


Research Article

EFFECTS OF LEARNING DIRECTION IN RETRIEVAL PRACTICE ON EFL VOCABULARY LEARNING

Masato Terai *

Nagoya University

Junko Yamashita

Nagoya University

Kelly E. Pasich

Nagoya University

Abstract

In paired-associate learning, there are two learning directions: L2 to L1 (L2 words as stimuli and L1 words as responses) and L1 to L2 (L1 words as stimuli and L2 words as responses). Results of previous studies that compared the effects of the two learning directions are not consistent. We speculated that the cause of this inconsistency may be L2 proficiency, as the strengths of the lexical links between L2 and L1 are different depending on the learner's L2 proficiency. This hypothesis was examined with 28 native speakers of Japanese learning English. Participants studied novel English words in the two learning directions. The results of posttests showed that for lower-proficiency learners, L2-to-L1 learning was more effective than L1-to-L2 learning, while for higher-proficiency learners, L1-to-L2 learning was more effective. The findings suggest that L2 proficiency influences the effects of learning direction on vocabulary learning.

INTRODUCTION

Vocabulary knowledge plays an important role in the process of acquiring a second language: without vocabulary knowledge, we cannot read, listen, speak, or write. For that reason, the effectiveness of vocabulary learning methods has been widely studied in second language acquisition research (e.g., Nakata, 2017; Nation, 2013; Webb, 2007; Webb & Chang, 2015). One of the most favored L2 vocabulary learning methods is paired-associate learning: learners study second language (L2) words and first language

* Correspondence concerning this article should be addressed to Masato Terai, Graduate School of Humanities, Nagoya University, Nagoya 464–0814 Japan. E-mail: teraimasato0915@gmail.com

(L1) words as a pair (for Japanese learners of English, e.g., *DOG* - 犬). In psychology, *retrieval* in paired-associate learning has gained attention for its central role in retaining a learned word in memory (e.g., Karpicke & Roediger, 2008; for a review, see Roediger & Karpicke, 2006). Retrieval is defined as the process of accessing stored information (e.g., Roediger & Guynn, 1996). An example of retrieval in paired-associate learning would be a learner seeing a word in one language, then retrieving the translation equivalent in the other (e.g., *DOG* = ?? or 犬 = ??). Previous studies have revealed that practices that include retrieval facilitate memory retention more than *study*, in which both a target (e.g., an L2 word) and its answer (e.g., the corresponding L1 word) are presented simultaneously (e.g., Barcroft, 2007; Carpenter et al., 2006). In addition, it has been speculated that the difficulty of each retrieval affects paired-associate learning, with more difficult retrieval leading to better retention than less difficult retrieval (Pyc & Rawson, 2009).

One factor influencing the difficulty of retrieval in paired-associate learning is the learning direction. There are two possible learning directions: in L2 to L1 learning, the L2 word is presented first, followed by the corresponding L1 word (e.g., *DOG* → 犬); in L1 to L2 learning, the L1 word is presented first, followed by the L2 word (e.g., 犬 → *DOG*). There is no consensus regarding the effectiveness of the two learning directions (e.g., Griffin & Harley, 1996; Schneider et al., 2002; Webb, 2009). Some studies have found that L1 to L2 learning led to better posttest performance, whereas others have found that L2 to L1 learning was superior. What may be more problematic in understanding the effect of learning direction is the lack of a common theoretical framework to serve as the basis on which researchers evaluate and explain their findings. We draw on two theoretical models, the Revised Hierarchical Model (Kroll & Stewart, 1994) and the retrieval effort hypothesis (Pyc & Rawson, 2009), to guide this study and consolidate the diverse findings of previous studies.

LITERATURE REVIEW

EFFECTS OF RETRIEVAL ON SECOND LANGUAGE VOCABULARY LEARNING

Retrieval is defined as the process of accessing stored information (e.g., Roediger & Guynn, 1996). This process is used in vocabulary learning, and due to its pedagogical implications, the learning effectiveness of this process has long caught researchers' attention (e.g., Barcroft, 2007; Karpicke & Roediger, 2007a, 2007b; Nakata, 2017; Royer, 1973). One of the earliest studies was conducted by Royer (1973), who investigated whether retrieval practice is more effective in L2 vocabulary learning than nonretrieval practice. In the study, three groups of participants studied 20 Turkish-English words pairs. The first group was asked to first study the word pairs simultaneously, then to check the English words after seeing the Turkish words (self-test-like condition). The second and third groups were allowed to see both Turkish and English pairs simultaneously, with the third group alone permitted to take as much time as they needed. Results of an immediate posttest showed that the self-test-like condition (the first group) outperformed the study-alone condition (the second and third groups). In Royer's study, the self-test-like learning can be considered *retrieval* practice, as the participants were required to retrieve the corresponding answers. However, the study-alone procedure can be considered a *non-retrieval* condition, as both cues and answers were presented simultaneously.

Barcroft (2007) utilized picture-based word learning to expand the study of retrieval practice to other learning paradigms. Barcroft compared a retrieval-oriented condition, in which a picture was presented for 6 seconds followed by the corresponding word for 6 seconds, and a control condition, in which a picture and a word were presented simultaneously for 12 seconds. After the treatment, participants took an immediate posttest and two delayed posttests (2 days and 1 week later). The results revealed that participants in the retrieval-oriented condition scored higher on all three posttests and retained more target words than those in the control condition.

Several studies have shown that increasing the repetition of retrievals will lead to better novel word retention (e.g., Nakata, 2017; Peters, 2014). Nakata (2017) investigated whether the effects of retrieval change based on the repetition of retrievals. Nakata compared four retrieval frequency groups: Retrieval 1, 3, 5, and 7. Participants in all retrieval practice groups studied 16 English–Japanese word pairs in L2 production conditions (from L1 to L2), and all retrieval attempts were followed by feedback. After the treatment, participants took one immediate posttest and two delayed posttests (1 week and 4 weeks later). Nakata found that the Retrieval 5 group and Retrieval 7 group significantly outperformed the Retrieval 1 group and Retrieval 3 group on all three posttests. However, no significant differences were found between the Retrieval 5 and 7 groups, nor between the Retrieval 1 and 3 groups. Nakata concluded that repetition of retrievals leads to short-term and long-term benefits in vocabulary retention.

Thus, the effects of retrieval practice on the retention of novel words have been widely acknowledged in previous research. To explain this positive effect of retrieval practice, Pyc and Rawson (2009) proposed the retrieval effort hypothesis. This hypothesis applies the desirable difficulty framework, which claims that difficult but successful processing leads to better retention than easier and successful processing (Bjork, 1994, 1999), to the specific process of retrieval, stating that effortful but successful retrievals are more effective than effortless retrievals. That is, a retrieval that imposes a higher cognitive demand on the learner is more effective than one imposing a lower cognitive demand; however, the cognitive demand cannot be so high that it impedes retrieval entirely.

