Effects of massing and spacing on the learning of semantically related and unrelated words

Tatsuya Nakata

Faculty of Foreign Language Studies

Kansai University, Japan

Yuichi Suzuki

Faculty of Foreign Languages

Kanagawa University, Japan

Abstract

Although researchers argue that studying semantically related words simultaneously (semantic clustering) inhibits vocabulary acquisition, recent studies have yielded inconsistent results. This study examined the effects of semantic clustering while addressing the limitations of previous studies (e.g., confounding of semantic relatedness with other lexical variables). Furthermore, the study investigated the effects of spacing because spacing might facilitate the learning of semantically related items by alleviating interference. In this study, 133 Japanese university students studied 48 English-Japanese word pairs under two conditions: massed and spaced. Half of the words were semantically related to each other while the other half were not. Although there were no significant differences between semantically related and unrelated items in posttest scores, semantically related items led to more interference errors than unrelated items. Furthermore, contrary to the authors' hypothesis that spacing is particularly beneficial for semantically related items, spacing benefited unrelated items more than it did related items.

In second language (L2) classrooms, learners are often exposed to words that are semantically related to each other, such as coordinates (e.g., *apple, orange*), synonyms (e.g., *fast, rapid*), or antonyms (e.g., *increase, decrease*). In many textbooks, words related to a particular topic or situation (e.g., colors, animals) are usually introduced together (Bolger & Zapata, 2011; Nation & Webb, 2011). Introducing semantically related words simultaneously, sometimes referred to as *semantic clustering*, is not only popular but also considered helpful for learning (Bolger & Zapata, 2011; Erten & Tekin, 2008; Folse, 2004; Waring, 1997).

Vocabulary researchers, however, argue against semantic clustering based on the interference theory (e.g., Baddeley, 1997). According to this theory, semantic clustering hinders vocabulary learning because it causes interference between similar meanings of related words. For instance, when learners are taught the Japanese words *neko* (cat) and *inu* (dog) at the same time, they might have difficulty remembering which of the two words means *cat* (cross-association; Schmitt, 2007). Advice against semantic clustering can be found in many books and journal articles authored by vocabulary researchers (e.g., Barcroft, 2015; Folse, 2004; Nation, 2000, 2013; Nation & Webb, 2011; Schmitt, 2007, 2010). However, although earlier studies supported the negative impact of semantic clustering (Tinkham, 1993, 1997; Waring, 1997), recent studies have yielded mixed results. Ishii (2015), for instance, found no significant difference between semantically related and unrelated items in the number of correct responses on a translation posttest. Four studies found that semantic clustering resulted in significantly higher posttest scores (Hashemi & Gowdasiaei, 2005; Hoshino, 2010; Schneider, Healy, & Bourne, 1998, 2002). These conflicting findings warrant further research.

Considering the prevalence of semantic clustering in classrooms, it is worth investigating how to facilitate the learning of semantically related words. One possible way might be to temporally space the opportunities for studying them (Folse, 2004; Nation, 2000; Schmitt, 2007). For instance, when learners are taught two semantically related words simultaneously (massed learning), they may have difficulty distinguishing between them. In contrast, introducing a second word only after the first one is mostly known (spaced learning) might result in less interference, thereby facilitating the learning of the two related words. This suggests that semantically related words might benefit more from spacing than unrelated words. Previous studies, however, have not

examined whether spacing benefits semantically related and unrelated words differently.

The present study has two objectives: First, given the inconsistent results of previous studies, this study aims to examine whether semantic clustering affects the amount of vocabulary learned as well as the amount of interference caused. Second, this study examines whether spacing has differential effects on the learning of semantically related and unrelated words. The findings of this study might allow us to identify how to effectively teach and learn semantically related words.

Literature Review

Effects of Semantic Clustering on L2 Vocabulary Learning

Vocabulary is commonly introduced in two ways: thematic and semantic clustering. In thematic clustering, vocabulary items related to a theme are presented together. For instance, for the theme of vacation, words such as *island, sunny, swim,* or *hotel* might be introduced. In semantic clustering, words that are semantically related to each other, such as coordinates, synonyms, or antonyms, are introduced together. Semantic clustering is common in L2 classrooms perhaps because it is thought to facilitate learning for at least three reasons. First, semantic clustering reflects how vocabulary is stored in the mental lexicon (Nation, 2000); semantically related words are interconnected and form a network in the mental lexicon, as suggested by research on word associations (Meara, 2009). Therefore, presenting semantically related words together is considered a natural and, hence, effective way to introduce vocabulary. Second, semantic clustering also might enhance learning by introducing difficulty (Finkbeiner & Nicol, 2003). As noted earlier, semantic clustering often causes interference between semantically related words, which increases difficulty. According to the desirable difficulty framework (Bjork, 1999), a condition that makes learning difficult can be effective over time. As a result, semantic clustering, which causes interference and slows down initial learning, might facilitate learning in the long term. Third, due to the difficulty caused by interference, semantically related items may receive more attention, effort, or engagement than unrelated words, which potentially results in better learning (Finkbeiner & Nicol, 2003; Hashemi & Gowdasiaei, 2005; Tagashira, Kida, & Hoshino, 2010).

Despite its potential pedagogical values, some researchers argue against semantic clustering because it might inhibit learning by causing interference between related words (e.g., Barcroft, 2015; Folse, 2004; Nation, 2000; Schmitt, 2007). Six studies found negative effects of semantic

clustering, supporting their claim. In Tinkham (1993), for instance, 20 English-speaking participants studied artificial words paired with English translations. Half of the items were semantically related to each other (e.g., *apple, apricot*), while the other half were semantically unrelated (e.g., *mouse, sky*). Participants heard an English translation and were asked to say the corresponding pseudoword. The treatment continued until the participants correctly answered all items in both sets. Tinkham (1993) found that participants required significantly more trials to learn semantically related sets than unrelated sets. Negative effects of semantic clustering were also observed in five other studies (Erten & Tekin, 2008; Finkbeiner & Nicol, 2003; Tinkham, 1997; Waring, 1997; Wilcox & Medina, 2013). Two studies showed negative effects of semantic clustering under limited conditions. Higa (1963) found negative effects of two types of semantic clustering (i.e., synonyms and free associations), but not with other four types (i.e., antonyms, coordinates, partial-response-identity, and connotations). Papathanasiou (2009) found the negative impact of semantic clustering with beginner adults, but not with intermediate children.

In contrast, Ishii (2015) found no significant differences between the semantically related and unrelated sets in the number of correct responses on a translation posttest. Four studies found positive effects of semantic clustering (Hashemi & Gowdasiaei, 2005; Hoshino, 2010; Schneider et al., 1998, 2002). In Schneider et al. (1998), English-speaking college students studied 25 French words from five semantic categories (body parts, vehicles, kitchen utensils, food, clothes). Participants were assigned to two conditions: blocked (semantically related) and mixed (unrelated). Contrary to the findings of earlier research (Higa, 1963; Tinkham, 1993, 1997; Waring, 1997), Schneider et al. (Experiment 2) found that the blocked condition produced more correct translations than the mixed condition during the initial learning phase. In a follow-up study, Schneider et al. (2002) found that the blocked (related) condition led to more correct translations than the mixed (unrelated) condition on an immediate posttest. Similarly, Hashemi and Gowdasiaei (2005) and Hoshino (2010) found the positive effects of semantic clustering on delayed posttests.

In summary, among 13 empirical studies, six studies found that semantic clustering inhibits vocabulary learning (Erten & Tekin, 2008; Finkbeiner & Nicol, 2003; Tinkham, 1993, 1997; Waring, 1997; Wilcox & Medina, 2013), and two studies found negative effects of semantic clustering under limited conditions (Higa, 1963; Papathanasiou, 2009). One study found no effect

(Ishii, 2015), and four studies found positive effects of semantic clustering (Hashemi & Gowdasiaei, 2005; Hoshino, 2010; Schneider et al., 1998, 2002). These findings suggest that the negative effects of semantic clustering might not be as robust as many vocabulary researchers claim.

The inconsistent results of the existing research might be partially due to methodological differences. Existing studies differ in how vocabulary knowledge was measured. Among 13 previous studies, five used receptive translation (i.e., translate from L2 to L1), three used productive translation (i.e., translate from L1 to L2), and three used both receptive and productive translation. Hashemi and Gowdasiaei (2005) used the Vocabulary Knowledge Scale (Wesche & Paribakht, 1996), and Erten and Tekin (2008) used a picture-word matching task. Previous studies also differ in other methodological factors such as learning stimuli (e.g., Spanish-English: Wilcox & Medina, 2013; pseudoword-Japanese: Ishii, 2015), age of participants (e.g., fourth graders: Erten & Tekin, 2008; university students: Hoshino, 2010), L2 proficiency of participants (e.g., novice: Wilcox & Medina, 2013; beginner / intermediate: Papathanasiou, 2009), and duration of the treatment (e.g., 20 minutes: Wilcox & Medina, 2013; 3-4 days: Hoshino, 2010). These methodological differences could partially be responsible for the inconsistent results of earlier studies.

Previous experiments have not only produced inconsistent results but also suffered from at least two limitations. One limitation is that some studies only examined trials to criterion (i.e., number of trials needed to reach the criterion of correct recalls) during the learning phase and did not administer a posttest (Higa, 1963; Tinkham, 1993, 1997; Waring, 1997). When examining the effects of semantic clustering, investigating retention on posttests is critical. This is because the desirable difficulty framework (Bjork, 1999) predicts that semantic clustering, which causes interference and slows down initial learning, facilitates learning in the long term. A meta-analysis of existing studies confirms the importance of investigating not only performance during the learning phase but also on posttests. When the results of previous studies are meta-analyzed, a synthetized effect size (Cohen's *d*) of semantically related sets over unrelated sets in the trials-to-criterion studies is 0.73 [0.41, 1.05] (the values inside the brackets indicate 95% confidence intervals). This indicates that semantically related items required more trials than unrelated items, producing a medium-sized effect ($d = 0.7$; Plonsky & Oswald, 2014). When learning is measured by posttests

administered after the learning phase, however, a synthetized effect size of semantic clustering is -0.24 [-0.71, 0.23]. This suggests that, although posttest scores were generally lower for related items, the effect size did not reach the criterion of a small effect (d = -0.40). The very small effect size and the 95% confidence interval that crosses zero suggest that, although semantic clustering might affect the initial rate of acquisition (i.e., trials to criterion), it may not necessarily influence subsequent retention as measured by posttests (see Appendix A in the online supplementary materials for further details of the meta-analysis).

