttest 1 whereas only 31% of the participants in the S-ST group expected it. Expecting a later test can enhance the quality of subsequent learning sessions and test performance (Yang et al., 2019; Weinstein et al., 2014). As mentioned above, this expectancy gap was attributed to the order of the Initial Test and the restudy session. The S-TS group had the Initial Test without any warning, and the restudy afterward. They were highly likely to expect another test while having the restudy session. Their expectation of another posttest may have prevented them from being overconfident and made them concentrate on learning the target items. On the other hand, the S-ST group had the opposite order: they had the restudy before the Initial Test. They may have had the impression that their learning procedure ended with the Initial Test, which led to their

Table 8
Results from the logistic regression analysis.

	B	SE	Z	p	95% CI
Intercept	0.97 (2.64)	0.42	2.32	.02	[0.20; 1.84]
the S-ST group	−1.55 (0.21)	0.57	−2.71	<.01	[-2.72; −0.46]

Note. Parentheses after estimated coefficients show odds ratios.

low expectation of a subsequent test.

Another reason for the higher expectation of the S-TS group can be found in the failure-encoding-effort theory (Yang et al., 2019). Retrieval failures inform test-takers that they have not reached their learning goals, which makes them feel dissatisfied. "[T]his dissatisfaction motivates people to commit more effort to encode new information" (Yang et al., p. 811). The retrieval failures by the S-TS group in the Initial Test may have triggered their more serious effort in the restudy session. This serious commitment of theirs could have raised their expectation of another test.

At the end of the discussion section, we would like to mention some limitations of this study. First, the number of participants was small. It is true that we used the GLMM and obtained Results that were less affected by the characteristics of the participants and the items than an ANOVA. However, to confirm the results of this study, it is desirable to conduct a larger-scale experiment. Second, it is necessary to conduct a longitudinal study. This study included only one test and two study sessions. In classroom settings, this test-and-study cycle is usually repeated. It would be beneficial to conduct a longitudinal experiment to compare the repeated S-TS cycles and the repeated S-ST cycles. The present study conducted only one learning session and one review, which did not reflect a real leaning situation which has abundant repetitions. Finally, we would like to refer to different learning environments for the two different groups. When the experiment was conducted, each group students took three different English learning courses (including the course where the experiment was carried out), each of which was taught by a different teacher. Different lexical items or different learning strategies they learned in these different lessons might have affected the results of the experiment. However, the target words used in this experiment were so low in frequency that there was little chance for these lessons to deal with the target words.

6. Conclusion

This study replicated the Results of Kanayama and Kasahara (2018): having a test before a restudy can be more effective for long-term retention of learned lexical items than having a restudy before a test. The point is that this study showed that this effectiveness can last longer than the period found in Kanayama and Kasahara. Having a test before a restudy enables test-takers to make the most of the indirect effects of testing. Retrieval failures in the test make them exert great effort in the restudy session. Another strength of this study is that it revealed that this effectiveness is partly due to the test expectancy of learners. The order of a test and a restudy can make them expect another test, which can lead to their greater effort in the restudy.

A pedagogical implication for vocabulary instructions from this study is that a test should be done before a restudy. If you introduce several lexical items in one lesson and conduct a review of these items in the next lesson, you can give your students a vocabulary quiz before the review. Another possible way is to introduce checking in pairs. One person in a pair asks the partner about L1 meanings or L2 forms of target items. The partner answers without looking at the vocabulary list. It is crucial to let students know that experiencing retrieval failures could lead to long-term retention of target items.

Declaration of interest statement

None.

Acknowledgement

This study was supported by JSPS KAKENHI Grant Number 17K03003.

References

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguish between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 940–945. https://doi.org/10.1037/a0029199

Baddeley, A. D. (2014). *Essentials of Human memory*. Psychology Press.

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

Bates, D., Macchler, M., Bolker, B., & Walkers, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

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

Bjork, R. A. (2018). Being suspicious of the sense of ease and undeterred by the sense of difficulty: Looking back at Schmidt and Bjork (1992). *Perspectives on Psychological Science, 13*, 417–444.