Pyc and Rawson (2009) conducted research to test this hypothesis using paired-associate learning. In the experiment, native English speakers were asked to memorize 70 Swahili–English translation word pairs. The difficulty of retrievals was operationalized by manipulating two variables: the interstimulus interval (ISI, defined as the duration of time between each practice trial with a given item) and the criterion level (CL, defined as the number of times that items were required to be correctly recalled). If the retrieval effort hypothesis is correct, two assumptions can be made. The first assumption is that a longer ISI is more difficult than a shorter ISI. The previous study reported that shorter ISIs led to faster response latencies for correct retrievals than longer ISIs (Karpicke & Roediger, 2007a); thus, a longer ISI condition is more difficult, which should lead to a better learning outcome than a shorter ISI. The second assumption is that a lower CL is more difficult than a higher CL. Karpicke and Roediger found that as the CL increased, response latencies in retrieval practice decreased; therefore, a lower CL is more difficult, which should lead to more effective learning than a higher CL. Indeed, the results showed that longer ISIs and lower CLs led to more difficult but correct retrievals and higher levels

of performance in the final test. Thus, the results supported Pyc and Rawson's retrieval effort hypothesis.

The central claim of the retrieval effort hypothesis is similar to that of the involvement load hypothesis, which is used as a theoretical framework in a number of L2 vocabulary studies. The involvement load hypothesis assumes that retention of unfamiliar words depends on the degree of task involvement, which consists of motivational (*need*) and cognitive (*search* and *evaluation*) components (Laufer & Hulstijn, 2001). Thus, the hypothesis speculates that words will be better retained in a higher involvement load than in a lower involvement load. However, the involvement load hypothesis's intent is to explain the learning effects of *incidental* vocabulary learning (Laufer & Hulstijn, 2001; Yanagisawa & Webb, 2021). By contrast, the retrieval effort hypothesis does not have such restrictions (Pyc & Rawson, 2009). Therefore, the retrieval effort hypothesis is a more appropriate theoretical framework for retrievals in paired-associate learning because it is *intentional* vocabulary learning.

EFFECTS OF LEARNING DIRECTION ON SECOND LANGUAGE VOCABULARY LEARNING

In paired-associate learning, there are two learning directions: either the L2 form is shown to learners first, followed by the L1 equivalent (e.g., *DOG* → 犬), or the L1 is presented first, followed by the L2 form (e.g., 犬 → *DOG*). The former, which might be the most common way to learn L2 vocabulary, is often called receptive vocabulary learning, while the latter is often called productive vocabulary learning (e.g., Nation, 2013; Webb, 2009). However, some research uses the terms somewhat differently, defining "receptive learning" as learning through reading and listening and "productive learning" as learning through speaking and writing (see Nation, 2013, pp. 46–47, for review of this point). Thus, the definitions of "receptive" and "productive" vary in different research areas, and there is no consensus on the definition. To avoid confusion, the current study uses the following definitions: learning from L2 to L1 (e.g., *DOG* → 犬) is simply called *L2 to L1 learning*, and learning from L1 to L2 (e.g., 犬 → *DOG*) is called *L1 to L2 learning*.

Previous studies have investigated the efficacy of the two types of vocabulary learning (L2 to L1 learning and L1 to L2 learning) (e.g., Griffin & Harley, 1996; Schneider et al., 2002; Webb, 2009); however, the findings are inconsistent.

Some researchers have found no significant differences between the two learning directions. Griffin and Harley (1996) recruited English native speakers learning French. The participants, who had been learning French for less than a year, can be considered to have been below intermediate proficiency. In the experiment, participants had to learn French–English word pairs, either in the L2 to L1 direction or in the L1 to L2 direction. After the learning session, participants took either an L1 production test, which required them to write L1 words in response to L2 cues, or an L2 production test, which required them to write L2 words in response to L1 cues. There was no significant difference between the two learning directions.

Other studies have found a time component to the effect of learning directions. Schneider et al. (2002) investigated which of the two learning directions leads to better short-term and long-term retention of newly learned L2 vocabulary. The participants in

their study were native English speakers who had never studied French before (i.e., complete beginners). The participants learned English and French word pairs, then took an immediate and delayed post-test (1 week after the treatment). The L2 to L1 learning group took L1 production tests, whereas the L1 to L2 learning group took L2 production tests. The scores of the L2 to L1 learning group were higher than those of the L1 to L2 learning group in the immediate posttest; however, this result reversed in the delayed posttest, with the L1 to L2 learning group outperforming the L2 to L1 learning group. Thus, Schneider et al. reported that although L2 to L1 learning is effective in the short term, L1 to L2 learning is more effective in the longer term.

Finally, some research has found decisive differences between the two learning directions. For example, Webb (2009) compared the two types of learning with Japanese native speakers learning English. The participants' proficiency was measured with Version 1 of the Vocabulary Levels Test (Schmitt et al., 2001) at the 2000-word level and Version C of the Productive Levels Test (Laufer & Nation, 1999) at the second 1,000-word level. According to the results of the two tests, the L2 proficiency of the participants can be considered to have been high-intermediate or above. The participants studied 10 words in either the L2 to L1 or the L1 to L2 condition. The target words were 10 English-like nonsense words (e.g., *masco*). The participants then took 10 different vocabulary tests (L2 to L1/L1 to L2 knowledge of orthography, meaning and form, association, syntax, and grammatical functions). The participants in the L1 to L2 learning group achieved significantly higher scores in L2 knowledge of orthography when the scores of L2 to L1 and L1 to L2 tests were combined. By contrast, the participants in the L2 to L1 learning group did not statistically outperform the L1 to L2 learning group in any test of L2 vocabulary knowledge.

Although it is not easy to reconcile these conflicting results, a possible explanation is suggested by the Revised Hierarchical Model (RHM) (e.g., Kroll & Stewart, 1994). RHM postulates that the strength of the lexical connection from the L2 lexicon to the L1 lexicon is stronger than the lexical link from the L1 lexicon to L2 lexicon. Kroll and Stewart, who investigated whether there is a difference in translation latency and accuracy in bilingual oral translation depending on the translation direction, revealed an asymmetry of the strength of the lexical links. Dutch native speakers of L2 English were asked to orally translate both English and Dutch words into their equivalents. The results showed that the bilinguals' translation latency from L1 to L2 was slower than from L2 to L1. In addition, L1 to L2 translation accuracy was lower than in the L2 to L1 direction. Kroll and Stewart concluded that these differences in latency and accuracy reflected the difference of strength of the connection between the two lexicons. That is, the L1 to L2 lexical link is weaker than the L2 to L1 lexical link because L2 words are initially associated with the L1 (Kroll & Stewart, 1994). From this asymmetry in the lexical links between L2 to L1 and L1 to L2, we can infer that the cognitive demand is higher when retrieving the L2 equivalent of an L1 word rather than the other way around.

Furthermore, the aforementioned assumption of the cognitive demand, combined with the retrieval effort hypothesis, suggests that the learning effect will be bigger in L1 to L2 learning than in L2 to L1 learning. However, as shown in the preceding text, the previous studies returned differing results. A factor that may account for the discrepancies may be the participants' L2 proficiency. One aspect that makes RHM unique

compared to other influential models of bilingual mental lexicon, such as the Bilingual Interactive Activation + Model (Dijkstra & Van Heuven, 2002) and the Multilink Model (Dijkstra et al., 2019), is that RHM involves a developmental hypothesis about the lexical connections between L2 lexicon and L1 lexicon. Namely, RHM asserts that the asymmetry of the links will diminish as the speaker's L2 proficiency increases (Kroll et al., 2002; Kroll & Sunderman, 2003). Kroll et al. (2002) conducted experiments to test the developmental aspect of RHM by comparing learners at different levels of L2 proficiency. In the experiment, participants were instructed to translate words, either L2 to L1 or L1 to L2. The results showed that the participants were faster to translate words from L2 to L1 than L1 to L2; however, the magnitude of the difference was larger for the less proficient participants than for the more proficient learners. Thus, the results supported the developmental aspect of RHM: the more proficient learners are, the smaller the asymmetry in the lexical links will be (Kroll & Sunderman, 2003).