Second, previous studies on semantic clustering are also limited in that they failed to control item difficulty. Earlier studies have attempted to examine the effects of semantic clustering using one of the two approaches. The first approach is to examine the effects of blocking and mixing (Finkbeiner & Nicol; 2003; Hashemi & Gowdasiaei, 2005; Schneider et al., 1998, 2002). In these studies, in a semantically related (blocked) condition, items from the same semantic category were studied sequentially, whereas in a semantically unrelated (mixed) condition, opportunities for studying semantically related items were distributed across the treatment. The second approach is to examine the effects of semantic clustering by comparing the learning of semantically related and unrelated items (e.g., Higa, 1963; Tinkham, 1993, 1997; Waring, 1997; Wilcox & Medina, 2013). Unlike in the first approach, where the same sets of target items were used for both the semantically related and unrelated conditions, studies employing the second approach used different sets of items for the semantically related and unrelated conditions. These studies assume that the semantically related and unrelated items are controlled for factors other than semantic relatedness that might affect learning. Otherwise, any difference between the related and unrelated sets cannot be confidently attributed to semantic relatedness.

Previous studies using the second approach, however, have often failed to control item difficulty between the semantically related and unrelated sets. Research suggests that a number of factors affect the learning burden of L2-L1 word pairs (e.g., Barcroft & Rott, 2010; Laufer, 2012; Schmitt, 2010). Factors related to L2 words include L2 word frequency (Lotto & de Groot, 1998), L2 word length (Ellis & Beaton, 1993), and pronounceability (de Groot, 2006; de Groot & van Hell, 2005; Ellis & Beaton, 1993). Factors related to L1 translation equivalents include the part of speech (Ellis & Beaton, 1993; Rodgers, 1969), L1 word frequency (de Groot, 2006; Lotto & de Groot,

1998), L1 word length (Ellis & Beaton, 1993), familiarity (Tagashira et al., 2010), and imageability (de Groot, 2006; de Groot & Keijzer, 2000; de Groot & van Hell, 2005; Ellis & Beaton, 1993). None of the previous studies comparing semantically related and unrelated sets have controlled all the lexical variables mentioned above. For instance, while most studies controlled L2 word length and part of speech, none of the studies controlled the pronounceability of L2 words, L1 word length, or L1 familiarity.

Some studies attempted to control item difficulty by using pseudowords, assigning different conditions to the same items for different participants (for instance, giving some participants a pseudoword as a semantically related item and giving other participants the same pseudoword as a semantically unrelated item). This eliminates the need to control L2-related factors because it can be assumed that the effects of the pseudowords are counterbalanced across participants. However, L1-related factors (i.e., frequency, length, familiarity, and imageability of L1 translation equivalents) still must be controlled. Although Higa (1963) and Tinkham (1997) controlled L1 frequency, other L1-related variables (L1 word length, familiarity, and imageability) were not controlled by any of the studies that used pseudowords (Higa, 1963; Ishii, 2015; Tinkham, 1993, 1997; Waring, 1997). The results of earlier studies, therefore, might be at least partly attributable to possible differences in item difficulty rather than semantic relatedness per se.

With the conflicting results and limitations of previous studies in mind, this study attempted to examine whether semantic clustering inhibits or facilitates vocabulary learning. This study expanded on previous studies by assessing long-term effects, giving posttests not only immediately but also 1 week after the treatment. To more rigorously examine the effects of semantic relatedness, this study also controlled semantically related and unrelated sets for lexical factors other than semantic relatedness that might affect learning.

Effects of Spacing on the Learning of Semantically Related Words

Considering the prevalence of semantic clustering in textbooks and classrooms, how can teachers facilitate the learning of semantically related words? One possible way might be to temporally space the opportunities for studying them (Folse, 2004; Nation, 2000; Schmitt, 2007). When discussing the effects of spacing, two concepts must be distinguished: spacing effect and lag effect (Rogers, 2017). The spacing effect refers to a phenomenon in which spaced learning (which

involves intervals between repetitions of a given item) yields superior retention as compared to massed learning (which does not involve any intervals). The lag effect, in contrast, is concerned with the question of whether long spacing facilitates learning better than short spacing. (The spacing effect and lag effect are collectively referred to as the distributed practice effect.) Studies have found that the lag effect is sensitive to changes in the posttest timing. Specifically, although long spacing tends to be effective when the posttest is given after a long delay, short spacing tends to be effective when the posttest is given after a short delay (e.g., Nakata, 2015; Nakata & Webb, 2016). The spacing effect and lag effect are found to affect L2 vocabulary learning (e.g., Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Nakata, 2015; Nakata & Webb, 2016). Bahrick et al. (1993), for instance, compared the effects of the following three spacing intervals: 14, 28, and 56 days. Learning was measured 1, 2, 3, and 5 years after learning. Bahrick et al. found that longer spacing resulted in better retention than shorter spacing.

Given the positive effects of spacing observed in vocabulary learning studies, spacing might also be expected to facilitate the learning of semantically related words. Spacing might be particularly effective for semantically related words because it might reduce interference. For instance, when teaching two semantically related words, interference might be alleviated by introducing a second word only after the first one is mostly known (Folse, 2004; Nation, 2000; Schmitt, 2007). Although the purpose of Bolger and Zapata's (2011) study was not to evaluate the effects of spacing, they compared the effects of two treatments that differed in the amounts of spacing for the learning of semantically related words. In their study, 66 participants read four short English stories. Thirty-two pseudowords from four semantic categories (animals, kitchen utensils, furniture, body parts) appeared throughout the stories. Bolger and Zapata found that learning was enhanced when items from a given semantic category were distributed across four stories (unrelated condition) rather than concentrated in one story (related condition). The findings demonstrate the value of spacing for the learning of semantically related words. However, because their treatment involved incidental vocabulary learning from context, it is not clear whether the advantage of the unrelated condition is attributable to the effects of spacing alone or to the combined effects of spacing and context. (Their research design, of course, should not be considered a limitation because the purpose of Bolger and Zapata's study was to evaluate the effects of context, not

spacing.)

This study expands on Bolger and Zapata (2011) by examining the effects of spacing on semantic clustering in a paired-associate format, where the target items are studied in a decontextualized format. This format allows the effects of spacing and context to be separated. Because most studies on semantic clustering employed paired-associate learning (e.g., Schneider et al., 1998, 2002; Tinkham, 1993, 1997; Waring, 1997), the use of a paired-associate format also allows the results of this study to be directly compared with those of earlier research. Furthermore, this study investigates the effects of spacing not only on semantically related but also on unrelated items. By examining whether spacing has differential effects on semantically related and unrelated words, this study attempts to test the view that spacing is particularly beneficial for semantically related words because it reduces interference.

Research Questions and Hypotheses

The current study addresses the following two research questions (RQs):

RQ1: Does semantic clustering facilitate or hinder L2 vocabulary learning?

RQ2: Does spacing have differential effects on the learning of semantically related and unrelated words?

The first research question is concerned with the effects of semantic clustering. Previous studies have yielded mixed results regarding the effects of semantic clustering on vocabulary learning. This study differs from previous studies in two important respects. First, unlike some previous studies (Higa, 1963; Tinkham, 1993, 1997; Waring, 1997), this study examined the effects of semantic clustering not only on performance during the learning phase but also on performance on posttests administered immediately and 1 week after learning. This is critical because the desirable difficulty framework (Bjork, 1999) predicts that semantic clustering, which causes interference and slows down initial learning, facilitates learning in the long term. The results of the meta-analysis, which found significant negative effects of semantic clustering on learning-phase performance but not on posttest scores (see Appendix A in the online supplementary material for details), also support the importance of investigating retention as well as the initial rate of acquisition. Second, previous studies comparing semantically related and unrelated sets were limited in that two sets of items were not tightly controlled for variables that might affect item

difficulty such as L1 word frequency, familiarity, or imageability. The results of these earlier studies, therefore, might be at least partly attributed to possible differences in item difficulty rather than semantic relatedness per se. With this in mind, the present study controlled lexical variables that were found to affect item difficulty. This allows for a more rigorous investigation of the effects of semantic relatedness.

The second research question asks whether the effects of spacing interact with the semantic relatedness of lexical items. By examining whether semantically related and unrelated words benefit differently from spacing, this study tests the view that spacing enhances the learning of semantically related words by reducing interference. Furthermore, unlike Bolger and Zapata (2011), this study examines the effects of spacing on semantic clustering in a paired-associate format to isolate the effects of spacing and context.

The following two hypotheses were formed regarding the above two research questions:

Hypothesis 1: Semantic clustering hinders the retention of L2 vocabulary.

Hypothesis 2: Spacing more greatly facilitates the learning of semantically related words than unrelated words.

Hypothesis 1 is based on the interference theory, according to which semantic clustering inhibits learning because it causes cross-associations between related words. This hypothesis is incongruent with the results of our meta-analysis of existing research, which indicates that, although semantic clustering might slow down the initial rate of acquisition, it might not affect subsequent retention as measured by posttests (see Appendix A in the online supplementary material for details). However, the authors predict negative effects of semantic clustering, based on the argument against semantic clustering made by many vocabulary researchers. Hypothesis 2 predicts that spacing more greatly facilitates the learning of semantically related words than unrelated words. Although spacing can be expected to benefit both semantically related and unrelated words (distributed practice effect), it might also reduce interference for semantically related items (Folse, 2004; Nation, 2000; Schmitt, 2007), resulting in larger positive effects for related items.

