CrossMark

Studies in Second Language Acquisition, 2017, **39**, 653–679.
doi:10.1017/S0272263116000280

DOES REPEATED PRACTICE MAKE PERFECT? THE EFFECTS OF WITHIN-SESSION REPEATED RETRIEVAL ON SECOND LANGUAGE VOCABULARY LEARNING

Tatsuya Nakata
Kansai University, Japan

Although research shows that repetition increases second language vocabulary learning, only several studies have examined the long-term effects of increasing retrieval frequency in one learning session. With this in mind, the present study examined the effects of within-session repeated retrieval on vocabulary learning. The study is original in that it (a) gave posttests after a delay greater than 2 weeks, (b) employed a paired-associate format to exert strict control over the treatment, (c) considered time-on-task as a factor, and (d) used the same target items for different frequency levels. In this study, 98 Japanese learners studied 16 English-Japanese word pairs using one of the following four retrieval frequency levels: one, three, five, or seven. Results showed that five and seven retrievals contributed to significantly higher scores than one and three retrievals regardless of the posttest timing. When time-on-task was controlled, however, one retrieval led to the largest gain.

This research was supported by the Faculty Research Grant (#109605) and the Victoria PhD Scholarship from Victoria University of Wellington as well as Grant-in-Aid for Research Activity Start-up (#15H06746) and Grant-in-Aid for Young Scientists (A) (#16H05943) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. This article is based on part of the author's doctoral dissertation, which was submitted to Victoria University of Wellington in 2013. I am very grateful to Stuart Webb, Paul Nation, Jan Hulstijn, Rod Ellis, Stuart McLean, Laura Huston, and anonymous reviewers for their invaluable advice and Tomohiro Tsuchiya for his cooperation with data collection.

Correspondence concerning this article should be addressed to Tatsuya Nakata, Faculty of Foreign Language Studies, Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680 Japan. E-mail: nakata@kansai-u.ac.jp

© Cambridge University Press 2016

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Research shows that retrieval increases second language (L2) vocabulary learning (e.g., Barcroft, 2007, 2013; Folse, 2006; Karpicke & Roediger, 2008; Royer, 1973). Retrieval refers to the process of accessing information about L2 words from memory. Barcroft (2007), for instance, found that presenting L2 word forms 6 seconds after their meanings, which provides learners opportunities for retrieval, increased learning compared with presenting L2 words and their meanings simultaneously, which does not encourage retrieval. Royer (1973) also found that studying of L1 and L2 word pairs printed on different sides of word cards increases learning more than studying of L1 and L2 word pairs printed on the same side of word cards, presumably because the former encourages retrieval while the latter does not. Given that retrieval facilitates learning, one may wonder whether increasing retrieval frequency in one learning session (hereafter referred to as *within-session repeated retrieval*) increases L2 vocabulary learning. Although it seems reasonable to expect repeated retrieval to increase learning, the distributed practice effect, retrieval effort hypothesis, and desirable difficulty framework suggest that within-session repeated retrieval is not necessarily desirable (for details, see Literature Review). The present study examined the effects of within-session repeated retrieval on L2 vocabulary learning. Findings of this study will be valuable because they may enable us to determine how we can optimize L2 vocabulary learning from retrieval.

LITERATURE REVIEW

Theoretical Perspectives Regarding the Effects of Within-Session Repeated Retrieval

Although it seems reasonable to expect repeated retrieval to increase learning, the distributed practice effect (e.g., Bahrick & Phelps, 1987; Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bloom & Shuell, 1981; Cepeda et al., 2009), retrieval effort hypothesis (Pyc & Rawson, 2009), and desirable difficulty framework (e.g., Bjork, 1994) suggest that within-session repeated retrieval is not necessarily desirable. First, according to the distributed practice effect, larger spacing between retrievals results in superior long-term retention than shorter spacing or no spacing at all (pure massing). The distributed practice effect is a robust phenomenon that has been found to affect the learning of a wide range of verbal information such as L1 vocabulary, name pairs, face-name pairs, general facts, and text materials (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Cepeda et al. (2009), for instance, found that distributing retrieval opportunities of L2 vocabulary over multiple days increases retention more than practicing retrieval repeatedly in one day.

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Other L2 vocabulary studies (e.g., Bahrick & Phelps, 1987; Bahrick et al., 1993; Bloom & Shuell, 1981) also found long-term benefits of distributing practice over a longer period of time. These findings suggest that practicing retrieval multiple times over a month, for instance, would lead to better long-term retention than practicing retrieval for the same number of times in one learning session. Note that the distributed practice effect states that larger spacing enhances long-term retention more than massed practice only when other variables such as the retrieval frequency or time-on-task are held constant. Hence, the distributed practice effect does not necessarily rule out the possibility that within-session repeated retrieval enhances learning relative to fewer retrievals, at least in the short term. However, because within-session repeated retrieval inevitably involves shorter spacing than retrievals distributed over multiple days, the distributed practice effect suggests that the benefits of within-session repeated retrieval will increasingly diminish and may completely disappear in the long term.

While the distributed practice effect is concerned with the *effectiveness* of within-session repeated retrieval (i.e., the issue of whether repeated retrieval allows learners to learn a large number of words), the retrieval effort hypothesis (Pyc & Rawson, 2009) and desirable difficulty framework (e.g., Bjork, 1994) are concerned with the *efficiency* of within-session repeated retrieval (i.e., the issue of whether repeated retrieval allows learners to learn a large number of words when time-on-task is controlled). The desirable difficulty framework states that more demanding tasks result in long-term retention superior to retention achieved through less demanding tasks. The retrieval effort hypothesis, which is derived from the desirable difficulty framework, states that difficult and effortful retrievals increase learning more than easy and effortless retrievals. When learners practice retrieval repeatedly in one learning session, retrieval will be progressively easier (e.g., the second retrieval is easier than the first one, and the third retrieval is even easier than the second one). Because according to the desirable difficulty framework and retrieval effort hypothesis, easy retrievals enhance learning less than difficult retrievals, these two theories predict that the positive effects of retrievals attenuate as a function of the retrieval frequency. The second retrieval enhances learning less than the first one, and the third retrieval enhances learning even less than the second one, and so forth. This suggests that when learners practice retrieval repeatedly in one learning session, most learning occurs in the first few retrieval attempts, and practicing retrieval beyond a certain frequency level does not provide a benefit that justifies its cost (i.e., increased study time). The retrieval effort hypothesis and desirable difficulty framework, as a result, predict that within-session repeated retrieval is not necessarily efficient.

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Effects of Input Frequency on L2 Vocabulary Learning

A number of studies have examined the effects of input frequency, or the number of encounters with target words, on L2 vocabulary learning. Most of these studies found that increasing input frequency within one learning session increases vocabulary learning (e.g., Bisson, van Heuven, Conklin, & Tunney, 2014; Chen & Truscott, 2010; Webb, 2007a). A listening study conducted by van Zeeland and Schmitt (2013), however, constitutes an exception. Van Zeeland and Schmitt compared the effects of four levels of input frequency (3, 7, 11, or 15) in one learning session. Although they found rather small positive effects of repeated input on the immediate posttest, no significant difference was found on the posttest conducted 2 weeks after learning. Their results are consistent with the prediction made from the distributed practice effect. Although increasing frequency in one learning session enhances learning in the short term, its advantage diminishes over time. Another significant finding of van Zeeland and Schmitt is that although 7, 11, and 15 encounters were generally more effective than three encounters, little difference was found among the effects of 7, 11, and 15 encounters. This is consistent with the desirable difficulty framework, according to which the positive effects of repetition attenuate as a function of frequency, and increasing frequency beyond a certain frequency level does not provide further benefit that justifies its cost. A similar pattern of diminishing returns is also reported by Bisson et al. (2014). Although the findings of these studies are informative, they manipulated the frequency of input (i.e., number of encounters that do not necessarily involve retrieval), not the frequency of retrieval. As a result, these studies do not necessarily demonstrate the value of repeated *retrieval* for L2 vocabulary learning.