The developmental perspective can also be applied to the change in cognitive demand in the two learning directions. As stated previously, L1 to L2 learning is much more demanding than L2 to L1 learning. We can predict that this difference in cognitive demand should be larger for low-proficiency learners. For high-proficiency learners, however, the gap should be smaller; though L1 to L2 learning may still be the more difficult direction, the cognitive demand is more reasonable for high-proficiency learners because their L1 to L2 lexical connection is stronger.

Hence, we infer two predictions from RHM and the retrieval practice hypothesis: first, the main effect should be bigger in L1 to L2 learning than in L2 to L1 learning; and second, the effects of L1 to L2 learning should be bigger for high-proficiency learners than for low-proficiency learners. The inconsistency in the previous studies may be due to the variety of L2 proficiency levels in their participants. A study comparing the effects of the two learning directions in learners of high-intermediate or advanced proficiency is likely to find a superior learning effect in L1 to L2 learning (e.g., Webb, 2009). However, a study will likely find a superior effect in L2 to L1 learning if the participants are low-proficiency learners (e.g., Schneider et al., 2002). Finally, the L2 to L1 and L1 to L2 learning effects might be offset for learners of lower-intermediate proficiency, which would result in no significant difference (e.g. Griffin & Harley, 1996). Table 1 summarizes the previously mentioned predictions by listing L2 proficiency and the effectiveness of learning direction of previous studies.

TABLE 1. Summary of previous research

| Research | Proficiency | Results |
|---------------------------|--|--|
| Griffin and Harley (1996) | Below intermediate (first year in learning L2) | No significant differences |
| Schneider et al. (2002) | Very low (no prior L2 knowledge) | L2 to L1 learning is more effective (Immediate) L1 to L2 learning is more effective (Delayed) |
| Webb (2009) | High-intermediate | L1 to L2 learning is more effective (L2 orthography) |

RESEARCH QUESTIONS

This study focuses on the relationship between L2 proficiency and the effectiveness of the two learning directions in paired-associate learning in L2 vocabulary acquisition. Although our review suggests a possible impact of L2 proficiency, any comparison of the relative levels of participants' L2 proficiency across different studies would be mere speculation. By examining the effect of L2 proficiency in a single experiment, we can test its influence more rigorously by including it in statistical modeling and controlling various potentially confounding factors (linguistic, educational, and sociocultural). The following research questions guided this study:

1. Which learning direction is more effective for vocabulary acquisition, L1 to L2 learning or L2 to L1 learning?
2. How does L2 proficiency influence the effects of L2 to L1 and L1 to L2 learning?

Based on previous studies (Schneider et al., 2002; Webb, 2009) and predictions based on RHM and the retrieval effort hypothesis, we established the following hypotheses:

1. There is no significant difference between L2 to L1 learning and L1 to L2 learning in the retention rate of novel words when no distinction is made between proficiency levels. The effects of learning direction will vary according to the developmental change in the strength of the L1 to L2 and L2 to L1 lexical links in the mental lexicon; thus, the learning effect of the two methods will be offset if we do not consider L2 proficiency.
2. The effect of learning direction depends on the learner's L2 proficiency. More specifically, the impact of L1 to L2 learning will be more pronounced for high-proficiency learners because the difficulty of L1 to L2 learning is moderated as the strength of the lexical link from L1 to L2 increases with proficiency (i.e., the cognitive demand of L2 to L1 learning is not sufficiently high for high-proficiency learners). However, the effects of L2 to L1 learning will be larger for low-proficiency learners, as the L2 to L1 lexical link is solid from the initial stage of vocabulary knowledge (i.e., the cognitive demand of L1 to L2 learning is too high for low-proficiency learners).

METHOD

PARTICIPANTS

A total of 28 Japanese native speakers learning English at a Japanese university participated in this study. The number of participants was statistically decided based on power analysis (Green & MacLeod, 2016) (see Appendix A). Participants came from a variety of majors, including agriculture, law, and mathematics. Because this study examined L2 proficiency as an important factor, we collected data from learners of various proficiencies. A questionnaire that included questions about participants' English learning history revealed that participants had studied English for at least 6 years in school. Although some participants reported that they had studied in an English-speaking country for about one month, the participants studied English largely as a foreign (as opposed to second) language. After the main experiment, participants took the V_YesNo v1.0 test, which measures L2 vocabulary size (Meara & Miralpeix, 2016) and was used as a proxy for participants' English proficiency. The test has a maximum score of 10,000 and a minimum score of 2,500. Based on their criteria for score estimation, scores from 2,500 to

TABLE 2. Descriptive statistics of the participants

| | <i>M</i> | <i>SD</i> | <i>Mdn</i> | <i>Minimum</i> | <i>Maximum</i> | <i>Skew</i> | <i>Kurtosis</i> |
|------------------------|----------|-----------|------------|----------------|----------------|-------------|-----------------|
| Vocabulary size | 5,002.71 | 945.96 | 4,877.50 | 3,466.00 | 6,970.00 | 0.28 | -0.52 |
| Age | 21.11 | 2.25 | 20.50 | 18.00 | 27.00 | 0.95 | 0.30 |
| Years learning English | 10.29 | 4.10 | 9.50 | 5.00 | 24.00 | 1.86 | 4.31 |

Note: $n = 28$. Vocabulary Size refers to V_YesNo v1.0 test scores (Meara & Miralpeix, 2016).

3,500 are considered beginner level, 3,500 to 6,000 intermediate level, and 6,000 to 10,000 proficient level. The results of the vocabulary size test revealed that there was a considerable variation in the L2 proficiency of the participants, ranging from the beginner level to the proficient level. Table 2 shows the descriptive statistics of the participants.

MATERIALS

The target items were 40 low-frequency English words paired with their Japanese translation equivalents (e.g., *bluff* = 絶壁). The target English items were neither loanwords nor cognates in Japanese. The target items were selected from Nakata and Suzuki (2019) and Nakata and Webb (2016), who also studied L1 Japanese learners of English and controlled L2-related variables (L2 frequency, L2 letters, syllables) and L1-related variables (L1 frequency, L1 letters, mora, L1 familiarity) in their materials. After selecting items from them, we extracted the L2 frequency of the target words from the Corpus of Contemporary American English (COCA) (Davies, 2008). L2 word length was defined as the number of letters and syllables. L1 frequency and L1 familiarity were retrieved from Amano and Kondo (2000) and Amano and Kondo (1999), respectively. L1 familiarity was calculated using rating scores on a 7-point scale, where 1 means unfamiliar and 7 means familiar. Additionally, familiarity ratings from the participants in the current study were also obtained (see Table 3 for the item characteristics). As the participants