Method

Participants

The participants were 133 Japanese students from two universities in Japan. They had been

studying English for at least 6 years. Prior to the experiment, the participants took the paired-associate section of LABJT (Language Aptitude Battery for the Japanese; Sasaki, 1993), which is a Japanese translation of Part V of the MLAT (Modern Language Aptitude Test) and measures learners' ability to memorize vocabulary in a paired-associate format. To control the possible effects of vocabulary-learning aptitude, scores on the paired-associate section of LABJT were used as a covariate (see the Results section). Eleven participants who did not have a score for the LABJT test were excluded from analysis. The participants were randomly divided into two groups, massed (n = 66) and spaced (n = 56). There was no statistically significant difference in the average LABJT scores between the massed and spaced groups, t (120) = -0.71, p = .48.

Materials

The target items were 48 low-frequency English words paired with their Japanese translation equivalents (e.g., *otter*-カワウソ). Half of the pairs were semantically related and the other half were semantically unrelated items. The 24 semantically related items consisted of four sets of six coordinates; Set 1: *baboon, badger, otter, porcupine, raccoon,* and *weasel* (mammals); Set 2: *diaphragm, intestine, placenta, rectum, tympanum,* and *womb* (organs); Set 3: *bluff, estuary, plateau, ravine, shoal,* and *strait* (geographical features); and Set 4: *azalea, camellia, camphor, cedar, magnolia,* and *willow* (plants). Coordinates, rather than synonyms or antonyms, were used as semantically related items because most previous studies using coordinates found significant effects of semantic clustering (e.g., Erten & Tekin, 2008; Finkbeiner & Nicol, 2003; Tinkham, 1993, 1997; Waring, 1997); the use of coordinates would reveal the effects of semantic relatedness. The 24 semantically unrelated words were also divided into four sets; Set 5: *alloy, apparition, kerosene, kiln, plumage,* and *rudder*; Set 6: *cistern, insurgent, pall, parable, sardine,* and *venom*; Set 7: *alcove, pail, pigment, potassium, relic,* and *toupee*; Set 8: *berth, fuselage, ointment, ore, sentry,* and *tuberculosis*. None of the target English words were loanwords or cognates in Japanese, the participants' L1.

The semantically related and unrelated words were controlled for L2-related variables (L2 frequency, L2 word length, pronounceability) and L1-related variables (part of speech, L1 frequency, L1 word length, familiarity, imageability) that might affect the learning burden (also see Literature Review). First, L2 word frequency was operationally defined as frequency levels of the

English target words in BNC/COCA (Nation, 2012) and BNC frequency lists (Nation, 2006). Second, L2 word length was operationalized as the number of syllables and letters. Third, the pronounceability of L2 words was operationalized as the average biphoneme, triphoneme, and positional probability of the English target words calculated using the Irvine Phonotactic Online Dictionary (Vaden, Hickok, & Halpin, 2009). Fourth, because all of the target words were nouns, the part of speech was not a factor. Fifth, L1 word frequency, which is an index of conceptual frequency (Ellis & Beaton, 1993), was derived from Amano and Kondo (1999). Sixth, L1 word length was operationalized as the number of moras and letters in the Japanese translation equivalents. Lastly, the familiarity and imageability of the Japanese translation equivalents derived from Amano and Kondo (1999) were also controlled. Because the semantically related and unrelated words were matched for the above variables, it was assumed that the two types of words were controlled for lexical factors other than semantic relatedness that might affect learning (see Appendix B in the online supplementary materials for details).

To ensure that the L1 translations of the target words were familiar to the participants, the familiarity ratings of the Japanese translation equivalents derived from Amano and Kondo (1999) were examined. The average familiarity rating of the Japanese equivalents of the 48 target words was 5.08 (*SD* = 0.77) on a 7-point scale, where 1 means *unfamiliar* and 7 means *familiar*. The word with the lowest familiarity rating was the Japanese equivalent of *baboon*, which had a familiarity rating of 2.06. This low familiarity rating possibly was due to this word being presented in the familiarity survey using its uncommon orthographic form (マント狒狒) instead of its more standard form (マントヒヒ). With the exception of *baboon*, all target words had an L1 familiarity rating of 3.94 or higher. As a result, it might be reasonable to assume that the participants were familiar with the L1 translations of the target words. Please note also that because the familiarity ratings of the Japanese translation equivalents were controlled for the semantically related and unrelated sets (see above), it can be assumed that the L1 familiarity did not have a major effect on the results of this study.

Procedure

The study was conducted during two regular classes. Each student had access to a computer, and the students studied and were tested individually with computer software developed by the first

author. There were two sessions. The first session consisted of the pretest, learning phase, filler task, and immediate posttest. In the second session, which was conducted 1 week after the first session, the delayed posttest was administered. At the outset of the first session, the participants received explanations about the study and practiced using the software with four sample words. After the practice, the pretest was given. In the pretest, participants were presented with 48 English target words one by one, and asked to type the corresponding Japanese translations. The target words from the eight item sets were mixed and presented in a pseudo-random order.

Following the pretest, the participants studied 48 target words using computer software in a paired-associate format. In both massed and spaced groups, all 48 target words were encountered four times throughout the learning phase, and each target word was studied separately, resulting in a total of 192 trials (48 words × 4 times). The two groups, however, differed in the intervals at which the target words were repeated. In the massed group, six items from a given item set were studied four times sequentially. For instance, six items from one set (e.g., Set 1: *baboon, badger, otter, porcupine, raccoon, weasel*) were presented in four sequential blocks, with all items randomized within each block, and then six items from another set (e.g., Set 5: *alloy, apparition, kerosene, kiln, plumage, rudder*) were presented in four sequential blocks. To ensure that the trials for semantically related and unrelated items were distributed roughly equally across the learning phase, semantically related and unrelated sets were alternated throughout the learning phase (e.g., Set 1 × 4 [related], Set 5 × 4 [unrelated], Set 2 × 4 [related], Set 6 × 4 [unrelated] …). To reduce any order effect, half of the participants studied a semantically related set first and the other half studied an unrelated set first.

In the spaced group, trials for a given set (e.g., Set 1) were separated by trials for the other seven sets (e.g., Set 2-8). The 48 items were divided into two blocks of 24 items, and each block consisted of three items from each of the eight sets (3 items × 8 sets = 24). The same blocks were maintained throughout the learning phase. The items were studied in these two blocks of 24 items in the following order: Block 1 × 2, Block 2 × 2, Block 1 × 2, and Block 2 × 2. A block size of 24, rather than 48, was used because the pilot study showed that studying in a block of 48 items was too challenging for most participants, leading to ineffective learning and decreased motivation. The item order in the spaced group followed the same rules used for the massed group: (a) to ensure that

semantically related and unrelated items were distributed roughly equally across the learning phase, semantically related and unrelated items alternated throughout the learning phase; (b) to reduce any order effect, half of the participants were given a related item as the first item in the first block, while the other half were given an unrelated item as the first item in the first block; and (c) to further reduce any order effect, the item order within each block was randomized for each repetition.

In both the massed and spaced groups, the participants' first encounter with each item was the initial presentation, where an English target word and its Japanese translation (e.g., *otter*-カワウソ) were presented for 8 seconds. In their second, third, and fourth encounters, the target items were practiced in a receptive translation format. In other words, the participants were presented with an English target word and asked to type in the corresponding Japanese translation (e.g., *otter*-_____?). After each response, the correct answer (e.g., *otter*-カワウソ) was provided as feedback for 5 seconds. Receptive (i.e., translate from L2 to L1) rather than productive translation (i.e., translate from L1 to L2) was used for two reasons. First, Schneider and colleagues (2002) used both receptive and productive translation during the learning phase and found larger effects of semantic clustering for receptive translation. They argue that this was possibly because responding in L1 is more likely to activate conceptual representations than responding in L2, because L2 lexical representations are often only weakly linked to conceptual representations. The use of receptive translation, therefore, might reveal the effects of semantic relatedness. Second, unlike receptive translation, productive translation requires the productive knowledge of orthography as well as the knowledge of form-meaning connections. Semantic clustering is expected to have larger effects on the knowledge of form-meaning connections than that of orthography. For instance, when learning *cat* and *dog* simultaneously, the concept of *dog* may be erroneously associated with *cat,* while the concept of *cat* may be associated with *dog* due to interference, thus affecting form-meaning connections. However, learning *cat* and *dog* simultaneously perhaps may not significantly affect the learning of how to spell these two words. Because the effects of semantic clustering may be more pronounced in the learning of form-meaning connections than that of orthography, the use of productive translation might obscure possible effects of semantic clustering. Receptive translation, therefore, was used as the treatment task and dependent measure. Please note also that receptive

translation has been used as a treatment task in existing studies on semantic clustering (Higa, 1963; Papathanasiou, 2009; Schneider et al., 1998, 2002; Tinkham, 1997).

The translation task was self-paced, and participants were allowed to take as much time as they needed to type a response. No time limit was set for the translation task for three reasons. First, in normal learning conditions, it is common for learners to pace practice by themselves (Nation & Webb, 2011). Self-pacing of translation tasks thus reflects authentic learning and increases ecological validity. Second, the amount of time needed for translation tasks might vary, depending on the participants or target items. Self-pacing of translation tasks might enable learners to learn effectively regardless of possible individual or item differences. Third, earlier studies on semantic clustering (e.g., Finkbeiner & Nicol, 2003; Hoshino, 2010; Papathanasiou, 2009; Schneider et al., 1998, 2002) typically used self-pacing of treatment. The self-paced practice, therefore, might enable us to better compare the results of the present and previous studies. Because the translation task was self-paced, time-on-task was not controlled. The translation latency, therefore, was modelled as a covariate in the data analysis (see Results).