Effects of Retrieval Frequency on L2 Vocabulary Learning

Although smaller in number, several studies have examined the effects of retrieval frequency, not input frequency, on L2 vocabulary learning. These studies also found that repetition increases vocabulary learning. (Note that because the focus of this study is the effects of retrieval frequency, studies that manipulated the frequency of *successful* retrievals during learning [e.g., Karpicke & Roediger, 2008] will not be discussed in detail.) Laufer and Rozovski-Roitblat (2011), for instance, examined the effects of frequency over a 13-week course in two conditions: T+F and T+Fs. In the T+F condition, learners read a text with occasional Focus-on-Form (e.g., asking the instructor for the meaning of unknown words or looking up unfamiliar words in dictionaries). In the T+Fs condition, learners first encountered unknown words in a text, followed by

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Focus-on-Forms activities in which target words were practiced in a number of exercises (e.g., matching the target words and definitions or choosing the correct meanings of target words), which involved retrieval. The target words were also assigned to one of the three frequency levels: (a) two or three, (b) four or five, and (c) six or seven. Note that the frequency of retrieval was manipulated in the T+Fs condition, whereas the frequency of input was manipulated in the T+F condition. Laufer and Rozovski-Roitblat found that frequency had a larger effect on learning in the T+Fs condition than on learning in the T+F condition. However, because Laufer and Rozovski-Roitblat examined the effects of increasing frequency over a 13-week course, their findings may not necessarily be applicable to *within-session* repeated retrieval.

Several studies examined the effects of increasing retrieval frequency in one learning session. In Barcroft (2013), for instance, 75 Spanish-speaking students read an English text containing six target words, each of which occurred three times in the text. Participants were assigned to one of the four groups: III, IIR, IRI, and IRR, where *I* denotes input (i.e., viewing the target L2 word form and its L1 translation) and *R* denotes retrieval (i.e., filling a blank with the target L2 word form). In the III group, target words appeared together with their L1 translations in all three occurrences. In the IIR group, L1 translations were provided in the first and second encounters with each target word. In the third occurrence, participants practiced retrieval. Specifically, target words were replaced with a blank, and participants were asked to write down one of the target words. In the IRI group, L1 translations were given in the first and third encounters with each target word, and learners practiced retrieval in the second encounter. In the IRR group, L1 translations were given only in the first occurrence, and learners practiced retrieval in the second and third occurrences. On the posttests conducted immediately after the treatment, Barcroft found that (a) IRI significantly outperformed III and IIR, and (b) IRR significantly outperformed III. The results suggest that although one (IRI) and two retrievals (IRR) are more effective than no retrieval (III), two retrievals (IRR) do not necessarily increase learning relative to one retrieval (IRI or IIR). It should be noted, however, that because Barcroft was interested in comparing the effects of different patterns of retrieval and input opportunities, higher retrieval frequency was naturally associated with lower input frequency in his study. More specifically, IRR had more opportunities for retrieval, but fewer opportunities for input, than IRI and IIR. As a result, the results of his study may be at least partly due to the input frequency rather than the retrieval frequency per se. (Barcroft's research design, of course, should not be considered a shortcoming because the purpose of his study was not to compare different retrieval frequency levels.)

In Folse (2006), 154 learners of English studied 18 target words under three conditions: (a) one completion, (b) three completions, and

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

(c) original sentences. In the one completion condition, participants answered one fill-in-the-blank exercise for each target word, which required participants to retrieve the L2 word form once. In the three completions condition, they answered three fill-in-the-blank exercises for each target word, which involved three retrievals. In the original sentences condition, participants wrote one sentence for each target word. Learning was measured immediately after the treatment by a modified Vocabulary Knowledge Scale (Paribakht & Wesche, 1997), which required participants to indicate their knowledge of target words on a three-point scale. Depending on their response, participants were asked to write original sentences using the target words and/or produce the meaning of the target words. Folse found an advantage of the three completions condition over the one completion and original sentences conditions. His result suggests that within-session repeated retrieval (three) increases L2 vocabulary learning relative to a single retrieval. Folse also reports that the three completions condition required 50% more time than the one completion condition. As a result, it is not clear whether the three completions condition would have been more effective than the one completion condition if time-on-task had been held constant.

In Peters (2012), 56 Dutch-speaking students read an English text containing 14 target words under message-oriented and vocabulary-oriented conditions. In the former, participants answered comprehension questions about the text. The questions were created so that the participants had to write the L1 translation of target words. As a result, the message-oriented condition required participants to retrieve the meaning of each target word once. In the vocabulary-oriented condition, participants answered translation and multiple-choice vocabulary questions after reading the text, which required them to retrieve the meaning of each target word twice. Peters found that the vocabulary-oriented condition resulted in significantly higher posttest scores than the message-oriented condition immediately and 2 weeks after the treatment. The result demonstrates the advantage of repeated (two) over a single retrieval in one learning session. In Peters (2014), 35 EFL university students studied 12 single words and 12 collocations. Target items were practiced either one, three, or five times in different exercises such as fill-in-the-blanks or collocation matching. Peters found that items practiced five times were consistently remembered better than those practiced only once immediately, 1 week, and 2 weeks after learning, supporting the value of repetition in deliberate vocabulary learning.

One positive feature of the studies conducted by Peters (2012, 2014) is that she administered the posttest not only immediately but also 1 and 2 weeks after learning. As noted previously, the distributed practice effect predicts that increasing within-session retrieval frequency facilitates short-term, but not necessarily long-term retention. Peters's (2012, 2014)

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

findings, however, were inconsistent with this prediction. She found that frequency increased vocabulary learning regardless of the retention interval (i.e., interval between the treatment and posttest). One possible explanation is that the retention interval used for the delayed posttests in her research (1 and 2 weeks) was not long enough. Previous research on the distributed practice effect suggests that short spacing tends to be effective at a short retention interval, whereas long spacing tends to be effective at a long retention interval (e.g., Bird, 2010; Cepeda et al., 2006, 2009). This suggests that the benefits of within-session repeated retrieval, which involves shorter spacing than retrievals distributed over multiple days, will increasingly diminish and may completely disappear in the long term. If the posttest had been given after a delay greater than 2 weeks, therefore, benefits of repetition might not have been found. One non-L2 vocabulary study supports this prediction (Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005). In Rohrer et al.'s (2005) study, American undergraduate students studied city-country pairs (e.g., *Chiba – Japan*; Experiment 1) or L1 word pairs (e.g., *acrogen – fern*; Experiment 2). Rohrer et al. found benefits of within-session repeated retrieval 1 week after the treatment, but not after a delay of 3 weeks or greater (3, 4, and 9 weeks after learning). Their findings suggest that (a) the benefits of within-session repeated retrieval are short-lived and (b) when examining the effects of within-session repeated retrieval, it may be useful to give a posttest after a delay greater than 2 weeks.[1]

Although Peters's (2012) findings are useful, the purpose of her study was to compare the effects of message-oriented and vocabulary-oriented conditions and not to compare different retrieval frequency levels. As a result, she compared the effects of only one and two retrievals, and her research was not intended to allow us to obtain a comprehensive picture regarding the effects of within-session repeated retrieval. (This, of course, does not constitute a limitation of Peters's study because the purpose of her research was not to compare different retrieval frequency levels.) Unlike her earlier study, Peters (2014) used a wider range of retrieval frequency levels (one, three, and five). However, a different set of items was used for each retrieval frequency level in her 2014 study. For instance, in Peters (2014), the target item *turnover* was practiced once, whereas *absenteeism* was practiced five times. As a result, as Peters acknowledges, differential posttest performance for these two items may be partly attributed to item difficulty rather than different frequency levels.