TABLE 3. Descriptive statistics of the target items

| | <i>M</i> | <i>SD</i> | <i>Mdn</i> | <i>Minimum</i> | <i>Maximum</i> | <i>Skew</i> | <i>Kurtosis</i> |
|----------------------|----------|-----------|------------|----------------|----------------|-------------|-----------------|
| L2-related variables | | | | | | | |
| Frequency | 1,025.33 | 851.93 | 813.50 | 51.00 | 3,930.00 | 1.57 | 2.68 |
| Syllables | 2.00 | 0.91 | 2.00 | 1.00 | 5.00 | 0.87 | 1.37 |
| Letters | 6.23 | 1.72 | 6.00 | 3.00 | 12.00 | 1.00 | 1.89 |
| L1-related variables | | | | | | | |
| Frequency | 596.90 | 885.49 | 275.00 | 6.00 | 5,109.00 | 3.73 | 17.34 |
| Letters | 3.68 | 1.07 | 4.00 | 1.00 | 6.00 | 0.18 | 0.72 |
| Mora (syllables) | 3.53 | 0.96 | 4.00 | 1.00 | 6.00 | 0.20 | 1.63 |
| Fami (A) | 5.22 | 0.64 | 5.36 | 3.72 | 6.38 | -0.65 | -0.02 |
| Fami (B) | 4.19 | 0.80 | 4.29 | 2.50 | 5.89 | -0.02 | -0.23 |

Note: $n = 40$. Frequency (L2-related variables) = raw frequency in COCA (Davies, 2008); Frequency (L1-related variables) = raw frequency (Amano & Kondo, 2000); Fami (A) = familiarity ratings from Amano and Kondo (1999); Fami (B) = familiarity ratings from the current study.

studied in both the L2 to L1 and L1 to L2 learning conditions, the 40 words were divided into two lists, Vocabulary A and Vocabulary B (see Appendix B). The order of the two sets of vocabulary was counterbalanced.

POSTTESTS

Two immediate posttests were administered: an L1 production test and an L2 production test. Both tests were word-form recall tests. The participants had to write the English or Japanese words corresponding to the Japanese or English stimuli. For example, in the L1 production test, the participants had to write the Japanese translation equivalents of the given English word (e.g., *DOG* = ??) (Laufer & Rozovski-Roitblat, 2011). In the L2 production test, the participants had to write the English words corresponding to the given Japanese word (e.g., 犬 = ??) (Webb, 2005). Because the 40 target words were divided into two sets of 20, there were four types of tests: L1 production test A, L1 production test B, L2 production test A, and L2 production test B. Both the L1 and L2 production tests were scored by the strict method, which does not allow for any spelling mistakes: a correct answer is scored as 1 point and an incorrect answer is scored as 0 points (Nakata, 2017).

PROCEDURE

The study was conducted using a computer program coded by the first author using Hot Soup Processor version 3.5. (<http://hsp.tv/>). The experiment had three phases: the participant saw the items (exposure), studied them (learning: L1 to L2 or L2 to L1), and were tested (testing). Experiments were conducted individually in a quiet room. Before the experiment, all participants signed a consent form and received instructions for the experiment in Japanese. The instructions were orally provided as follows:

In this experiment, you have to try to learn new English words and their Japanese translation equivalents. First, 20 words will be presented on the screen for 3 seconds each. Next, you will study those words again by typing their English or Japanese equivalents. After the learning phase, you will take two types of tests. You will repeat the process one more time with different words and in different conditions. (English translation of the Japanese instruction)

After a practice session, the experiment started. First, in the exposure phase, the participants saw 20 English–Japanese target word pairs on a computer screen. The English–Japanese word pairs were presented in the middle of the computer screen; the English words were placed on the left and the Japanese words were on the right (e.g., *bluff* = 絶壁). Each target word pair was presented for 3 seconds. After the exposure phase, participants studied the target words in either the L2 to L1 condition (e.g., *bluff* = ??) or the L1 to L2 condition (e.g., 絶壁 = ??). For example, in the L2 to L1 condition, the English word was presented and participants were asked to type the Japanese translation equivalent. After each response, the correct Japanese translation equivalent was provided whether the participant's response was correct or not. The participants were able to take as much time as they needed to learn the L1 and L2 word association in the learning phase. Immediately after the learning phase, participants took the L1 production test, then the L2 production test. Because the participants took two rounds of tests, the order of the practice types

(i.e., learning direction) and words (vocabulary list A or B) were counterbalanced to avoid the practice effect and the order effect. At the end of the experiment, all participants took the vocabulary size test (Meara & Miralpeix, 2016), completed familiarity ratings of the Japanese words on the tests, and took a known word check test and a background questionnaire.

ANALYSIS

First, all data for words known to each participant were deleted from the dataset. Then, the data were analyzed in a series of generalized linear mixed-effects models (GLMM) using RStudio 3.6.1 (R Core Team, 2019) and the *lme4* package (Bates et al., 2015). GLMM analysis was employed to examine three explanatory variables: Learning Condition (L2 to L1/L1 to L2), Test Type (L1 production/L2 production), and Vocabulary Size (L2 Proficiency), as well as interaction terms between two variables. The continuous variable (Vocabulary Size) was scaled, and the categorical variables (Learning Condition and Test Type) were contrast-coded before creating models to avoid convergence issues (Tamura et al., 2019). The models were built with variables chosen based on the research interests, and the final models were chosen by comparing the Akaike Information Criterion (AIC) of the models with and without interactions. The AIC shows the goodness of the model fit; the lower the AIC, the better the model.

In total, three models were chosen for analysis. The first model was built to analyze the relationship between the production tests and learning conditions (Research Question 1). The model contained Learning Condition and Test Type as explanatory variables, as well as the interaction of the two variables. Random effects (Subject and Item) were included, and production test answers were used as the response variable.

The second and third models were built to examine the effects of the two types of learning based on the results of the production tests (Research Question 2). The two models were the same except that the second model included L1 production test scores as the response variable, while the third model utilized L2 production test scores. Test Type, Vocabulary Size, and the interaction of the two variables were added as explanatory variables, including random effects (Subject and Item).

After each of the three models was built, Variance Inflation Factors (VIF) were checked to confirm that there were no multicollinearity issues. The VIF cutoff point was 5. Finally, if the GLMM analyses revealed a significant effect of the interaction of variables, the *phia* package (De Rosario-Martinez, 2015) was applied to find the simple main effects of variables. Then, the *emmeans* package (Lenth, 2019) was used to conduct multiple comparisons.

RESULTS

EFFECTS OF LEARNING CONDITION

Table 4 shows the descriptive statistics of the two types of post-tests. Cronbach's α showed adequate reliability of all the tests (ranging from .73 to .84) (Table 5).

The first model was applied to investigate the effectiveness of learning direction (Research Question 1). The first model includes Learning Condition and Test Type as

TABLE 4. Descriptive statistics of the tests

| | <i>M</i> | <i>SD</i> | <i>Mdn</i> | <i>Minimum</i> | <i>Maximum</i> | <i>Skew</i> | <i>Kurtosis</i> |
|--------------------|----------|-----------|------------|----------------|----------------|-------------|-----------------|
| L1 production test | | | | | | | |
| L2 to L1 | 9.71 | 4.16 | 9.00 | 4.00 | 18.00 | 0.44 | -0.92 |
| L1 to L2 | 9.00 | 3.85 | 9.00 | 0.00 | 16.00 | -0.15 | -0.30 |
| L2 production test | | | | | | | |
| L2 to L1 | 5.64 | 3.49 | 5.00 | 0.00 | 13.00 | 0.40 | -0.92 |
| L1 to L2 | 6.39 | 3.52 | 6.00 | 0.00 | 13.00 | -0.01 | -0.80 |

Note: *n* = 28. L2 to L1 refers to L2 to L1 learning; L1 to L2 refers to L1 to L2 learning.