Upon completing the learning phase, the participants answered ten 2-digit additions (e.g., 26 + 65 = ?) as a filler task. This task was included to neutralize the order effect. Subsequently, the participants took the immediate posttest. Apart from the randomized item order, the posttest was identical to the pretest. After the immediate posttest, participants were asked to estimate how many target words out of 48 they would remember 1 week later (judgements of learning; e.g., Kornell, 2009). One week after the learning phase, an unannounced delayed posttest was administered. Apart from the randomized item order, the delayed posttest was identical to the pretest and immediate posttest.

Scoring and Data Analysis

To ensure consistency in scoring, the responses on the pretest, posttest, and during the learning phase were first scored by computer software based on answer keys compiled by the authors. Responses that were scored as incorrect by the computer program were manually checked by the authors and a research assistant. Six participants (four participants from the massed and two participants from the spaced groups) who scored zero for both semantically related and unrelated items on at least one of the following were excluded from the analysis: the second retrieval attempt

during the learning phase, third retrieval attempt during the learning phase, and immediate posttest. The remaining participants consisted of 62 students from the massed group and 54 from the spaced group.

To determine if semantic relatedness caused interference, within-set errors were also analyzed. For semantically related sets, when participants provided an incorrect response from the same semantic category (within-category errors), this was categorized as a within-set error. For instance, when participants produced the Japanese translation of *raccoon* (アライグマ) when asked to translate *weasel* (イタチ), it was categorized as a within-set error because both items belong to the semantic category of mammals (Set 1). For semantically unrelated sets, in the massed group, the within-set error was defined as a within-block error. In other words, in the massed group, the unrelated items were studied in a block of six items from a given item set. As a result, when the participants produced the Japanese translation of one of the other five items from the same unrelated set, this was regarded as a within-set error. The frequency of within-set errors was not calculated for the unrelated items in the spaced group. This is because in the spaced group, the target items were studied in a block of 24 items as opposed to six items in the massed group. This means that, for a given word, there are 23 words that could be classified as a within-set error in the spaced group, whereas there are only five words that could be classified as a within-set error in the massed group. Because the comparison of within-set errors for unrelated items could be misleading, the within-set errors for the unrelated items were not calculated for the spaced group. See Figure 1 for a diagram illustrating how within-set errors were operationalized in the massed and spaced groups.

[Insert Figure 1 around here]

Results

Learning-Phase Performance

During the learning phase, the participants were presented with the English target word and asked to provide the corresponding Japanese translation (receptive translation). Figure 2 shows the mean translation accuracy rates (%) during the learning phase for the massed and spaced groups (see Appendix C in the online supplementary materials for detailed descriptive statistics). For instance, Figure 2 shows that, on average, the massed group correctly translated 54.77%, 74.36%,

and 83.23% of the items for the first, second, and third retrieval attempts, respectively. Since the translation task was self-paced, the translation latency differed between the two groups. The average translation latency per trial was 6.81 seconds (*SD* = 1.57) for the massed group and 6.47 seconds (*SD* = 1.89) for the spaced group. To account for individual differences in translation latency, the translation latency was modelled as a covariate in the following analyses. Figure 3 shows separate translation accuracy rates for semantically related and unrelated words (see Appendix C for detailed descriptive statistics). When collapsed across the three retrieval attempts, the related and unrelated items resulted in similar translation accuracy for both the massed (related = 69.38%; unrelated = 72.20%) and spaced groups (related = 36.83%; unrelated = 38.14%).

[Insert Figures 2 and 3 around here]

To determine if spacing and semantic relatedness affected learning-phase performance, the translation accuracy rates during learning were analyzed using a logistic mixed-effects model. The dependent variable was a binary response (correct / incorrect). Fixed-effect predictors were group (massed vs. spaced) and semantic relatedness (related vs. unrelated). The retrieval position (retrieval 1, 2, and 3) was also included as a fixed-effect factor. To control for possible differences in translation latency, translation latency was added as a covariate in the model. The scores on the paired-associate section of LABJT (see Participants section) were also included as a covariate. Participants and items were treated as random effects.[1] The effect sizes were interpreted using the following criteria (Plonsky & Oswald, 2014): small ($d = 0.4$), medium ($d = 0.7$), and large ($d = 1.0$). The mixed-effects logit model revealed a significant fixed effect of group, $z = -6.25$, $p < .001$. No significant effect of relatedness was found, $z = -0.62$, $p = .54$. The interaction between group and relatedness also was not significant, $z = 0.53$, $p = .60$ (see Appendix D in the online supplementary materials for further details about the model). The findings suggest that the massed group significantly outperformed the spaced group during learning regardless of the relatedness of items (massed: related = 69.38%, unrelated = 72.20%; spaced: related =36.83%, unrelated = 38.14%).

To determine if semantic relatedness caused interference, the proportion of within-set errors during the learning phase was also analyzed. When collapsed across the three retrieval attempts, for the massed group, the proportion of within-set errors was 8.22% (*SD* = 7.80%) for the related items and 2.69% (*SD* = 3.41%) for the unrelated items. In the spaced group, for the related items, 2.73%

(*SD* = 2.82%) of the responses were within-set errors (see Appendix E in the online supplementary materials for detailed descriptive statistics). The proportion of within-set errors was analyzed using mixed-effects logit models. Two models were constructed: Model A compared the within-set error rates of related items between the massed and spaced groups, and Model B compared the within-set error rates between related and unrelated items in the massed group (see Appendix F in the online supplementary materials for full results). Model A revealed that the fixed effect of group was significant, $z = -5.39$, $p < .001$, $d = 0.85$ [0.46, 1.22] (the values inside the brackets indicate 95% confidence intervals). The results show that, for the related items, the spaced group produced significantly fewer within-set errors than did the massed group, demonstrating that spacing might potentially reduce interference among related items. According to the results of Model B, the massed group produced significantly more within-set errors for the related words than for the unrelated words, $z = 2.71$, $p = .01$, $d = 0.86$. This suggests that, in the massed group, semantically related words caused more interference than unrelated words during learning, which replicated the interference effect observed in previous studies (e.g., Tinkham, 1993, 1997; Waring, 1997).

Posttest Performance

Figures 4-6 illustrate the mean translation accuracy (%) on the posttests (see Appendix C in the online supplementary materials for detailed descriptive statistics). The translation accuracy rates on the immediate and delayed posttests were analyzed using separate mixed logit models. All fixed and random effects were identical to those included for the analysis of learning-phase translation accuracy, except that the retrieval position during the learning phase (retrieval 1, 2, and 3) was not included as a fixed-effect factor. Table 1 presents the results of the models. The fixed effect of relatedness, by itself, was not significant in either the immediate or delayed posttest (Figure 5; for the immediate posttest, related = 53.70%, unrelated = 56.49%; for the delayed posttest, related = 21.95%, unrelated = 22.54%). The findings suggest that semantic relatedness neither inhibited nor facilitated retention when collapsed across the massed and spaced groups.

[Insert Figures 4-6 around here]

For the model with the immediate posttest, the fixed effect of group was significant ($p = .02$), suggesting that the spaced group significantly outperformed the massed group when the semantically related and unrelated items were combined. The interaction between group and

relatedness was also statistically significant (p = .02). A post-hoc comparison showed that, on the immediate posttest, the spaced group significantly outperformed the massed group for the unrelated items (z = -2.31, p = .02, d = 0.35 [-0.02, 0.72]) but not for the related items (z = -0.87, p = .39, d = 0.12 [-0.25, 0.48]). In addition, no statistically significant difference was found between the semantically related and unrelated words for both the massed and spaced groups (massed: z = -0.01, p = .99, d = 0.00; spaced: z = 1.85, p = .06, d = 0.52). This suggests that semantic relatedness did not significantly affect the immediate posttest scores.

[Insert Table 1 around here]

The model for the delayed posttest also revealed a significant effect for group (p < .001). This indicates that the spaced group significantly outperformed the massed group on the delayed posttest when the semantically related and unrelated items were combined. The interaction between group and relatedness was also significant (p < .001). A post hoc comparison showed that (a) on the delayed posttest, the spaced group significantly outperformed the massed group for both semantically related and unrelated items (related: z = -2.86, p = .004; unrelated: z = -4.65, p < .001), and (b) whereas a medium effect size was found between the massed and spaced groups for the unrelated words (d = 0.72 [0.34, 1.09]), only a small effect size was observed for the related words (d = 0.38 [0.01, 0.75]). The findings suggest that, although spacing led to superior long-term retention compared with massing regardless of semantic relatedness, the advantage of spacing was more pronounced for the unrelated words (massed: 15.99%, spaced: 29.09%) than for the related words (massed: 17.74%, spaced: 24.15%). A post hoc comparison also found no statistically significant difference between the semantically related and unrelated words in delayed posttest scores for both the massed and spaced groups (massed: z = -0.69, p = .49, d = -0.15; spaced: z = 1.66, p = .10, d = 0.38). The findings demonstrate that semantic relatedness neither hindered nor facilitated long-term retention.

To determine if the interference observed during the learning phase persisted until the posttest, the proportions of within-set errors on the posttests were analyzed. Figure 7 illustrates the within-set error rates on the posttests (see Appendix E in the online supplementary materials for detailed descriptive statistics). As was done for the analysis of within-set error rates during the learning phase, two mixed logit models were constructed to analyze within-set error rates in the

immediate and delayed posttests: Model A compared the within-set error rates of related items between the massed and spaced groups, and Model B compared the within-set error rates between related and unrelated items in the massed group. Table 2 presents the results of the models for the immediate and delayed posttests.

[Insert Table 2 and Figure 7 around here]

Results of Model A demonstrated that the fixed effect of group was not significant on the immediate posttest (z = -0.90, p = .37, d = 0.12 [-0.24, 0.49]), indicating no significant difference in within-set error rates on related items between the two groups (massed = 3.83%, spaced = 2.70%). However, the fixed effect of group was significant on the delayed posttest (z = -2.46, p = .01, d = 0.31 [-0.06, 0.68]). This suggests that, on the delayed posttest, the massed group committed significantly more within-set errors on related items than the spaced group (massed = 6.32%, spaced = 3.70%). Model B, which examined the massed group's within-set error rates for related and unrelated items, revealed a significant effect of relatedness for both immediate and delayed posttests (immediate: z = 3.02, p = .002, d = 0.45; delayed: z = 5.32, p < .001, d = 0.89). As illustrated in Figure 7, the massed group produced more within-set errors for the related items than for the unrelated items on both immediate (related: 3.83%, unrelated: 0.54%) and delayed posttests (related: 6.32%, unrelated: 0.47%). To summarize, the findings suggest that (a) spacing significantly reduced interference among semantically related items particularly on the delayed posttest, and (b) semantically related sets caused more interference than unrelated sets among massed learners on both the immediate and delayed posttests.