PRESENT STUDY

The preceding review of literature shows that although many studies have investigated the effects of repetition on L2 vocabulary learning,

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

only two studies examined the long-term effects of increasing retrieval frequency in one learning session (Peters, 2012, 2014). With this in mind, the present study examined the long-term effects of within-session repeated retrieval. The present study is original in four respects. First, this study gave posttests after a delay greater than 2 weeks. The distributed practice effect predicts that increasing retrieval frequency in one learning session facilitates short-term, but not necessarily long-term retention. Indirect evidence for this prediction comes from L2 vocabulary research on within-session repeated input (van Zeeland & Schmitt, 2013) and non-L2 vocabulary research on within-session repeated retrieval (Rohrer et al., 2005), both of which failed to find an advantage of repetition at a retention interval of 2 weeks or greater. Considering that Peters (2012, 2014) found benefits of within-session repeated retrieval 2 weeks after the treatment, the present study gave posttests 4 weeks as well as immediately and 1 week after the treatment to investigate the long-term effects of within-session repeated retrieval.

Second, the treatment in this study was conducted in a paired-associate learning (hereafter PAL) format. In PAL, target items are presented outside message-focused tasks, and learners link the L2 word form and its meaning such as the first language (L1) translation or L2 definition (e.g., Barcroft, 2007; Elgort, 2011; Royer, 1973; Steinel, Hulstijn, & Steinel, 2007). Examples of PAL include learning from word lists and flashcards. (Although PAL does not necessarily have to involve retrieval, this study is concerned only with PAL that involves a retrieval component. This is because retrieval is found to significantly increase L2 vocabulary learning, e.g., Barcroft, 2007; Karpicke & Roediger, 2008; Royer, 1973.) Employing PAL is valuable because it allows experimenters to exert strict control over the treatment. For instance, when frequency is manipulated in reading texts, variables such as the contextual richness, importance of the target words in the comprehension of texts, or distribution of the occurrences of target words may affect learning (e.g., Elgort & Warren, 2014; Webb, 2008). Most of these variables, however, are either irrelevant or can be controlled if the treatment involves PAL. At the same time, it should be noted that there are some disadvantages to employing a PAL format. One disadvantage is that this study may have limited ecological validity. This is because although it is common for students to study vocabulary in a PAL format outside of classrooms using word lists, flashcards, or computer software (e.g., Nakata, 2011; Schmitt, 1997; Wissman, Rawson, & Pyc, 2012), the repeated use of PAL in a foreign language classroom is perhaps uncommon.

It should also be pointed out that PAL is not without criticisms. One common criticism is that PAL does not allow learners to remember L2 words successfully (e.g., Judd, 1978; Lewis, 1993; Oxford & Crookall, 1990). Empirical studies, however, suggest that PAL is one of the most

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

effective vocabulary learning strategies (e.g., Laufer & Shmueli, 1997; Nation, 2013; Webb, 2007b). Vocabulary learned in PAL is resistant to decay (Fitzpatrick, Al-Qarni, & Meara, 2008) and can be retained over several years (Bahrick & Phelps, 1987). Studies have also shown that in PAL, large numbers of words can be learned in a very short time (e.g., Fitzpatrick et al., 2008; Nation, 2013). Another common criticism is that PAL does not necessarily result in successful use of the target items (e.g., Anderson & Nagy, 1992; Larson & Smelley, 1972; Oxford & Crookall, 1990). This criticism is largely true possibly because PAL focuses on linking form and meaning, and other aspects of word knowledge such as the spoken form, associations, and constraints on use are usually not dealt with (Nation, 2013). Despite these limitations, most researchers agree that PAL should have a place in L2 vocabulary instruction (e.g., Barcroft, 2012; Folse, 2004; Laufer, 2005; Nation, 2013; Nation & Webb, 2011). Considering the role that PAL plays in vocabulary development as well as the methodological advantages that it offers, the present study examined the effects of within-session repeated retrieval in PAL.

Third, the present study considered time-on-task as a factor. Because repeated retrievals naturally take more time than fewer retrievals, the results of earlier research may be partly attributed to the difference in the study time. L2 vocabulary studies underscore the importance of considering time-on-task when comparing the effects of different treatments (Ellis, 1995; Folse, 2006; Hulstijn & Laufer, 2001; Webb, 2005). Webb (2005), for instance, compared the effects of receptive (reading) and productive (writing) tasks on L2 vocabulary learning. Although Webb found the superiority of receptive learning when time-on-task was controlled, he found the advantage of productive learning when time-on-task was not controlled. Webb's findings, along with those of other studies (e.g., Ellis, 1995; Folse, 2006; Hulstijn & Laufer, 2001), suggest that it is of value to take time-on-task into account when comparing the effects of different treatments. The present study considered the effects of time-on-task by examining the efficiency as well as the effectiveness (number of items learned) of different retrieval frequency levels. Efficiency was operationalized as the number of items learned per minute (e.g., Ellis, 1995; Kornell, 2009; Mondria, 2003) and calculated by dividing the posttest score by the study time (hereafter referred to as *efficiency scores*). Examining the efficiency of different frequency levels is useful because it allows us to test the view that within-session repeated retrieval may not necessarily be efficient according to the retrieval effort hypothesis and desirable difficulty framework (see Theoretical Perspectives Regarding the Effects of Within-Session Repeated Retrieval). Examining both effectiveness and efficiency will also help us to reach a more informed decision regarding whether within-session repeated retrievals are more desirable

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

than fewer retrievals. Fourth, in some previous studies, different target items were used for different retrieval frequency levels (e.g., Laufer & Rozovski-Roitblat, 2011; Peters, 2014). With this limitation in mind, the same target items were used for all retrieval frequency levels in this study.

In summary, the present study is significant in that it is the first study to examine the effects of within-session repeated retrieval on L2 vocabulary learning while (a) giving posttests at a retention interval greater than 2 weeks, (b) employing a PAL format to exert strict control over the treatment, (c) considering time-on-task as a factor, and (d) using the same target items for all retrieval frequency levels. Findings of the present study will be useful because they may enable us to determine the value of within-session repeated retrieval for L2 vocabulary learning and instruction.

RESEARCH QUESTIONS AND HYPOTHESES

The following two research questions were addressed:

> RQ1: In L2 vocabulary learning, does increasing within-session retrieval frequency increase posttest scores immediately, 1 week, and 4 weeks after learning?

> RQ2: In L2 vocabulary learning, does increasing within-session retrieval frequency increase efficiency scores immediately, 1 week, and 4 weeks after learning?

Based on the review of literature, the following two hypotheses were formed:

> Hypothesis 1: Within-session repeated retrieval leads to higher posttest scores than fewer retrievals immediately and 1 week after learning, but not 4 weeks.