TABLE 5. Alpha coefficients for L1 production test and L2 production test

| | L1 production test | | L2 production test | |
|--------------|---------------------|------------|---------------------|------------|
| | Cronbach's <i>α</i> | 95% CI | Cronbach's <i>α</i> | 95% CI |
| Vocabulary A | .84 | [.75, .92] | .82 | [.73, .92] |
| Vocabulary B | .74 | [.60, .88] | .73 | [.59, .87] |

Note: Vocabulary A refers to the L1 or L2 production test for Vocabulary A; Vocabulary B refers to the L1 or L2 production test for Vocabulary B.

explanatory variables. The first model also includes the interaction of the two explanatory variables, as the model with the interaction showed lower AIC than the model without interaction (with interaction: 2,427.10; without interaction: 2,429.77). Participant and item intercepts were used as crossed random effects. The results showed a significant main effect of Test Type (Estimate = -0.976, *SE* = 0.105, *z* = -9.315, *p* < .001), and the interaction of Test Type and Learning Condition was also significant (Estimate = -0.446, *SE* = 0.206, *z* = -2.169, *p* = .030); however, there was no main effect of Learning Condition (Estimate = -0.038, *SE* = 0.103, *z* = -0.366, *p* = .714). As the interaction of the two variables was significant, the *phia* package (De Rosario-Martinez, 2015) and the *emmeans* package (Lenth, 2019) were applied to find the simple main effects of the variables and conduct multiple comparisons. The results revealed that there was a statistically significant difference between the scores of the two tests, suggesting that L1 production test scores were higher than L2 production test scores in both the L2 to L1 and L1 to L2 learning conditions (L2 to L1 learning: *p* < .001, *d* = 1.20, 95% CI [0.91, 1.49]; L1 to L2 learning: *p* < .001, *d* = 0.75, [0.47, 1.04]). However, there were no simple main effects of Learning Condition (L1 production test: *p* = .188, *d* = -0.19, 95% CI [-0.46, 0.09]; L2 production test: *p* = .082, *d* = 0.26, [-0.03, 0.55]) (Figure 1). All variables had VIF scores of approximately 1. Because of the main effect of Test Type, the two test types were analyzed separately in the following analyses.

The results revealed no significant difference in learning effects between L2 to L1 learning and L1 to L2 learning, which replicates the findings of Griffin and Harley (1996).

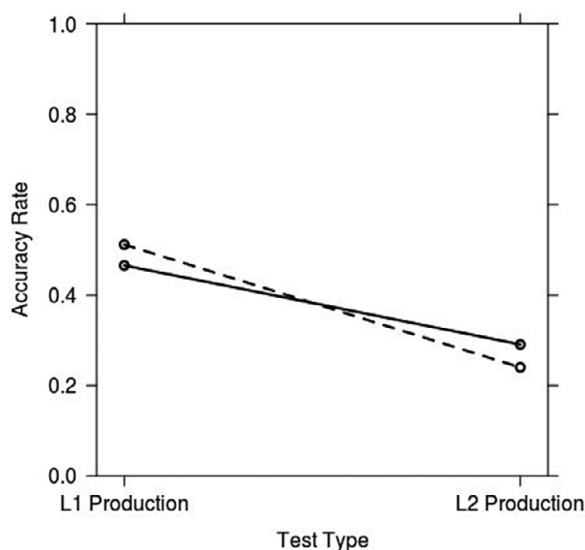


FIGURE 1. Effect plot of test type and learning condition in the GLMM.

Note: The solid line represents the L1 to L2 learning condition; the dotted line represents the L2 to L1 learning condition. The y-axis represents accuracy rates on both the L1 production test and the L2 production test. The x-axis represents Test Type: L1 Production refers to the L1 production test, and L2 Production refers to the L2 production test.

EFFECTS OF LEARNING DIRECTIONS AND L2 PROFICIENCY

The second and third models were applied to explore the influence of L2 proficiency on the two types of learning (Research Question 2).

L1 Production Test

The second model was used to analyze the scores of the L1 production test (e.g., *bluff* = ??). The second model contained Learning Condition and Vocabulary Size (English proficiency) as explanatory variables and Accuracy of the L1 production test as the response variable. The model with interaction was chosen because it showed lower AIC (with interaction: 1,311.56; without interaction: 1,316.26). All variables had VIF scores of approximately 1. Participant and item intercepts were used as crossed random effects. The results indicated that there were no significant main effects of Vocabulary Size (Estimate = 0.103, $SE = 0.186$, $z = 0.556$, $p = .058$) or Learning Condition (Estimate = 0.184, $SE = 0.141$, $z = 1.300$, $p = .194$); however, the interaction of Learning Condition and Vocabulary Size was significant (Estimate = -0.367 , $SE = 0.141$, $z = -2.596$, $p = .009$). There was no simple main effect (Figure 2).

The results showed that L2 to L1 learning was more effective for the lower-proficiency learners, while L1 to L2 learning was more effective for the higher-proficiency learners. More specifically, L1 to L2 learning was more effective than L2 to L1 learning for learners who scored more than 5,419 on the vocabulary size test (the crossover point of the solid and dotted lines in Figure 2). The plot lines also show that L2 to L1 learning is less likely to

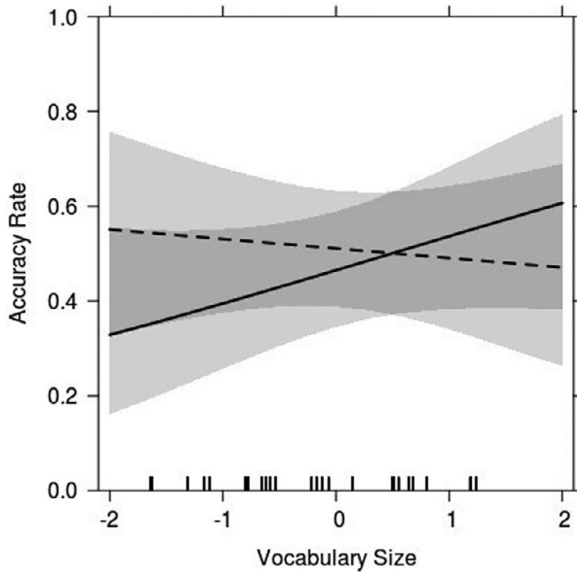


FIGURE 2. Effect plot of vocabulary size and learning condition in the GLMM (L1 production test).
 Note: The solid line represents the L1 to L2 learning condition; the dotted line represents the L2 to L1 learning condition. The response variable only includes L1 production test scores. The y-axis represents L1 production test accuracy rates. The x-axis represents scaled scores of the vocabulary size test. Values on the x-axis are standard deviations.

be influenced by L2 proficiency (the line is nearly horizontal). By contrast, L1 to L2 learning is significantly influenced by L2 proficiency (the line has a positive slope).