Judgements of Learning

After the immediate posttest, the participants were asked to estimate how many target words out of 48 they would expect to remember 1 week later. On average, the participants in the massed and spaced groups predicted that they would remember 13.74 (SD = 10.27) and 13.21 (SD = 8.42) words, respectively. (Twenty-three participants did not provide responses.) The estimated and actual scores were analyzed using a 2 × 2 mixed analysis of covariance (ANCOVA) with group (massed / spaced) as a between-participant variable and score type (estimated / actual) as a within-participant variable. The LABJT score was included as a covariate. The ANCOVA revealed a significant interaction between the group and the score type, F (1, 100) = 7.92, p = .01, η_p^2 = .04.

The follow-up analysis found no statistically significant difference in the estimated scores between the massed and spaced groups with a very small effect size, F (1, 90) = 0.79, p = .38, η_p^2 = .01. The findings suggest that, although the spaced group recalled significantly more words than the massed group 1 week after the learning phase, the participants were not aware of the benefits of spacing. In the massed group, the difference between the estimated and actual performance was statistically significant with a large effect size (Cohen, 1988), F (1, 89) = 22.64, p < .001, η_p^2 = .20. In the spaced group, however, the difference between the estimated and actual performance was not statistically significant and only a very small effect size was found, F (1, 91) = 0.08, p = .78, η_p^2 = .001. The findings suggest that, although the massed group overestimated retention compared to their actual delayed posttest performance (estimated: 31.70%; actual: 16.12%), no significant overestimation was found for the spaced group (estimated: 28.63%; actual: 27.53%).

Discussion

The current study investigated the effects of semantic clustering on L2 vocabulary learning (RQ1). It was hypothesized that semantic clustering would hinder retention because it would cause interference between the similar meanings of related words (Hypothesis 1). This study found no significant differences between semantically related and unrelated words in translation accuracy rates either during the learning phase or on the posttests. However, semantically related sets caused more within-set errors than unrelated sets both during the learning phase and on the posttests. Hypothesis 1, therefore, was not supported for translation accuracy but was supported for within-set error rates.

The lack of a significant difference between the semantically related and unrelated sets in posttest scores is consistent with the results of our meta-analysis of earlier studies (see Appendix A in the online supplementary materials). Why did semantic clustering not decrease translation accuracy even though it caused more interference? One possible explanation is that the negative effects of interference were offset by the positive effects of semantic clustering. As discussed above in the Literature Review section, researchers argue that semantic clustering has both advantages and a disadvantage. The disadvantage is that, by causing interference between similar meanings of related words, semantic clustering might hinder learning. At the same time, semantic clustering might facilitate retention because (a) it reflects how vocabulary is stored in the mental lexicon, (b) it

introduces desirable difficulty, and (c) it leads to extra attention, effort, or engagement from learners. In this study, semantic clustering did not decrease translation accuracy possibly because these positive effects of semantic clustering compensated for the negative effects of interference.

The second research question asked whether spacing has differential effects on semantically related and unrelated words. Hypothesis 2 predicted that semantically related words would benefit more from spacing than unrelated words. While spacing did interact with semantic relatedness, it was unrelated words, not related, that benefited more from spacing, which is contrary to our hypothesis. The benefits of spacing on the translation accuracy were almost twice or three times as large for the unrelated words (immediate: $d = 0.35$; delayed: $d = 0.72$) as for the related words (immediate: $d = 0.12$; delayed: $d = 0.38$). Our hypothesis was based on two premises: (a) spacing would reduce interference among related words, and (b) due to the diminished interference, related words would benefit more from spacing than unrelated words. The first premise was supported by the current study because on the delayed posttest, the spaced group produced significantly fewer within-category errors for the related words than the massed group. However, this diminished interference effect did not translate into better retention, probably because interference was not necessarily harmful and may even have been beneficial to some extent. As discussed above, semantic clustering has both advantages and a disadvantage. Two of the advantages assume that semantic clustering causes interference. First, by causing interference, semantic clustering makes learning desirably difficult, which might potentially lead to better long-term retention (Bjork, 1999). Second, due to the difficulty caused by interference, semantically related items may receive more attention, effort, or engagement than unrelated words (Finkbeiner & Nicol, 2003; Hashemi & Gowdasiaei, 2005; Tagashira et al., 2010). As predicted by Hypothesis 2, spacing significantly reduced interference among related words. Although the reduced interference effect facilitated learning to some extent, at the same time it might have decreased learning by diminishing the advantages resulting from interference (desirable difficulty and extra attention, effort, or engagement). This was possibly the reason why the benefits of spacing were more pronounced for the unrelated words than for the related words.

The above interpretation is partially supported by the descriptive statistics of the translation accuracy rates on the delayed posttest. On the descriptive level, the related items led to higher

translation accuracy (17.74%) than unrelated items (15.99%) for the massed schedule, which was associated with more interference. Furthermore, despite the reduced interference effect, the related items resulted in lower translation accuracy (24.15%) than unrelated items (29.09%) for the spaced schedule. These results suggest that, by alleviating interference, spacing might have diminished the advantages resulting from interference. At the same time, because no statistically significant difference was found between the semantically related and unrelated words in delayed posttest scores, this interpretation remains only speculative and must be tested empirically in future research. One way to do so would be to compare the effects of multiple spacing schedules that differ in the amounts of spacing. For instance, suppose that the effects of short and long spacing were compared, and unrelated items led to higher scores than related items for long spacing, whereas related items resulted in higher scores than unrelated items for short spacing. These findings would support the supposition that interference has not only negative but also positive effects, and reducing interference through spacing might diminish the potential benefits of interference.

Pedagogical Implications

Despite its popularity, many researchers have recommended against semantic clustering because it might inhibit learning by causing interference between related words. This study did not find any significant differences between semantically related and unrelated sets in translation accuracy on the posttests, which is consistent with the results of our meta-analysis of earlier research (see Appendix A in the online supplementary materials). Semantically related items, however, resulted in a larger number of within-set errors than unrelated items. Within-set errors may be problematic because they might cause non-understanding or even miscommunication (e.g., producing *teacher* when meaning *student*). The findings of this study, therefore, support the widely held view that semantic clustering should be avoided (e.g., Barcroft, 2015; Folse, 2004; Nation, 2000; Schmitt, 2007, 2010).

From a broader perspective, this study further underscores the value of spacing for vocabulary learning. The delayed posttest results showed that spacing enhanced long-term retention regardless of the semantic relatedness of words. When collapsed across the related and unrelated items, spacing was 1.6 times as effective as massing 1 week after the learning phase (massed:

16.87%, spaced: 26.62%). This adds to existing literature suggesting that spacing facilitates L2 vocabulary learning (e.g., Bahrick et al., 1993; Nakata, 2015; Nakata & Webb, 2016). Given its robustness, teachers, learners, and materials developers should take advantage of the benefits of spacing.

At the same time, the questionnaire survey conducted after the immediate posttest revealed that learners were not necessarily aware of the positive effects of spacing. Although the spaced group recalled significantly more words than the massed group 1 week after the learning phase, the participants predicted that both treatments would lead to similar levels of retention (massed: 31.70%; spaced: 28.63%). The findings support raising awareness of the importance of spacing. One potential reason why participants were unaware of the benefits of spacing is that spaced learning led to significantly lower translation accuracy during learning (37.78%) than massed learning (70.32%). Because learners tend to equate learning-phase performance with long-term retention (e.g., Bjork, 1999; Kornell, 2009), the learners in the spaced group perhaps felt that they were not learning effectively, possibly leading to judgements of learning that were similar to those reported by the massed group. In contrast, the massed group significantly overestimated retention. This is probably because massed learning led to higher translation accuracy during learning, which made them overconfident and created "an illusion of effective learning" (Kornell, 2009, p. 1302). The findings highlight the importance of raising awareness that treatments that induce many incorrect responses during learning can be effective in the long term, while treatments that produce many correct responses during learning can be detrimental to long-term retention (desirable difficulty framework; Bjork, 1999).

Concluding Remarks

The present study compared the learning of semantically related and unrelated words while assessing long-term retention. By controlling lexical variables that might affect item difficulty, this study isolated the effects of semantic relatedness from other possibly confounding variables. Another goal of this study was to explore an interaction between semantic relatedness and spacing. The study found that, although spacing facilitated the learning of both semantically related and unrelated words, the advantage of spacing was more pronounced for the unrelated words.

Although the findings of this study are valuable, the present study is not without limitations.

One limitation is that this study used only one type of posttest in which participants were asked to translate L2 words into their L1. In future research, it might be useful to give another type of posttest (e.g., translate from L1 to L2) to examine whether semantic relatedness and spacing have differential effects on different aspects of word knowledge. Another limitation is that the treatment in this study involved paired-associate learning. Although the paired-associate format is useful because it allows for strict control over the treatment, the findings of this study might not necessarily be generalizable to other learning conditions. In future research, it might be valuable to examine the effects of semantic clustering and spacing on other kinds of vocabulary learning tasks (e.g., incidental learning through reading and listening). Considering the popularity of semantic clustering and the robustness of the distributed practice effect, further research along this line is valuable because it has direct and immediate application for teachers, learners, and materials developers.

References

Amano, S., & Kondo, T. (1999). *Nihongo-no Goitokusei [Lexical properties of Japanese: Word familiarity]*. Tokyo, Japan: Sanseido.