> Hypothesis 2: Within-session repeated retrieval leads to lower efficiency scores than fewer retrievals regardless of the retention interval.

Hypothesis 1 predicts that within-session repeated retrieval is more effective than fewer retrievals immediately and 1 week after learning, but not 4 weeks after learning. This is based on the distributed practice effect and on previous studies (Peters, 2012, 2014; Rohrer et al., 2005; van Zeeland & Schmitt, 2013). Hypothesis 2 predicts that within-session repeated retrievals are less efficient than fewer retrievals because the retrieval effort hypothesis (Pyc & Rawson, 2009) and desirable difficulty framework (e.g., Bjork, 1994) state that although increasing retrieval

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

frequency in one learning session increases learning initially, its benefits may disappear after a certain frequency level.

METHOD

Participants

The original pool of participants consisted of 116 Japanese first-year engineering students (aged 15–16) from three EFL classes at a technical college in Japan. Eighteen students were excluded because they were absent on the day of the experiment, leaving 98 students. The students from the three classes were randomly assigned to one of the four retrieval frequency groups: Retrieval 1, 3, 5, and 7. Although no data were available regarding their English proficiency, an analysis of learning phase performance found no statistically significant difference among the four retrieval frequency groups in their ability to learn L2 vocabulary (see Appendix S1 in the online supplementary materials for details). Each of the Retrieval 1, 3, and 5 groups consisted of 24 students, whereas the Retrieval 7 group consisted of 26 students. The difference in the number of students resulted from the absence of students.

Target and Filler Items

Sixteen English-Japanese word pairs (e.g., *husk* – 皮) and three additional word pairs were chosen as target and filler pairs, respectively. The number of target word pairs was determined based on the results of pilot studies conducted with 53 Japanese learners of English. The target and filler word pairs were chosen from three previous experiments (Nakata, 2015a; Nakata & Webb, in press, Experiments 1 & 2), which involved 301 Japanese college students. To ensure that items would be unfamiliar to the participants, English words that were beyond the most frequent 9,000 word families in the British National Corpus frequency lists (Nation, 2006) were used (target items: *apparition*, *billow*, *citadel*, *dally*, *gouge*, *grig*, *husk*, *jibe*, *levee*, *loach*, *mane*, *mirth*, *rue*, *toupee*, *urn*, *warble*; filler items: *lava*, *tyro*, *valor*). Ten of the target items were nouns, and six were verbs, approximating the 6:4 ratio of nouns to verbs in natural text (e.g., Webb, 2005, 2009). To lower the learning burden, relatively short words were selected. Filler pairs were used as primacy and recency buffers during the treatment (see Procedure for details). Although filler pairs were studied and tested like target pairs, they were excluded from the analysis.

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Pretest

Two types of pretests, productive (i.e., form recall) and receptive (i.e., meaning recall), were administered in that order in this study. The productive pretest required learners to produce the English (L2) word form corresponding to the Japanese (L1) translation given (e.g., 皮 = ____?). So that participants would not supply synonyms for a target word (e.g., *skin* for *husk*), one letter in the target word and the number of letters in the word were provided as a hint (e.g., _ *u* _ _ for *husk*). The receptive pretest required students to provide the Japanese translation corresponding to the English word form given (e.g., *husk* = ____?). Sixteen target word pairs and three filler word pairs were tested in the pretest. The order of items in the test was randomized by the computer program for each participant.[2]

Posttest

Productive and receptive posttests were administered in that order at three retention intervals: immediately, 1 week, and 4 weeks after the treatment. Unlike in the pretest, the hints (e.g., _ *u* _ _ for *husk*) were not given in the productive posttest. Apart from this, the posttests were exactly the same as the pretests. The same 19 word pairs (16 target and three filler) were tested on the immediate and delayed posttests. The delayed posttests were conducted with no prior notice. As in the pretest, the item order was randomized by the computer program for each participant.

Procedure

The study was administered during regular class hours using computer software developed by the author. Participants studied and were tested individually. Participants first took productive and receptive pretests. After the pretests, students studied 19 English-Japanese word pairs (including three filler pairs) using a computer-based flashcard program. At the beginning of the treatment, the English-Japanese word pairs (e.g., *husk* – 皮) appeared on the participants' individual monitors for 8 seconds each (initial presentation). The three filler pairs were presented at the beginning, thus serving as primacy buffers (e.g., Nakata, 2015a). The initial presentation was followed by one or more retrieval cycles. There were one, three, five, and seven retrieval cycles for the Retrieval 1, 3, 5, and 7 groups, respectively. Each word pair was practiced

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

only once in each retrieval cycle. During the retrieval cycles, word pairs were practiced in a productive format. More specifically, participants were required to type the English (L2) word form corresponding to the Japanese (L1) translation given (e.g., 皮 = ____?). Participants were allowed to take as much time as they needed to respond. (Because the treatment was self-paced by participants, the average study time was analyzed after the experiment. See Efficiency, the Results section for details.) A productive format (from L1 to L2) was chosen because research shows that productive learning leads to adequate increases in receptive knowledge as well as large increases in productive knowledge and may be more effective than receptive learning (e.g., Mondria & Wiersma, 2004; Steinel et al., 2007; Webb, 2005, 2009).

Each retrieval attempt was followed by feedback, where the target English-Japanese word pair (e.g., *husk* – 皮) was displayed for 5 seconds per response. For each participant, for half of the target pairs, feedback was given immediately after each response (immediate feedback). For the other half, feedback was withheld until all other word pairs were practiced (delayed feedback). The timing of feedback was manipulated because the current study was part of a larger study that looked at feedback timing as a variable (Nakata, 2015b). The treatment concluded with the final review, where the English-Japanese word pairs appeared on the screen for 5 seconds each. The three filler pairs were presented at the end, thus serving as recency buffers (e.g., Nakata, 2015a). The treatment was followed by a distractor task, which required participants to complete 10 two-digit additions (e.g., 27 + 66 = ?). The immediate posttest followed the distractor task. As a measure of long-term retention, participants took productive and receptive posttests 1 and 4 weeks after the treatment.

Scoring

Two scoring protocols, strict and sensitive, were employed in this study. In the productive test, strict scoring did not allow any misspellings. Sensitive scoring was a replication of the lexical production scoring protocol (LPSP-written) developed by Barcroft (2002). In sensitive scoring, scores of 0.00, 0.25, 0.50, 0.75, or 1.00 were given based on the number of letters that are *correct* (i.e., a given letter in the target word is placed in the same position in the response) and *present* (i.e., a letter in the target word is found in the response regardless of the position; for further details, see Barcroft, 2002, 2007). The responses on the productive test were scored automatically by a computer program developed by the author. In the receptive test, with strict scoring, responses were marked as incorrect if (a) an intransitive verb was provided for

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

a transitive verb (e.g., 調和させる [transitive] for *jibe*) and vice versa or (b) responses were the wrong part of speech (e.g., 後悔 [noun] for *rue*). With sensitive scoring, the part of speech and transitive/intransitive distinction were ignored. To maintain consistency, responses on the receptive test were first scored by a computer program based on answer keys compiled by the author, a native speaker of Japanese. The answer keys were verified by two other native speakers of Japanese. Based on the answer keys, the responses on the receptive test were automatically categorized by the computer software into the following five categories: (a) correct (i.e., correct in both strict and sensitive scoring), (b) partially correct (i.e., correct only in the sensitive scoring system), (c) blank responses, (d) intralist errors (i.e., providing correct or partially correct Japanese translations for other target or filler items), and (e) others. The responses that were classified as (e) others (0.5% of all responses) were manually checked by the researcher and another native speaker of Japanese.