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

Carpenter, S. K. (2020). *Distributed practice or spacing effect.* The Oxford Research Encyclopedias of Education. https://doi.org/10.1093/acrefore/9780190264093.013.859

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin and Review, 13*(5), 826–830.
Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation, 39*, 860–864. https://DOI:101080/03610911003650383.
Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology, 104*, 268–294.
DeKeyser, R. M. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press.
Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning, 61*, 367–413. https://doi.org/10.1111/j.1467-9922.2010.00613.x
Kanayama, K., & Kasahara, K. (2015). The effect of word retrieval on L2 vocabulary learning: Which are better tools, word lists or word cards? *Journal of the Hokkaido English Language Education Society, 15*, 21–33.
Kanayama, K., & Kasahara, K. (2018). The indirect effects of testing: Can poor performance in a vocabulary quiz lead to long-term L2 vocabulary retention? *Vocabulary Learning and Instruction, 7*(1), 1–13. https://doi: 10.7820/vli.v07.1Kanayama.
Kapler, V. I., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning. *Learning and Instruction, 36*, 38–45. https://doi.org/10.1016/j.learninstruc.2014.11.001
Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review, 24*, 401–418. https://doi.org/10.1007/s10648-012-9202-2
Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162. https://doi:10.1016/j.jml.2006.09.004.
Kasahara, K., & Kabara, M. (2018). Examining an appropriate burden between in a vocabulary quiz and an optimal interval between two quiz sessions. *Journal of Hokkaido University of Education, 69*(1), 15–25.
Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 989–998.
Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.
Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Laredo Publishing Company.
Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review, 76*(1), 82–96. https://psycnet.apa.org/doi/10.1037/h0026746.
Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal, 28*, 89–108.
Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(Suppl.,1), 185–207. https://doi:10.1111/lang.12117.
Lindstromberg, S. (2016). Inferential statistics in language teaching research: A review and ways forward. *Language Teaching Research, 20*(6), 741–768. https://doi:10.1177/1362168816649979.
Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session prepared retrieval on second language vocabulary learning. *Studies in Second Language Acquisition, 39*, 653–679. https://doi:10.1017/S0272263116000280.
Nakata, T., & Suzuki, Y. (2018). 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., Tada, S., Mclean, S., & Kim, Y. A. (2020). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly online*. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/tesq.596.
Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition, 38*(3), 523–552. https://doi.org/10.1017/S0272263115000236
Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning, 65*(Suppl.,1), 97–126. https://doi:10.1111/lang.12114.
Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition, 35*(4), 655–687. https://www.jstor.org/stable/26328389.
Prince, P. (1996). Second language vocabulary learning: The role of context versus translation as a function of proficiency. *The Modern Language Journal, 80*, 478–493.
R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.
Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011a). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*, 382–395.
Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.
Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.
Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011b). Ten benefits of testing and their applications to educational practice. In J. Mestre, & B. Ross (Eds.), *Psychology of learning and motivation, 55* pp. 1–36). Elsevier.
Rowland, C. A. (2014). The effects of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559
Schmitt, N. (2010). *Researching vocabulary*. Palgrave Macmillan.
Shimizu, H. (2014). *Kojin to shuudan no maruchireveru moderu [Multilevel modelings for individual and group data*. Nakanishiya-shuppan.
Sobel, S. H., Cepeda, J. N., & Kapler, V. I. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology, 25*, 763–767. https://doi.org/10.1002/acp.1747
Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99–115. https://doi.org/10.1016/j.jml.2014.03.003
Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1980 to the present (and beyond). In J. Dunlosky, & R. A. Bjork (Eds.), *A hand-book of memory and metamemory* (pp. 333–351). Psychology Press.
Strong, B., & Boers, F. (2019). Weighing up exercises on phrasal verbs: Retrieval versus trail-and-error practices. *The Modern Language Journal, 103*(3), 562–579. https://DOI:10.1111/modl.12579.
Suzuki, Y., Nakata, T., & Dekeyser, R. M. (2019a). Optimizing second language practice in the classroom: Perspectives from cognitive psychology. *The Modern Language Journal, 103*(3), 551–561. https://DOI:10.1111/modl.12582.
Suzuki, Y., Nakata, T., & Dekeyser, R. M. (2019b). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal, 103*(3), 713–720. https://DOI:10.1111/modl.12585.
Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258.
Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46–65. https://doi:101093/applin/aml048.
Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1039–1048. https://doi:10.1037/a0036164.
Yang, C., Chew, S.-J., Sun, B., & Shanks, D. R. (2019). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology, 111*(5), 809–826. https://doi.org/10.1037/edu0000320
van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System, 41*(3), 609–624. https://doi.org/10.1016/j.system.2013.07.012
Zhang, G. (2014). Serial position effects and forgetting curves: Implications in word memorization. *Studies in English Language Teaching, 2*(3), 306–313.