Baddeley, A. D. (1997). *Human memory: Theory and practice* (Revised ed.). East Sussex, UK: Psychology Press.

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

Barcroft, J., & Rott, S. (2010). Partial word form learning in the written mode in L2 German and Spanish. *Applied Linguistics*, *31*, 623-650.

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). Cambridge, MA: MIT Press.

Bolger, P., & Zapata, G. (2011). Semantic categories and context in L2 vocabulary learning. *Language Learning*, *61*, 614-646.

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

de Groot, A. M. B. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, *56*, 463-506.

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 Groot, A. M. B., & van Hell, J. G. (2005). The learning of foreign language vocabulary. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 9-29). New York, NY: Oxford University Press.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign-language vocabulary learning. *Language Learning*, *43*, 559-617.

Erten, I. H., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets. *System*, *36*, 407-422.

Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied Psycholinguistics*, *24*, 369-383.

Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor, MI: University of Michigan Press.

Hashemi, M. R., & Gowdasiaei, F. (2005). An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC Journal*, *36*, 341-361.

Higa, M. (1963). Interference effects of intralist word relationships in verbal learning. *Journal of Verbal Learning and Verbal Behavior*, *2*, 170-175.

Hoshino, Y. (2010). The categorical facilitation effects on L2 vocabulary learning in a classroom setting. *RELC Journal*, *41*, 301-312.

Ishii, T. (2015). Semantic connection or visual connection: Investigating the true source of confusion. *Language Teaching Research*, *19*, 712-722.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297-1317.

Laufer, B. (2012). Second language word difficulty. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 5151-5156). Oxford, UK: Wiley-Blackwell.

Lotto, L., & de Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, *48*, 31-69.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam, Netherlands: John Benjamins.

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*, 677-711.

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. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. *TESOL Journal, 9(2),* 6-10.

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. (2012). The BNC/COCA word family lists. Retrieved March 3, 2017, from http://www.victoria.ac.nz/lals/about/staff/paul-nation

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.

Papathanasiou, E. (2009). An investigation of two ways of presenting vocabulary. *ELT Journal, 63,* 313-322.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64,* 878-912.

Rodgers, T. S. (1969). On measuring vocabulary difficulty: An analysis of item variables in learning Russian-English vocabulary pairs. *International Review of Applied Linguistics in Language Teaching, 7,* 327-343.

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics, 38,* 906-911.

Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language Learning, 43,* 313-344.

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave Macmillan.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E. Bourne (Eds.),

Foreign language learning: Psycholinguistic studies on training and retention (pp. 77-90). Mahwah, NJ: Erlbaum.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46*, 419-440.

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.

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System, 21*, 371-380.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research, 13*, 138-163.

Vaden, K. I., Hickok, G. S., & Halpin, H. R. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. Retrieved October 1, 2017, from http://www.iphod.com

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System, 25*, 261-274.

Wesche, M. B., & Paribakht, T. S. (1996). Assessing L2 vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*, 13-40.

Wilcox, A., & Medina, A. (2013). Effects of semantic and phonological clustering on L2 vocabulary acquisition among novice learners. *System, 41*, 1056-1069.

Note

[1] Nine participants scored correctly on one or two items on the pretest (five items in the massed group and seven items in the spaced group); these items were treated as missing values across the entire experiment by participant.

Table 1

Logistic Mixed-Effects Model of Translation Accuracy for Immediate and Delayed Posttests

	Immediate Posttest				Delayed Posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	0.21	0.27	0.77	.44	-2.39	0.26	-9.10	.00
Group	0.71	0.31	2.31	.02	1.32	0.28	4.65	.00
Relatedness	0.00	0.29	-0.01	.99	0.21	0.31	0.69	.49
Translation Latency	0.65	0.13	4.85	.00	0.10	0.11	0.87	.38
LABJT	0.28	0.12	2.40	.02	0.21	0.11	1.88	.06
Group × Relatedness	-0.47	0.20	-2.41	.02	-0.61	0.21	-2.91	.00

Note. Intraclass correlation coefficients on the random effect variances were .27 (Subject) and .17 (Item) on the immediate posttest and .18 (Subject) and .19 (Item) on the delayed posttest.

Table 2

Logistic Mixed-Effects Model of Within-Set Error Rate on Immediate and Delayed Posttests

Model A: Comparison of related items in massed and spaced groups

	Immediate Posttest				Delayed Posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	-3.73	0.27	-13.79	.00	-3.21	0.25	-13.09	.00
Group	-0.27	0.30	-0.90	.37	-0.64	0.26	-2.46	.01
Translation Latency	0.13	0.25	0.52	.60	0.27	0.16	1.70	.09
LABJT	-0.11	0.17	-0.64	.52	0.17	0.14	1.14	.25

Note. Intraclass correlation coefficients on the random effect variances were .46 (Subject) and .16 (Item) on the immediate posttest and .35 (Subject) and .24 (Item) on the delayed posttest.

Model B: Comparison of related and unrelated items in massed group

	Immediate Posttest				Delayed Posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	-14.78	2.93	-5.04	.00	-12.28	1.62	-7.59	.00
Relatedness	16.06	5.32	3.02	.00	17.17	3.23	5.32	.00
Translation Latency	0.35	0.74	0.48	.63	0.41	0.10	3.93	.00
LABJT	0.33	1.13	0.29	.77	0.35	0.18	1.97	.05

Note. Intraclass correlation coefficients on the random effect variances were .84 (Subject) and .08 (Item) on the immediate posttest and .50 (Subject) and .43 (Item) on the delayed posttest.

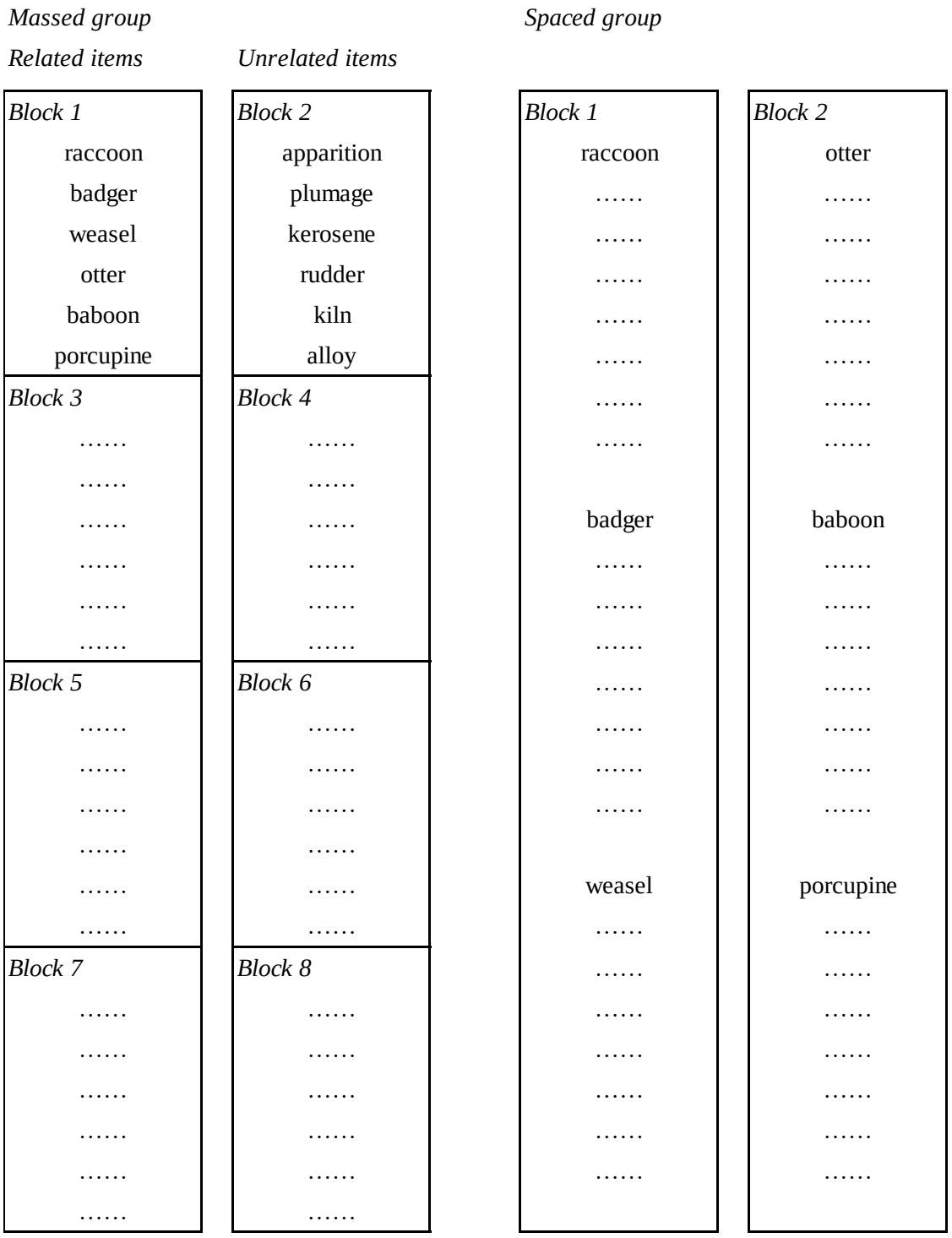


Figure 1. Operationalization of within-set errors in the massed and spaced groups. For semantically related sets, when participants provided an incorrect response from the same semantic category, it was categorized as a within-set error (e.g., producing the Japanese translation of *raccoon* for *weasel*). For semantically unrelated sets, in the massed group, when the participants produced the Japanese translation of one of the other five items from the same unrelated set (e.g., producing the Japanese translation of *apparition* for *plumage*), this was regarded as a within-set error. The

frequency of within-set errors was not calculated for the unrelated items in the spaced group.