L2 Production Test

The third model was applied to investigate the scores of the L2 production test. The model with interaction had a lower AIC (with interaction: 1,145.32; without interaction: 1,146.52). Thus, the third model was identical to the second model except that it used L2 production test Accuracy as the response variable (VIF approximately 1 for all variables). The results revealed that there were no significant main effects of Vocabulary Size (Estimate = 0.227, SE = 0.193, z = 1.179, p = .238) or Learning Condition (Estimate = -0.262, SE = 0.154, z = -1.699, p = .089), nor was the interaction significant (Estimate = -0.278, SE = 0.154, z = -1.810, p = .070). These results are consistent with Griffin and Harley (1996): there is no difference in the learning effects of the two learning directions, and there is no influence of L2 proficiency. However, the effects plot shows an interaction trend similar to the results of the L1 production test (Figure 3). For lower-proficiency learners, L2 to L1 learning leads to higher performance than L1 to L2 learning. In addition, L2 to L1 learning is less influenced by L2 proficiency, evidenced by the line being almost completely parallel to the x-axis. However, L1 to L2 learning is more influenced by L2 proficiency, as can be seen by the line’s positive slope.

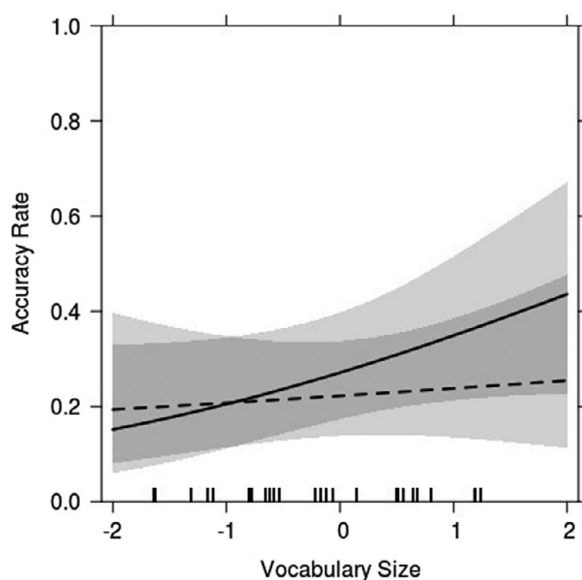


FIGURE 3. Effect plot of vocabulary size and learning condition in the GLMM (L2 production test).

Note: The solid line represents the L1 to L2 learning condition; the dotted line represents the L2 to L1 learning condition. The response variable only includes L2 production test scores. The y-axis represents L2 production test accuracy rates. The x-axis represents scaled scores of the vocabulary size test. Values on the x-axis are standard deviations.

DISCUSSION

This study investigated the effect of learning direction on paired-associate learning in L2 vocabulary acquisition (L1 to L2 vs. L2 to L1) with a special interest in the interaction between L2 proficiency and learning direction based on RHM and the retrieval effort hypothesis.

EFFECTS OF LEARNING DIRECTION ON SECOND LANGUAGE VOCABULARY LEARNING

Before discussing our main research interests (learning direction and L2 proficiency), the effect of test type deserves some attention. As shown in the preceding text, participants scored significantly higher on the L1 production test than the L2 production test regardless of learning condition. This result is consistent with studies reporting that L2 meaning recognition (operationalized by the L1 production task in this study) is acquired earlier than L2 form recall (operationalized by the L2 production task in this study) (e.g., González-Fernández & Schmitt, 2019). According to González-Fernández and Schmitt, there seems to be an acquisition order of word knowledge components:

1. Form-Meaning link meaning recognition
2. Collocate form recognition
3. Multiple-Meanings meaning recognition

4. Derivative form recognition
5. Collocate form recall
6. Form-Meaning link form recall
7. Derivative form recall
8. Multiple-Meanings recall

(p. 493)

Another possible explanation is that L1 forms (operationalized by the L1 production task in this study) are entrenched more deeply in the mental lexicon than L2 forms (operationalized by the L2 production task in this study). As established by RMH, L2 production is more difficult than L1 production. The participants in this study had neither cause nor opportunity to more deeply entrench the L2 forms on a level at all comparable to their L1: they spent most or all of their lives in the EFL context of Japan rather than in English-speaking countries, mostly used Japanese in their daily lives and at school, and received much more Japanese input than English input. It is hardly a surprise that it would be easier for them to retrieve L1 forms than L2 forms.

Now, moving on to our main research purposes, there was no difference in the learning effect between the two learning directions. Thus, the finding echoes those of Griffin and Harley (1996) and Hypothesis 1 was supported. The reason for this result may be the substantial variation in the participants' proficiency levels. Specifically, the developmental aspect of RHM may play a role. Although the lexical link from L2 to L1 is solid from the initial stages of acquisition, the strength of the lexical connection from L1 to L2 increases with the advancement of L2 proficiency (Kroll et al., 2002; Kroll & Stewart, 1994). Therefore, lower-proficiency learners are likely to benefit from L2 to L1 learning but not from L1 to L2 learning because their L1 to L2 connection is too weak to support the effort needed. However, higher-proficiency learners may benefit from both learning directions because the strengths of the L1 to L2 connection and L2 to L1 connection are similar; alternatively, higher-proficiency learners may benefit from L1 to L2 learning more than L2 to L1 learning because the latter may provide a cognitive demand that is heavy, but reasonably so. Because of the complex interaction between L2 proficiency and learning direction effectiveness, the real effect of learning direction may not be observed if learner proficiency is not considered. Thus, participant L2 proficiency must be considered when examining the effect of learning condition.

THE INFLUENCE OF L2 PROFICIENCY ON EFFECTS OF LEARNING DIRECTION

Regarding Research Question 2, which addresses the effect of L2 proficiency on learning direction, different results were obtained according to Test Type. First, the analysis of the L1 production test found a significant interaction of L2 proficiency and learning condition. The cutoff point affecting the relative effectiveness of learning direction was the vocabulary size 5,419: L1 to L2 learning or L2 to L1 learning was more effective for learners who scored above or below this cutoff point, respectively. In this test, participants who score from 3,500 to 6,000 are considered to be of intermediate level (Meara & Miralpeix, 2016). Thus, broadly, if a learner's proficiency exceeds an intermediate level, L1 to L2 learning will be more effective than L2 to L1

learning. The results of the high-proficiency participants accord with RHM and the retrieval effort hypothesis. According to RHM, the L1 to L2 lexical link is weaker than the L2 to L1 link (Kroll & Stewart, 1994). Thus, we can infer that a weaker lexical link creates a higher cognitive demand than a stronger one. In addition, the retrieval effort hypothesis maintains that more difficult retrieval leads to better memory retention than easier retrieval (Pyc & Rawson, 2009). Thus, the learning condition utilizing the weaker lexical link—the L1 to L2 link—should result in better learning outcomes. Why, then, does this positive learning effect appear only for higher-proficiency learners?