Analysis

The posttest scores and efficiency scores were analyzed by a three-way 4 (retrieval frequency: 1/3/5/7) × 3 (retention interval: immediate/1-week delayed/4-week delayed) × 2 (feedback timing: immediate/delayed) ANOVA. The alpha level was set to .05 for all analyses. When a significant effect was found, the Bonferroni method of multiple comparisons was used to investigate where the significant differences existed. Effect sizes and 95% confidence intervals of difference were also estimated. Because some target pairs were answered successfully on the receptive pretest, relative gains were analyzed for the receptive test. Relative gains were calculated based on the following formula (e.g., Rogers, Webb, & Nakata, 2015):

$$\text{Relative gain (\%)} = (\text{posttest score} - \text{pretest score})/(16 - \text{pretest score}) \times 100$$

Note that in the above formula, 16 indicates the number of target items.

RESULTS

Pretest

None of the participants answered correctly on any of the target words on the productive pretest. The average receptive pretest score (*SD*s in

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

parentheses) was 0.13 (0.34), 0.13 (0.34), 0.08 (0.28), and 0.12 (0.33) out of 16 with strict scoring and 0.13 (0.34), 0.17 (0.38), 0.08 (0.28), and 0.15 (0.37) out of 16 with sensitive scoring in the Retrieval 1, 3, 5, and 7 groups, respectively.

Posttest Performance

Table 1 provides the immediate and delayed posttest results for the four retrieval frequency groups. Cronbach's alpha was .84 or higher (.84–.92) for all dependent measures, which indicates good reliability. The productive and receptive test scores were analyzed by a three-way 4 (retrieval frequency: 1/3/5/7) × 3 (retention interval: immediate/1-week delayed/4-week delayed) × 2 (feedback timing: immediate/delayed) ANOVA. Table 2 provides the results of the ANOVAs. As shown by the table, the ANOVAs found significant main effects of retrieval frequency on both productive and receptive tests regardless of the scoring procedure. None of the interactions involving retrieval frequency was significant. Due to the significant main effect of retrieval frequency, the Bonferroni method of multiple comparisons was used to investigate

Table 1. Proportion of correct responses on the productive posttests and relative gains on the receptive posttests

Posttests	Retrieval frequency	Immediate		1 week		4 weeks	
		Strict	Sensitive	Strict	Sensitive	Strict	Sensitive
Productive	1 (n = 24)	59.1%	70.8%	27.6%	38.3%	24.2%	35.7%
		25.5%	*25.6%*	*24.0%*	*26.6%*	*23.1%*	*28.0%*
	3 (n = 24)	67.7%	77.3%	29.2%	38.0%	25.3%	37.0%
		26.6%	*24.0%*	*23.6%*	*25.3%*	*20.6%*	*20.8%*
	5 (n = 24)	96.9%	98.4%	49.0%	66.7%	47.9%	65.9%
		5.8%	*4.6%*	*21.4%*	*23.9%*	*27.6%*	*26.0%*
	7 (n = 26)	95.2%	97.6%	59.1%	69.2%	48.8%	66.3%
		10.5%	*6.9%*	*24.5%*	*24.2%*	*25.7%*	*27.9%*
Receptive	1 (n = 24)	71.1%	73.7%	62.8%	64.4%	59.6%	61.5%
		22.2%	*22.1%*	*25.3%*	*25.7%*	*22.2%*	*23.0%*
	3 (n = 24)	70.7%	73.3%	62.9%	63.3%	59.5%	60.2%
		24.6%	*25.6%*	*28.1%*	*28.1%*	*25.2%*	*25.6%*
	5 (n = 24)	89.5%	93.7%	84.2%	88.2%	85.0%	86.9%
		10.0%	*8.3%*	*12.1%*	*11.3%*	*14.3%*	*13.4%*
	7 (n = 26)	92.0%	95.4%	88.8%	90.8%	83.3%	85.4%
		11.0%	*6.6%*	*12.6%*	*11.9%*	*21.5%*	*22.1%*

Note. Standard deviations in italics. Strict = strict scoring; Sensitive = sensitive scoring.

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Table 2. Results of three-way ANOVAs for the posttest scores

		Strict scoring				Sensitive scoring			
Posttests	Effects	*df*	*F*	*p*	η_p^2	*df*	*F*	*p*	η_p^2
Productive	Retrieval frequency	3, 94	17.57	<.001	.36	3, 94	17.36	<.001	.36
	Retrieval frequency × Retention interval	6, 94	1.80	.100	.05	6, 94	0.36	.891	.01
	Retrieval frequency × Feedback timing	3, 94	0.78	.511	.02	3, 94	2.00	.119	.06
	Retrieval frequency × Feedback timing × Retention interval	6, 94	0.64	.702	.02	6, 94	0.93	.472	.03
Receptive	Retrieval frequency	3, 94	13.38	<.001	.30	3, 94	15.12	<.001	.33
	Retrieval frequency × Retention interval	6, 94	0.75	.594	.02	6, 94	0.56	.739	.02
	Retrieval frequency × Feedback timing	3, 94	1.25	.295	.04	3, 94	0.88	.454	.03
	Retrieval frequency × Feedback timing × Retention interval	6, 94	0.49	.815	.01	6, 94	1.23	.293	.04

where the significant differences existed at each retention interval. The results of the multiple comparisons are summarized in Tables S2 and S3 in the online supplementary materials (Appendix S2). The multiple comparisons showed that regardless of the posttest, scoring system, and retention interval, the Retrieval 5 and 7 groups significantly outperformed the Retrieval 1 and 3 groups, producing medium to large effect sizes ($0.88 \leq d \leq 2.04$). However, no statistically significant difference existed between the Retrieval 1 and 3 groups, on the one hand, and between the Retrieval 5 and 7 groups, on the other, yielding no more than small effect sizes ($d \leq 0.44$). The findings suggest the following order on all posttests: Retrieval 5 = 7 > 1 = 3.

Efficiency

Participants spent 6.48 (1.12), 12.03 (1.73), 17.98 (1.59), and 22.82 (3.31) minutes (*SD*s in parentheses) on average studying the target word pairs in the Retrieval 1, 3, 5, and 7 groups, respectively. To examine the efficiency of the four retrieval frequency levels, efficiency scores were calculated by dividing the posttest score by the study time (minutes; e.g., Ellis, 1995; Kornell, 2009; Mondria, 2003). Table 3 provides the efficiency scores in the four groups. The efficiency scores were entered into a three-way 4 (retrieval frequency: 1/3/5/7) × 3 (retention interval: immediate/1-week delayed/4-week delayed) × 2 (feedback timing: immediate/delayed) ANOVA. Table 4 shows the results of the ANOVAs. The ANOVAs showed significant main effects of retrieval frequency on both productive and receptive tests regardless of the scoring system. Although none of the interactions involving feedback timing were statistically significant, the interaction between the retrieval frequency and retention interval was significant on both productive and receptive tests regardless of the scoring procedure. Due to the significant interaction, the simple main effect of retrieval frequency was tested. The simple main effect of retrieval frequency fell short of significance with strict scoring on the 4-week delayed productive posttest, $F(94, 3) = 2.52$, $p = .062$, but was significant on all other posttests, $F(94, 3) \geq 2.92$, $p \leq .038$.