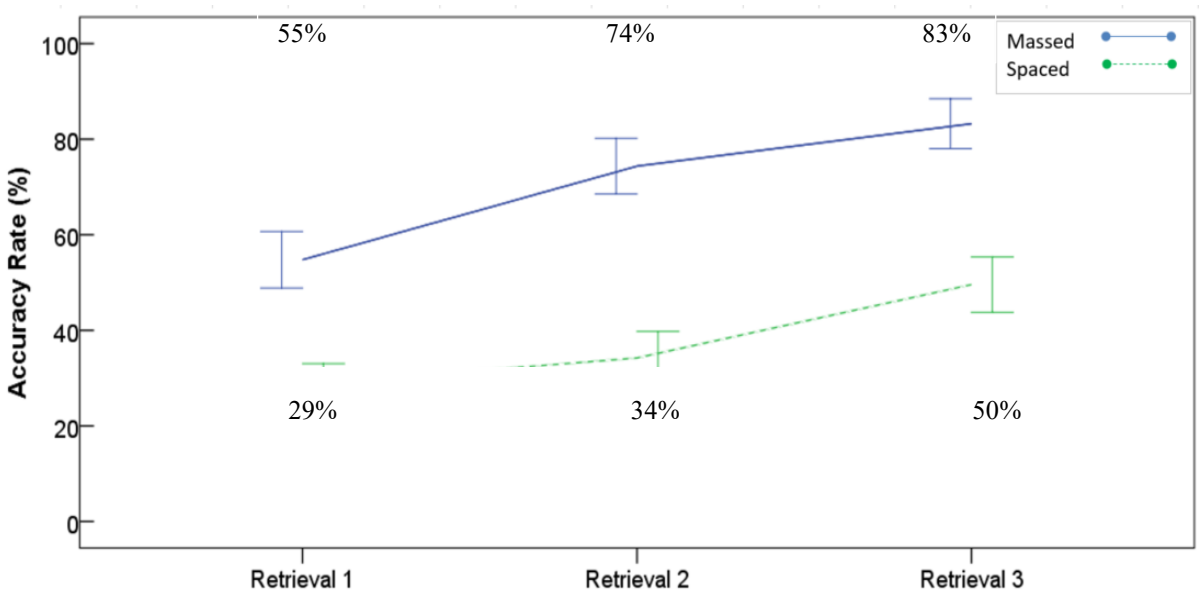
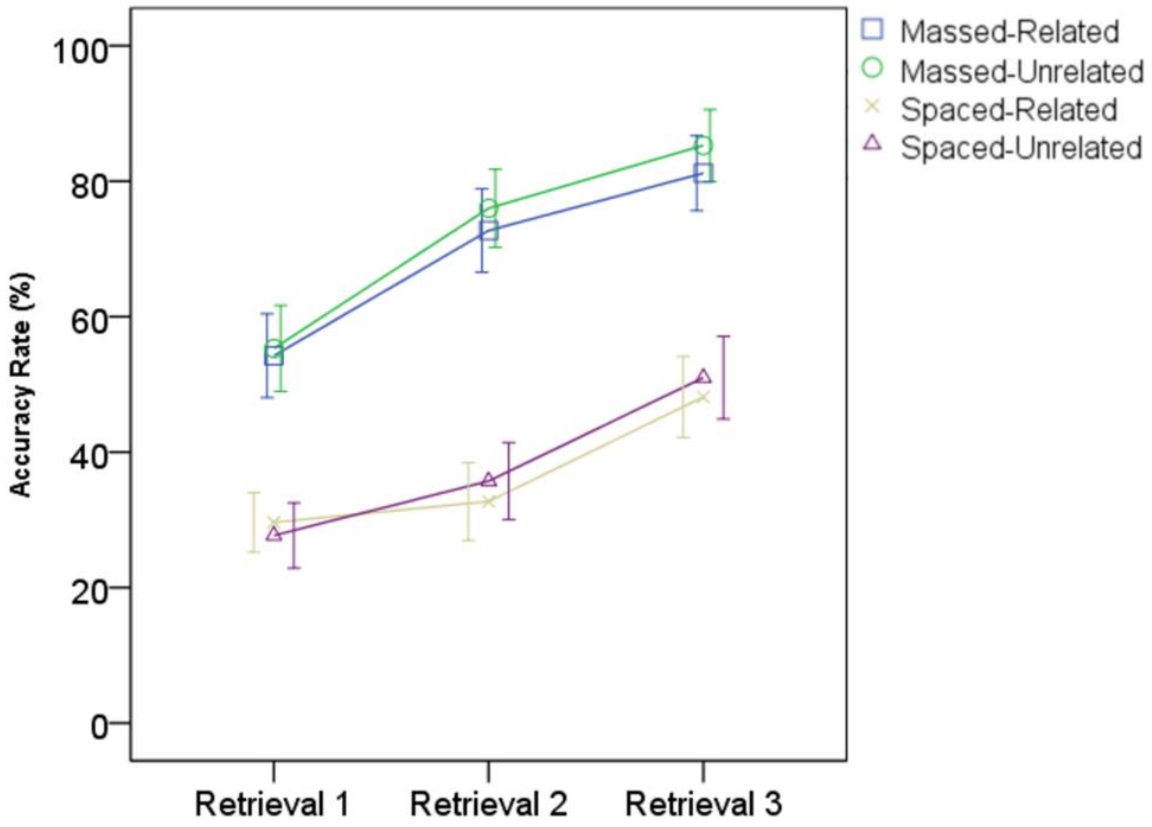


Figure 2. Mean translation accuracy rates during learning phase by spacing. The error bars indicate 95% confidence intervals (CIs).



	Retrieval 1		Retrieval 2		Retrieval 3	
	M	*SD*	*M*	*SD*	*M*	*SD*
Massed-Related	54.23%	24.31%	72.72%	24.26%	81.18%	21.84%
Massed-Unrelated	55.31%	24.99%	76.01%	22.69%	85.28%	20.93%
Spaced-Related	29.63%	16.14%	32.72%	21.04%	48.15%	21.86%
Spaced-Unrelated	27.70%	17.68%	35.73%	20.83%	51.00%	22.34%

Figure 3. Mean translation accuracy rates during learning phase by spacing and semantic relatedness. The error bars indicate 95% CIs.

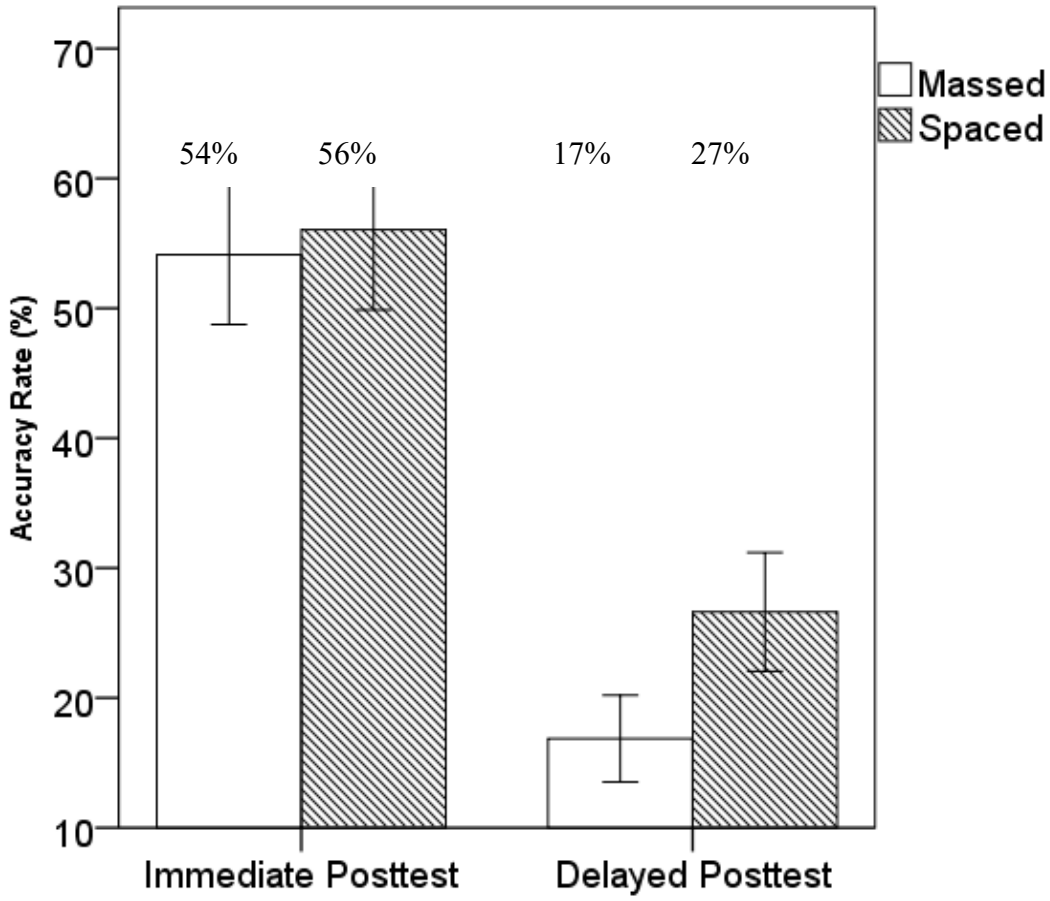


Figure 4. Mean translation accuracy rates on posttests by spacing. The error bars indicate 95% CIs. Since items answered correctly on the pretest were treated as missing values by participant, the pretest score was zero.