There is a possibility that the lower-proficiency learners' L1 to L2 lexical connection was too weak to receive the boosted learning effect from L1 to L2 learning. The important aspect of the retrieval effort hypothesis is not just that difficult retrieval is effective; rather, retrieval should be difficult *but successful* (e.g., Pyc & Rawson, 2009). It is likely that retrieval was too difficult for the lower-proficiency learners to accomplish. Thus, L1 to L2 learning was effective only for higher-proficiency learners, as they have already built an L1 to L2 lexical link that is solid enough to gain the expected learning effect from L1 to L2 learning. Meanwhile, L2 to L1 learning was effective for the lower-proficiency learners because the L2 to L1 lexical link is relatively solid even for low-proficiency learners (Kroll & Sunderman, 2003).

Endorsing our expectation from the review of previous studies, this study demonstrated that the L2 proficiency of the learners strongly affects the outcome of the effectiveness of learning directions. Thus, we offer an explanation for the inconsistent results of earlier studies. To reiterate, if a study compares the effects of learning direction on low-proficiency learners, it is more likely to find that L2 to L1 learning is more effective than L1 to L2 learning (e.g., Schneider et al., 2002 [immediate posttest]). If a study focuses on learners of lower-intermediate proficiency, there likely will not be any significant difference between the effects of the two learning directions (e.g., Griffin & Harley, 1996). Finally, if research is conducted with participants whose proficiency is intermediate or above, the results will suggest that L1 to L2 learning is more effective than L2 to L1 learning (e.g., Webb, 2009).

Unlike the L1 production test, the analysis of the L2 production test found no effect of learning direction. A possible explanation is the floor effect. The descriptive statistics of the two tests indicate that participants scored lower on the L2 production test than the L1 production test (see Table 4). As the González-Fernández and Schmitt (2019) study shows, L2 production knowledge (Form-Meaning link form recall) is acquired later than L1 production knowledge (Form-Meaning link meaning recognition). In this experiment, all the participants were exposed to each new word only twice (in the exposure phase and the learning phase), which was not sufficient for participants to build the ability to produce the L2 form. Interestingly, however, the effects plot shows a similar tendency to the L1 production results (Figure 3). Thus, a difference in learning effects between L2 to L1 learning and L1 to L2 learning might emerge with more exposures to the words.

Overall, the results of this study suggest that when the effects of learning direction are examined, the participants' proficiency should be considered. Otherwise, the results may produce an incomplete and unfortunately deceptive picture of how learning direction impacts vocabulary retention.

SUMMARY, PEDAGOGICAL IMPLICATIONS, AND SUGGESTIONS FOR FURTHER RESEARCH

In this study, we investigated whether there are differential effects of L2 to L1 and L1 to L2 learning on L2 vocabulary retention. A variety of previous studies returned conflicting results. However, little attention was paid to these divergences in the previous findings, and few attempts were made to explain them. We hypothesized that, according to RHM and the retrieval effort hypothesis, a learner's proficiency influences the effects of the two types of learning, and we considered learner proficiency as an important factor that might explain the effects of learning direction. The current study confirmed the effect of L2 proficiency, which led us to propose that when the effects of learning directions are examined, learner proficiency should be considered. We believe that the current research has successfully combined the knowledge of memory research (the retrieval effort hypothesis) and that of language research (the Revised Hierarchical Model) to examine L2 word learning. Vocabulary learning has largely been studied using two different approaches: those of memory research and language research. Memory researchers have examined the relationship between learning and memory retention (e.g., Karpicke & Roediger, 2007a, 2007b, 2008; Pyc & Rawson, 2009), while language researchers have focused on how bilinguals' mental lexicons develop and how the lexicons are used to understand language (e.g., Kroll et al., 2002; Kroll & Stewart, 1994). Bjork and Kroll (2015) stated that both approaches can provide useful accounts to reveal the optimal learning conditions for L2 word acquisition. This study supports their implications that L2 vocabulary research needs to consider the relationship of the mechanisms of memory retention and bilingual language processing.

This study also supports the latest framework of learning proposed by Suzuki et al. (2019), who state that practice conditions (e.g., retrieval practice, auditory/written input), linguistic difficulty (e.g., formal complexity, saliency), and individual differences (e.g., prior knowledge, cognitive aptitude) need to be considered to create optimal learning conditions. While it is, as they state, difficult to take all three factors into account when examining the learning effect, this study did include two factors (practice conditions and individual differences) and found that they interact in a significant and meaningful way. Thus, in teaching, practitioners should not choose a one-size-fits-all teaching method, but should consider their students' individual differences and linguistic factors, such as L2 proficiency, cognitive aptitude, and semantic relatedness. For example, in the EFL context, and especially in Japanese classrooms, learners' L2 proficiency is usually mixed. Taking account of the situation, L2 to L1 learning is recommended in this type of classroom context, as the effects of L2 to L1 learning are less influenced by learners' proficiency.

Despite the aforementioned contributions, this study suffers from three limitations. First, this study did not implement a delayed posttest. Schneider et al. (2002) found that the learning effect reversed from the immediate posttest to the delayed posttest. That is, L2 to L1 learning was more effective in the immediate posttest, whereas L1 to L2 learning was more effective in the delayed posttest conducted 1 week later. Therefore, further study is needed to examine the effects of learning direction in the longer term. Second, it is known that word characteristics such as concreteness, L1 frequency (e.g., De Groot & Keijzer, 2000), and L1 familiarity (Tagashira et al., 2010) affect vocabulary learning. We included some of these factors (L2- and L1-related variables) in our statistical modeling as covariates, but they were not in the final model because their inclusion did not improve the

model. However, it is worth considering word characteristics and investigating complex interactions among item features (word characteristics or linguistic difficulty), practice features, and learner features in the future research.

Finally, there is an argument that this study might have been better served by using sensitive rather than strict scoring. Strict scoring does not allow any spelling mistakes, whereas sensitive scoring provides more leeway (i.e., “partial credit”). We utilized strict scoring because of the challenges posed by deciding what types of mistakes should be counted as correct. For example, if participants produced *penomenom* or *fenomenon* instead of *phenomenon*, are these answers scored as correct? Thus, to make scoring as objective and reliable as possible, we used strict scoring. Furthermore, we were interested in whether learners can acquire correct spellings of L2 words; therefore, we told participants that only completely correct spellings would be marked as correct. However, analyzing answers using a more lenient scoring rubric, such as the lexical production scoring protocol-written (LPSP-Written), may yield interesting insights (see Barcroft, 2002, for more details). LPSP-Written determines the score using a 5-point scale (0.00, 0.25, 0.50, 0.75, or 1.00) based on the number of letters that are *correct* and *present*. This method of scoring might be useful to analyze the results of the L2 production test from another perspective.

REFERENCES

- Amano, S., & Kondo, T. (1999). Nihongo-no-Goitokusei Shinmitsudo [Lexical properties of Japanese: Word familiarity]. Sanseido.
- Amano, S., & Kondo, T. (2000). Nihongo-no-Goitokusei Hindo [Lexical properties of Japanese: Frequency]. Sanseido.
- 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.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–12*.
- 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). MIT Press.
- Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). MIT Press.
- Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, 128, 241–252.
- Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, 13, 826–830.
- Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. <https://www.english-corpora.org/coca>
- De Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, 50, 1–56.
- De Rosario-Martinez, H. (2015). *phia: Post-hoc interaction analysis. R package version 0.2–1*.
- Dijkstra, T., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5, 175–197.
- Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 22, 657–679.