To follow-up the significant simple main effect, the Bonferroni method of multiple comparisons was used. The results of the multiple comparisons are summarized in Tables S4 and S5 in the online supplementary materials (Appendix S3). The multiple comparisons showed the following three findings. First, the Retrieval 1 group was significantly more efficient than the other three groups on all posttests except with strict scoring on the 1-week and 4-week delayed productive posttests, yielding medium to large effect sizes ($0.71 \leq d \leq 3.15$). Second, the Retrieval 3 group was more efficient than the Retrieval 7 group with sensitive

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Table 3. Average efficiency scores in the four retrieval frequency groups

Posttests	Retrieval frequency	Immediate		1 week		4 weeks	
		Strict	Sensitive	Strict	Sensitive	Strict	Sensitive
Productive	1 (n = 24)	1.45	1.74	0.67	0.97	0.57	0.86
		0.58	*0.57*	*0.59*	*0.59*	*0.50*	*0.55*
	3 (n = 24)	0.90	1.01	0.39	0.54	0.34	0.50
		0.33	*0.30*	*0.33*	*0.33*	*0.28*	*0.27*
	5 (n = 24)	0.87	0.88	0.43	0.56	0.43	0.56
		0.10	*0.09*	*0.18*	*0.18*	*0.25*	*0.22*
	7 (n = 26)	0.68	0.70	0.43	0.50	0.35	0.47
		0.13	*0.11*	*0.21*	*0.20*	*0.21*	*0.20*
Receptive	1 (n = 24)	1.74	1.81	1.54	1.58	1.46	1.50
		0.50	*0.50*	*0.60*	*0.61*	*0.50*	*0.52*
	3 (n = 24)	0.93	0.96	0.84	0.85	0.80	0.80
		0.32	*0.33*	*0.38*	*0.38*	*0.34*	*0.34*
	5 (n = 24)	0.80	0.84	0.75	0.78	0.75	0.77
		0.10	*0.10*	*0.10*	*0.10*	*0.12*	*0.11*
	7 (n = 26)	0.66	0.68	0.63	0.65	0.59	0.61
		0.14	*0.12*	*0.14*	*0.14*	*0.19*	*0.20*

Note. Standard deviations in italics. Strict = strict scoring; Sensitive = sensitive scoring.

scoring on the immediate productive posttest and with strict and sensitive scoring on the immediate receptive posttest, producing large effect sizes ($1.16 \leq d \leq 1.39$). Third, the differences were not statistically significant in all other cases. Overall, the results indicate that (a) the Retrieval 1 group was the most efficient among the four groups, and (b) the Retrieval 3 group was more efficient than the Retrieval 7 group on the immediate posttests except with strict scoring on the productive test.

DISCUSSION

The first research question in this study was concerned with the effectiveness of the four retrieval frequency levels. Based on the distributed practice effect and previous research (e.g., Folse, 2006; Peters, 2012, 2014; Rohrer et al., 2005; van Zeeland & Schmitt, 2013), Hypothesis 1 predicted that increasing retrieval frequency in one learning session increases effectiveness immediately and 1 week after learning, but not 4 weeks after learning. The findings of this study, however, did not support this hypothesis: The Retrieval 5 and 7 groups significantly outperformed the Retrieval 1 and 3 groups not only immediately and 1 week after, but also 4 weeks after the treatment. As shown in Tables S2 and S3

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Table 4. Results of three-way ANOVAs for the efficiency scores

		Strict scoring				Sensitive scoring			
Posttests	Effects	*df*	*F*	*p*	η_p^2	*df*	*F*	*p*	η_p^2
Productive	Retrieval frequency	3, 94	9.90	<.001	.24	3, 94	25.57	<.001	.45
	Retrieval frequency × Retention interval	6, 94	7.90	<.001	.20	6, 94	9.64	<.001	.24
	Retrieval frequency × Feedback timing	3, 94	1.16	.329	.04	3, 94	1.02	.388	.03
	Retrieval frequency × Feedback timing × Retention interval	6, 94	0.36	.906	.01	6, 94	0.75	.610	.02
Receptive	Retrieval frequency	3, 94	49.14	<.001	.61	3, 94	51.85	<.001	.62
	Retrieval frequency × Retention interval	6, 94	2.84	.011	.08	6, 94	2.89	.010	.08
	Retrieval frequency × Feedback timing	3, 94	0.49	.688	.02	3, 94	0.17	.916	.01
	Retrieval frequency × Feedback timing × Retention interval	6, 94	1.16	.332	.04	6, 94	1.51	.178	.05

in the online supplementary materials, larger effect sizes were observed on the immediate posttest ($0.99 \leq d \leq 2.04$) than on the 1-week ($0.88 \leq d \leq 1.34$) or 4-week delayed posttest ($0.93 \leq d \leq 1.37$). This suggests that the advantage of within-session repeated retrieval perhaps diminished over time. At the same time, the effect sizes observed on the delayed posttests still reached the threshold of a medium ($d = 0.70$) or large effect ($d = 1.00$; Plonsky & Oswald, 2014). The findings of this study are consistent with previous L2 vocabulary studies that found long-term (Peters, 2012, 2014) as well as short-term benefits of within-session repeated retrieval (Folse, 2006).

Why was the advantage of repeated retrieval also found 4 weeks after learning, contrary to the prediction of Hypothesis 1? One possible explanation is that the negative effects of massing practice were outweighed by the positive effects of increased frequency. The findings add to the existing literature suggesting the importance of repetition in L2 vocabulary learning (e.g., Bisson et al., 2014; Chen & Truscott, 2010; Elgort & Warren, 2014; Folse, 2006; Peters, 2012, 2014; Webb, 2007a). Another possible explanation is that the retention interval for the delayed posttest was not long enough. Research on the distributed practice effect shows that although massed practice facilitates learning in the short term, it leads to poor retention over time (e.g., Bird, 2010; Cepeda et al., 2006, 2009). This suggests that the benefits of within-session repeated retrieval, which involves shorter spacing than retrievals distributed over multiple days, will increasingly diminish and may completely disappear in the long term. If the posttest had been given after a delay greater than 4 weeks (e.g., 9 weeks; Rohrer et al., 2005), therefore, we might not have observed benefits of within-session repeated retrieval. In future research, it may be useful to identify the retention interval at which the positive effects of within-session repeated retrieval disappear. Another possible explanation is that repetition affects retrieval-based and non-retrieval-based learning conditions differently. Although van Zeeland and Schmitt (2013) failed to find an advantage of repeated input 2 weeks after learning, the present study found that the benefits of within-session repeated retrieval persist at least 4 weeks after the treatment. Because retrieval increases learning more than input (e.g., Barcroft, 2007; Karpicke & Roediger, 2008; Royer, 1973), the positive effects of within-session repeated retrieval may be more durable than those of within-session repeated input.

Although the Retrieval 5 and 7 groups significantly outperformed the Retrieval 1 and 3 groups, no statistically significant difference existed between the Retrieval 1 and 3 groups, on the one hand, and between the Retrieval 5 and 7 groups, on the other. The results suggest that although within-session repeated retrieval generally facilitates vocabulary learning, its benefits may not necessarily be observed depending on the range of retrieval frequency levels used (hence, Retrieval 5 > 1, but Retrieval 3 = 1).