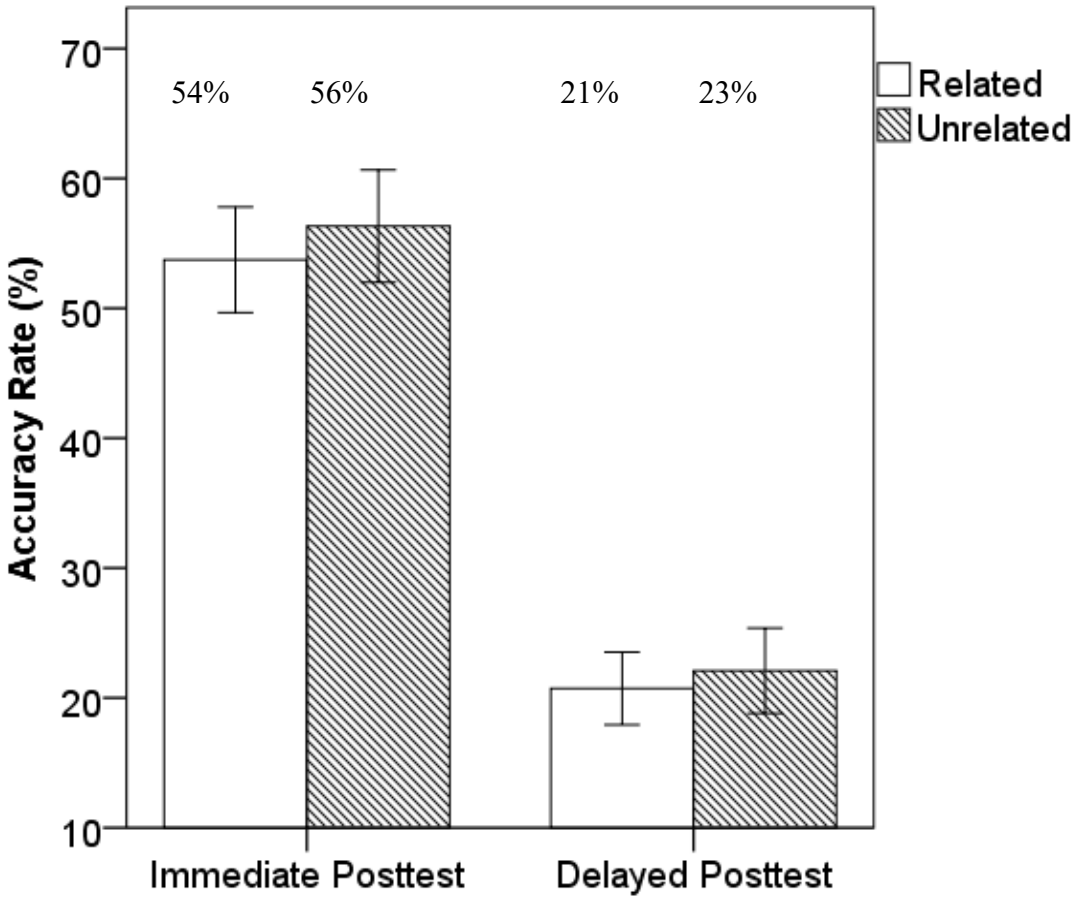


Figure 5. Mean translation accuracy rates on posttests by semantic relatedness. The error bars indicate 95% CIs.

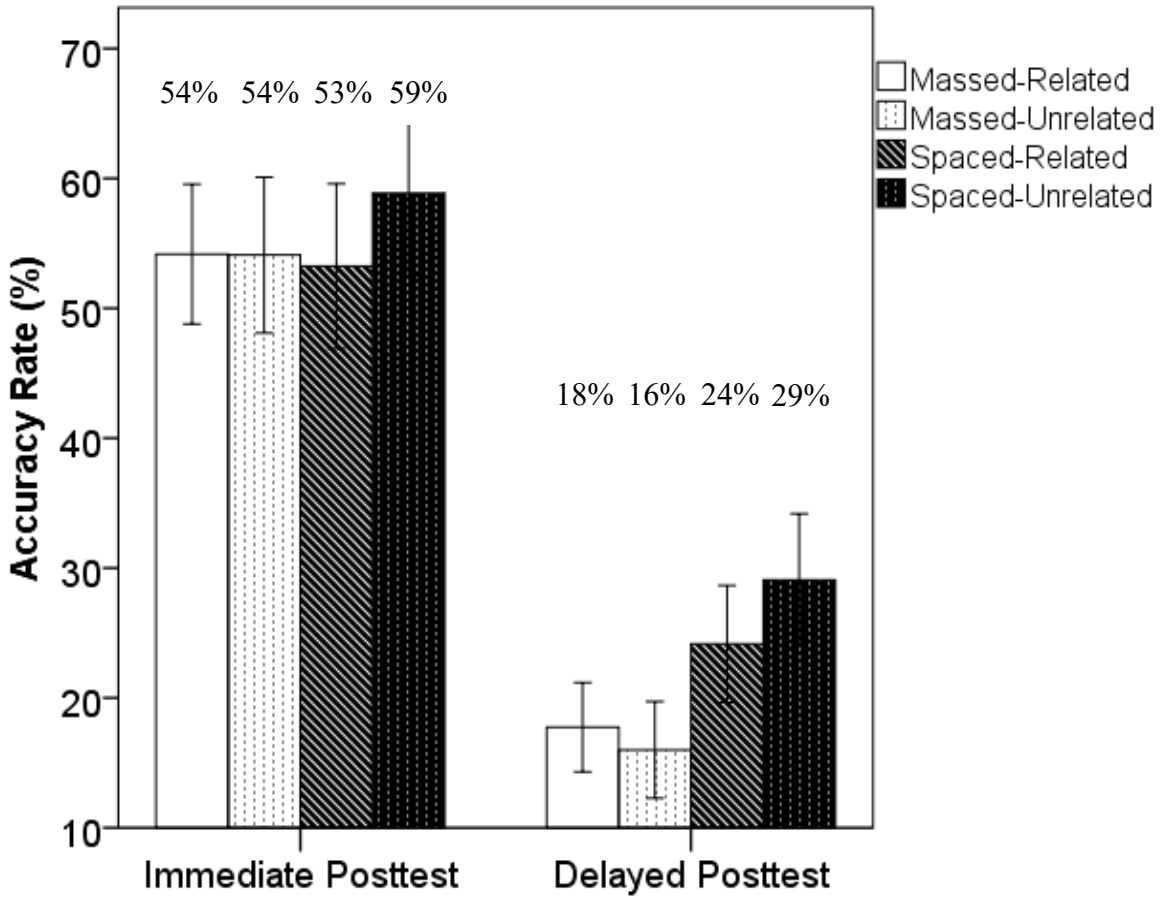


Figure 6. Mean translation accuracy rates on posttests by spacing and semantic relatedness. The error bars indicate 95% CIs.

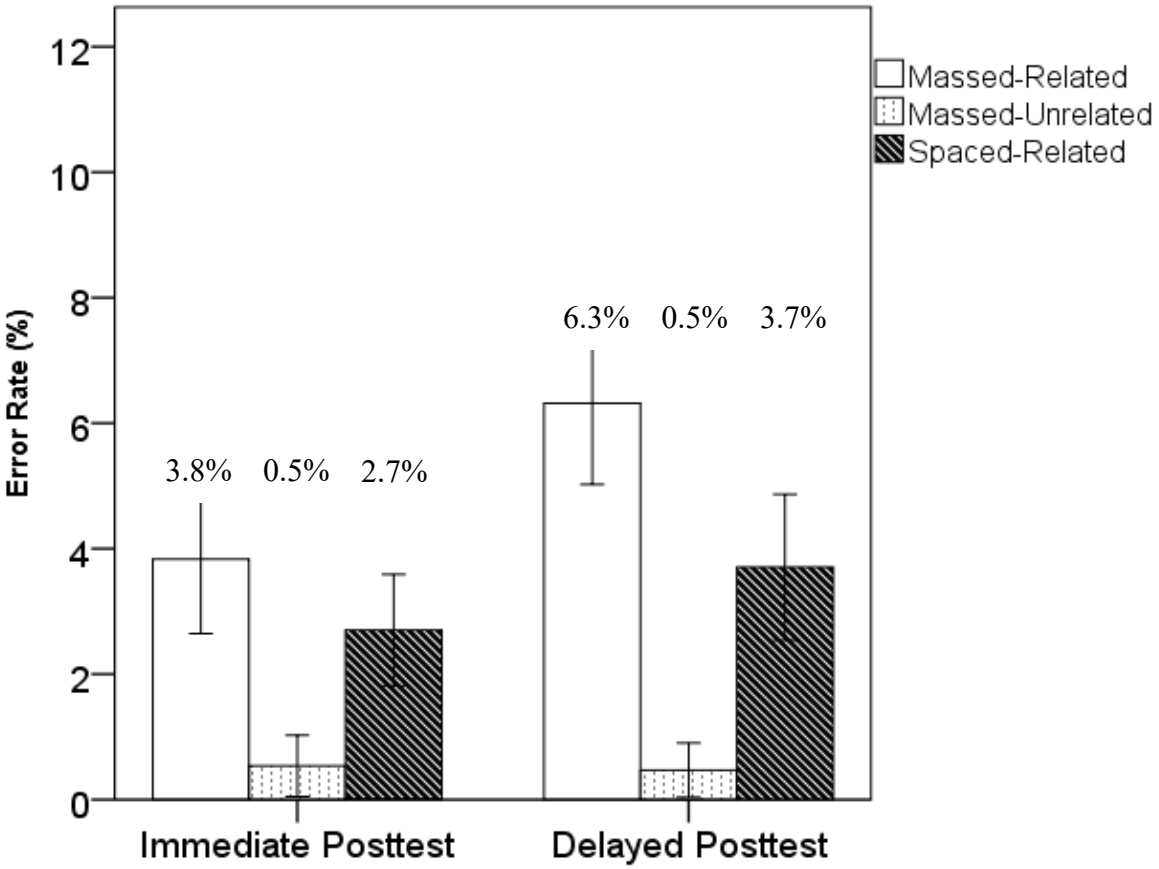


Figure 7. Mean within-set error rates on posttests by spacing and semantic relatedness. The error bars indicate 95% CIs.

Acknowledgements

This research was supported in part by Grant-in-Aid for Young Scientists (A) (#16H05943) awarded to the first author from Japan Society for the Promotion of Science. An earlier version of this paper was presented at the 27th Annual Conference of the European Second Language Association (EuroSLA), Reading, UK, 2017. We are very grateful to Paul Nation, Yu Tamura, Akira Murakami, and Steve Porritt for their invaluable advice.