- Gerard, P. D., Smith, D. R., & Weerakkody, G. (1998). Limits of retrospective power analysis. *The Journal of Wildlife Management*, 62, 801–807.
- González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, 41, 481–505.
- Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7, 493–498.
- Griffin, G., & Harley, T. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, 17, 443–460.
- Karpicke, J. D., & Roediger, H. L. (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 704–719.
- Karpicke, J. D., & Roediger, H. L. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57, 151–162.
- Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319, 966–968.
- Kroll, J. F., Michael, E., Tokowicz, N., & Dufour, R. (2002). The development of lexical fluency in a second language. *Second Language Research*, 18, 137–171.
- Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33, 149–174.
- Kroll, J. F., & Sunderman, G. (2003). Cognitive processes in second language learners and bilinguals: The development of lexical and conceptual representations. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 104–129). Blackwell.
- Kusanagi, K., Mizumoto, A., & Takeuchi, O. (2015). Nihon no Gaikokugo kyouiku kenkyu ni okeru koukaryou, kenteiryoku, hyouhonsaizu: Language Education & Technology keisai ronbun wo taisyou ni shita jirei bunseki [Reviewing effect sizes, statistical powers, and sample sizes of foreign language teaching research in Japan: A case analysis of Language Education & Technology]. *Language Education & Technology Journal*, 52, 105–131.
- Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22, 1–26.
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16, 33–51.
- 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.
- Lenth, R. V. (2019). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.4.1.
- Meara, P., & Miralpeix, I. (2016). *Tools for researching vocabulary*. Multilingual Matters.
- Mizumoto, A., & Takeuchi, O. (2010). Koukaryou to kenteiryoku bunseki nyuumon: Toukeiteki kentei wo tadashiku tsukau tame ni [Introduction to effect size and power analysis: For an appropriate use of the statistical test]. *Language Education & Technology, Kansai Chapter, Methodology Special Interest Group*, 47–73.
- Nakagawa, S., & Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica*, 7, 103–108.
- Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, 39, 653–679.
- Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, 41, 287–311.
- 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, 523–552.
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
- 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.
- 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.

- R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org>.
- Roediger, H. L., & Guynn, M. J. (1996). Retrieval processes. In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (2nd ed., pp. 197–236). Academic Press.
- Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.
- Royer, J. M. (1973). Memory effects for test-like-events during acquisition of foreign language vocabulary. *Psychological Reports, 32*, 195–198.
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*, 55–88.
- Schneider, V., Healy, L., & Bourne, E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46*, 419–440.
- Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal, 103*, 713–720.
- Tagashira, K., Kida, S., & Hoshino, Y. (2010). Hot or gelid? The influence of L1 translation familiarity on the interference effects in foreign language vocabulary learning. *System, 38*, 412–421.
- Tamura, Y., Fukuta, J., Nishimura, Y., Harada, Y., Hara, K., & Kato, D. (2019). Japanese EFL learners' sentence processing of conceptual plurality: An analysis focusing on reciprocal verbs. *Applied Psycholinguistics, 40*, 59–91.
- 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. (2007). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research, 11*, 63–81.
- Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal, 40*, 360–376.
- Webb, S., & Chang, A. C. S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition, 37*, 651–675.
- Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning, 71*, 487–536.

APPENDIX A

POWER ANALYSIS

The importance of statistical power analysis has been emphasized in previous studies (e.g., Gerard et al., 1998; Kusanagi et al., 2015). Statistical power is defined as the probability that the test will reject a null hypothesis when the null hypothesis is false and yield a significant result when the research hypothesis is true (e.g., Kusanagi et al., 2015; Nakagawa & Foster, 2004). Generally, increasing the sample size is likely to lead to statistically significant results; by contrast, when the sample size is too small, it leads to small statistical power (Mizumoto & Takeuchi, 2010). Therefore, deciding sample sizes subjectively is not appropriate for statistical analysis, and a statistical power analysis is needed to decide the appropriate sample size based on the aimed-statistical power (Kusanagi et al., 2015). In this study, the power analysis was conducted with the *SIMR* package using the data from the pilot study (Green & MacLeod, 2016). The pilot study was conducted with 12 participants who were comparable to the participants in the main study. The results of the power analysis are as follows.

For the first model, as the main effects of Learning Condition and Test Type were statistically significant in the pilot study, a retrospective observed power calculation,

where the target effect size comes from the observed data, was conducted with those variables (Green & MacLeod, 2016). However, the effect of interaction was not significant in the pilot test; thus, the effect size was changed from -0.50 (observed effect size) to -0.60 (target effect size), which led to a significant effect. The power analysis showed that the minimum number of each factor to reach $\alpha = .05$, power = 80% were as follows: Learning Condition: 87.00%, 95% CI [78.80, 92.89] for 19 participants; Test Type: 99.00%, [94.55, 99.97] for 7 participants; and the interaction of Learning Condition and Test Type: 82.00%, [73.05, 88.97] for 24 participants.

For the second and third models, the retrospective observed power calculation was conducted with those variables, as the main effects of Learning Condition and the interaction of Learning Condition and Vocabulary Size were statistically significant in the pilot study (Green & MacLeod, 2016). However, the main effect of Vocabulary Size was not significant in the pilot test; thus, the effect size was changed from -0.13 (observed effect size) to -0.40 (target effect size), which led to a significant effect, before implementing the power analysis. The power analysis showed that the minimum numbers of each factor to achieve $\alpha = .05$, power = 80% were as follows: Learning Condition: 82.00%, 95% CI [73.05, 88.97] for 24 participants; Vocabulary Size (English proficiency): 80.00%, [70.82, 87.33] for 11 participants; and the interaction of Learning Condition and Vocabulary Size: 92.00%, [84.84, 96.48] for 11 participants.

The current study recruited 28 participants, which is slightly more than the number of participants suggested by the results of the power analysis: the highest minimum number of required participants was 24 for the interaction of Learning Condition and Test Type in the first model and 24 for the main effect of Learning Condition in the second and the third model. Therefore, it is estimated that the current statistical analysis adequately detected a significant difference of the data.

APPENDIX B

TARGET ITEMS

Vocabulary A

| English | Japanese | English | Japanese |
|-----------|----------|--------------|----------|
| azalea | ツツジ | tuberculosis | 結核 |
| berth | 寝台 | loach | ドジョウ |
| billow | 大波 | otter | カワウソ |
| bluff | 絶壁 | pail | バケツ |
| camphor | クスノキ | plateau | 高原 |
| cistern | 水槽 | rudder | 舵 |
| citadel | 砦(とりで) | plumage | 羽 |
| fracas | けんか | shoal | 浅瀬 |
| fuselage | 胴体 | strait | 海峡 |
| insurgent | 暴徒 | tympanum | 鼓膜 |

Vocabulary B

| English | Japanese | English | Japanese |
|-----------|----------|-----------|----------|
| alcove | 床の間 | pall | 棺 |
| parable | 比喩 | porcupine | ヤマアラシ |
| badger | 穴熊 | potassium | カリウム |
| scowl | しかめっ面 | quail | ウズラ |
| diaphragm | 横隔膜 | ravine | 渓谷 |
| estuary | 河口 | rectum | 直腸 |
| mane | たてがみ | sentry | 見張り |
| levee | 堤防 | toupee | かつら |
| mirth | 歡喜 | ore | 鉱石 |
| kiln | 炉 | weasel | イタチ |