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Theoretically, the results suggest that learning may not occur incrementally but in an all-or-none fashion (one-trial learning controversy; see Roediger & Arnold, 2012, for a review).

One caveat to be considered is that scores of the Retrieval 5 and 7 groups approached the ceiling on the immediate productive and receptive posttests (Table 1). As a result, the lack of statistical significance between these two groups on the immediate posttest could be in part due to a potential ceiling effect. The delayed posttests, however, did not show a ceiling or floor effect. The lack of significant differences between the Retrieval 5 and 7 groups on the delayed posttest, thus, suggests that seven retrievals may not significantly increase long-term retention compared with five. Because scores on the delayed posttest are more important than those on the immediate posttest from a pedagogical standpoint, this study nonetheless has pedagogical significance. At the same time, although the present study found statistically significant differences, the 95% confidence intervals (CIs) of difference in the means were relatively wide. The 95% CI of difference in the means between repeated (five and seven) and fewer retrievals (one and three) for the posttest scores ranged from 7% to 49% (Tables S2 and S3 in the online supplementary materials). Given the relatively wide CI of difference, it may be useful to replicate this research with a larger sample size to gain a more precise estimate of treatment effect.

The second research question was concerned with the efficiency of the four retrieval frequency levels. Based on the retrieval effort hypothesis and desirable difficulty framework, Hypothesis 2 predicted that within-session repeated retrieval is less efficient than fewer retrievals. The findings of this study supported Hypothesis 2. The results showed that the Retrieval 1 group was the most efficient regardless of the retention interval. Theoretically, the findings are consistent with the retrieval effort hypothesis and desirable difficulty framework, according to which the positive effects of repeated retrieval attenuate as a function of the retrieval frequency. Most learning occurred in the first retrieval attempt, and practicing retrieval more than once did not provide a benefit that justified its cost (i.e., increased study time). The findings are also consistent with previous L2 studies showing that the incremental benefit to learning decreases as the number of repetition increases (Bisson et al., 2014; Pyc & Rawson, 2009; van Zeeland & Schmitt, 2013). Methodologically, the findings underscore the importance of controlling time-on-task when comparing the effects of different treatments. It should be noted that although the findings of this study showed that within-session repeated retrieval may not be efficient, they do not necessarily deny the value of increasing retrieval frequency over multiple sessions. This is because when spacing between retrievals is long enough to induce forgetting (e.g., 1 month), the subsequent retrieval can be as difficult as or more difficult than the initial retrieval. Because increasing retrieval

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

frequency over multiple sessions may allow learners to benefit from the positive effects of both increased frequency and retrieval difficulty, repeated retrieval over multiple sessions may be efficient unlike within-session repeated retrieval. In future research, it may be valuable to identify the interval after which repeated retrievals become more efficient than fewer retrievals.

Next, let us consider the 95% CI of difference for the efficiency scores. As shown by Table S5 in the online supplementary materials, on the receptive test, the CI of difference between Retrieval 1 and the other three groups ranged from 0.41 to 1.34. This means that one retrieval allows learners to acquire 24.6 (0.41 × 60 = 24.6) to 80.4 (1.34 × 60 = 80.4) more words per hour on the receptive test. Considering that PAL allows learners to learn around 33–166 words per hour on average (e.g., Nation, 2013; Thorndike, 1908), it may be possible to argue that the advantage of one retrieval over repeated retrievals is not only statistically but also practically significant. On the productive posttest, however, the results regarding the CI were not consistent across the retention intervals. On the immediate productive posttest, the CI of difference between Retrieval 1 and the other three groups ranged from 0.27 to 1.27. This corresponds to 16.2 to 76.2 words per hour and should be considered practically significant. With sensitive scoring on the delayed productive posttest, however, the CI ranged from 0.15 to 0.72 (1-week delayed) and 0.06 to 0.63 (4-week delayed), which correspond to 9.0 to 43.2 and 3.6 to 37.8 words per hour, respectively. (Because no significant difference was found in the efficiency scores with strict scoring on the delayed posttest, only results with sensitive scoring are reported for the delayed posttest.) Considering that the lower end of the CI is close to zero (1-week delayed: 0.15; 4-week delayed: 0.06), it may be reasonable to assume that one retrieval is not necessarily efficient 1 or 4 weeks after learning for the acquisition of productive knowledge. This is possibly because the delayed productive posttest was the most demanding dependent measure. Hence, practicing retrieval only once did not, perhaps, guarantee successful performance on this measure, and more retrievals were needed for the long-term retention of productive knowledge.

Pedagogical Implications

One pedagogical implication of this study is that increasing retrieval frequency in one learning session is not necessarily desirable, at least for the short-term retention of productive knowledge and short- and long-term retention of receptive knowledge. This is based on the findings that (a) the Retrieval 1 group led to the largest gain when time-on-task

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

was controlled and (b) one retrieval allows learners to acquire 16.2 to 76.2 more words per hour on the immediate productive test and 24.6 to 80.4 more words per hour on the immediate and delayed receptive tests. When learners practice retrieval repeatedly in one learning session, retrievals will be increasingly easier and more successful. This may increase learners' confidence and create "an illusion of effective learning" (Kornell, 2009, p. 1302). However, it should be noted that within-session repeated retrieval is not necessarily efficient.

At the same time, this study suggested that although within-session repeated retrieval is not necessarily efficient, it may be desirable in terms of effectiveness. This is because the Retrieval 5 and 7 groups significantly outperformed the Retrieval 1 and 3 groups regardless of the posttest, scoring system, and retention interval, producing medium to large effect sizes. If learners are concerned more with effectiveness rather than efficiency, hence, increasing retrieval frequency in one learning session is advisable. Within-session repeated retrieval may be particularly valuable for the long-term retention of productive knowledge. This is because the CIs of difference showed that the Retrieval 1 group was not necessarily more efficient than the other groups on the delayed productive posttests.

The findings of this study are useful because they may help us to propose guidelines for how vocabulary should be learned from word lists or flashcards (e.g., Elgort, 2011; Royer, 1973). Developers of computer-assisted language learning (CALL) materials, most of which include a retrieval component (Folse, 2006; Nakata, 2011), may particularly benefit from the findings of this study. For instance, based on the findings of this study, CALL developers may use single retrieval when efficiency is the goal and encourage within-session repeated retrieval when effectiveness is the goal. When repeated retrieval is used, it may be useful to ensure that L2 words are practiced in increasingly difficult formats because gradually increasing retrieval difficulty may maximize learning (Logan & Balota, 2008). For instance, when L2 words are practiced four times, they can be practiced in the following order: receptive recognition (multiple-choice), productive recognition, receptive recall, and productive recall (Laufer & Goldstein, 2004). Although this may sound cumbersome, it may be implemented relatively easily if computer programs are used.

Because retrieval is a major component of most vocabulary learning activities such as fill-in-the-blank exercises, true or false statements, multiple-choice questions, and word-definition matching besides PAL (Nation & Webb, 2011), the findings of this study are also useful for teachers and materials writers. At the same time, the issue of whether the findings of the present study also apply to instructional activities other than PAL awaits future research. It should also be noted that because this study was concerned with the frequency of retrieval, the findings of

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

this study may not necessarily be applicable to the frequency of input, which does not necessarily involve retrieval.

Limitations of the Present Study

Although the results of the present study are useful, the study suffers from some limitations. First, in the present study, the retention interval was manipulated within participants, and each participant took posttests three times (immediate, 1-week, and 4-week delayed). This might have affected posttest performance. It would be useful to manipulate the retention interval between participants in future research to avoid the effects that earlier tests have on later tests. Another limitation is that the effects of retrieval frequency may be influenced by a number of factors such as the difficulty or number of target items, nature of learning tasks, type of posttests, or learners' memory capacity. As a result, the findings of this study may not necessarily be applicable to contexts other than those of this study. Especially, it should be noted that the treatment in this study was restricted to PAL. Although the use of PAL is beneficial in that it allows the experimenter to exert strict control over the treatment, it may have limited ecological validity. Further research investigating the effects of within-session repeated retrieval on vocabulary learning during other kinds of tasks such as reading (Barcroft, 2013) would be a useful follow-up to this study.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263116000280.

Received 25 April 2015
Accepted 10 June 2016
Final Version Received 21 June 2016

NOTES

1. One anonymous reviewer noted that Rohrer et al.'s (2005) study should be excluded from the Literature Review because it does not involve L2 vocabulary learning. Nonetheless, Rohrer et al.'s study was included for three reasons. First, although Rohrer et al. did not examine L2 vocabulary learning, the treatment in their study involved associating novel forms (e.g., *Chiba*, *acrogen*) with known L1 words (e.g., *Japan*, *fern*), a process similar to that of L2 vocabulary learning. As a result, it was judged valuable to discuss their results. Please note also that the city names used in Rohrer et al.'s Experiment 1 were essentially L2 words because they come from languages other than English (participants' L1)

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

such as Chinese (*Yumen*), Japanese (*Chiba*), Italian (*Bari*), or Spanish (*Lugo*). Second, previous L2 vocabulary acquisition research has shown the value of drawing on non-L2 vocabulary research. For instance, non-L2 vocabulary studies examining the depth of processing theory, transfer-appropriate processing theory, testing effect, and generation effect have provided useful starting points for L2 vocabulary acquisition research (e.g., Barcroft, 2002, 2007, 2015; Nation, 2011, 2013). This suggests that it may be useful to survey the non-L2 vocabulary literature when researching L2 vocabulary acquisition. Third, Rohrer et al.'s study indicates that although within-session repeated retrieval is effective after a relatively short delay (1 week), its advantage disappears over time (3, 4, and 9 weeks after learning). Based on their findings, the present study gave posttests 4 weeks after the treatment to examine whether Rohrer et al.'s findings also apply to L2 vocabulary learning. As the preceding discussion suggests, although Rohrer et al. (2005) did not examine L2 vocabulary learning, their study offers valuable implications as to how L2 vocabulary acquisition research should be conducted.

2. As one anonymous reviewer points out, it is possible that the presentation of target word parts on the productive pretest (e.g., _ *u* _ _ for *husk*) and the provision of the target L2 word forms on the receptive pretest (e.g., *husk* = ____?) constitute (at least partial) input to target words and might have affected subsequent learning. However, it would be reasonable to assume that neither the presentation of target word parts nor that of target word forms on the pretests had a major effect on the results of this study because there is no theory to predict that input prior to the treatment affected the conditions examined in this study differentially. An analysis of learning phase performance, which found that all four groups performed roughly equally during the period when they underwent identical experimental procedures (see Appendix S1 in the online supplementary materials for details), supports this view.

REFERENCES

Anderson, R. C., & Nagy, W. E. (1992). The vocabulary conundrum. *American Educator*, *16*, 14–18; 44–47.

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

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*, 316–321.

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning*, *52*, 323–363.

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

Barcroft, J. (2012). *Input-based incremental vocabulary instruction*. Alexandria, VA: TESOL.

Barcroft, J. (2013). *Four input-retrieval patterns and incidentally oriented vocabulary learning during reading*. Paper presented at the Vocab@Vic Conference, Wellington, New Zealand.

Barcroft, J. (2015). *Lexical input processing and vocabulary learning*. Amsterdam, The Netherlands: John Benjamins.

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *31*, 635–650.

Bisson, M.-J., van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). The role of repeated exposure to multimodal input in incidental acquisition of foreign language vocabulary. *Language Learning*, *64*, 855–877.

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

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

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

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

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*, 236–246.
Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*, 693–713.
Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*, 367–413.
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*, 365–414.
Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics*, *16*, 409–435.
Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: A case study. *Language Learning Journal*, *36*, 239–248.
Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor, MI: University of Michigan Press.
Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*, 273–293.
Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*, 539–558.
Judd, E. L. (1978). Vocabulary teaching and TESOL: A need for re-evaluation of existing assumptions. *TESOL Quarterly*, *12*, 71–76.
Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968.
Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297–1317.
Larson, D. N., & Smelley, W. A. (1972). *Becoming bilingual: A guide to language learning*. Pasadena, CA: William Carey Library.
Laufer, B. (2005). Focus on Form in second language vocabulary learning. *EUROSLA Yearbook*, *5*, 223–250.
Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436.
Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research*, *15*, 391–411.
Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*, 89–108.
Lewis, M. (1993). *The Lexical Approach*. Hove, UK: Language Teaching Publications.
Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, *15*, 257–280.
Mondria, J. A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings: An experimental comparison of the "Meaning-Inferred Method" and the "Meaning-Given Method." *Studies in Second Language Acquisition*, *25*, 473–499.
Mondria, J. A., & Wiersma, B. (2004). Receptive, productive and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 79–100). Amsterdam, The Netherlands: John Benjamins.
Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*, 17–38.
Nakata, T. (2015a). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*, 677–711.
Nakata, T. (2015b). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research*, *19*, 416–434.
Nakata, T., & Webb, S. (in press). 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*. doi:10.1017/S0272263115000236.

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82.
Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching*, *44*, 529–539.
Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge, UK: Cambridge University Press.
Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle Cengage Learning.
Oxford, R., & Crookall, D. (1990). Vocabulary learning: A critical analysis of techniques. *TESL Canada Journal*, *7*, 9–30.
Paribakht, T. S., & Wesche, M. B. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174–200). Cambridge, UK: Cambridge University Press.
Peters, E. (2012). The differential effects of two vocabulary instruction methods on EFL word learning: A study into task effectiveness. *International Review of Applied Linguistics in Language Teaching*, *50*, 213–238.
Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research*, *18*, 75–94.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.
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*, 437–447.
Roediger, H. L., & Arnold, K. M. (2012). The one-trial learning controversy and its aftermath: Remembering Rock (1957). *The American Journal of Psychology*, *125*, 127–143.
Rogers, J., Webb, S., & Nakata, T. (2015). Do the cognacy characteristics of loanwords make them more easily learned than noncognates? *Language Teaching Research*, *19*, 9–27.
Rohrer, D., Taylor, K. M., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology*, *19*, 361–374.
Royer, J. M. (1973). Memory effects for test-like-events during acquisition of foreign language vocabulary. *Psychological Reports*, *32*, 195–198.
Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 199–227). Cambridge, UK: Cambridge University Press.
Steinel, M. P., Hulstijn, J. H., & Steinel, W. (2007). Second language idiom learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency. *Studies in Second Language Acquisition*, *29*, 449–484.
Thorndike, E. L. (1908). Memory for paired associates. *Psychological Review*, *15*, 122–138.
van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*, 609–624.
Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*, 33–52.
Webb, S. (2007a). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46–65.
Webb, S. (2007b). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research*, *11*, 63–81.
Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*, 232–245.
Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*, 360–376.
Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*, 568–579.

https://doi.org/10.1017/S0272263116000280 Published online by Cambridge University Press