Studies in Second Language Acquisition, 2016, **38**, 523–552.
doi:10.1017/S0272263115000236

DOES STUDYING VOCABULARY IN SMALLER SETS INCREASE LEARNING?

The Effects of Part and Whole Learning on Second Language Vocabulary Acquisition

Tatsuya Nakata
Victoria University of Wellington

Stuart Webb
The University of Western Ontario

The present study examined the effects of part and whole learning on the acquisition of second language (L2, English) vocabulary. In whole learning, the materials to be learned are repeated in one large block, whereas, in part learning, the materials are divided into smaller blocks and repeated. Experiment 1 compared the effects of the following three treatments: 20-item whole learning, four-item part learning, and 10-item part learning. Unlike previous studies, part and whole learning were matched in spacing. In Experiment 2, spacing as well as the part-whole

Tatsuya Nakata is now at Kansai University.

This research was supported in part by Faculty Research Grants (#98778 and #105920) and Victoria PhD Scholarship from Victoria University of Wellington and Student Exchange Support Program (Long-Term Study Abroad) Scholarship from Japan Student Services Organization awarded to the first author. An earlier version of this paper was presented at the First Auckland Postgraduate Conference in Linguistics and Applied Linguistics, Auckland, 2011. This article is based on part of the first author's doctoral dissertation, which was submitted to Victoria University of Wellington in 2013. We are very grateful to Paul Nation, Jan Hulstijn, Rod Ellis, and anonymous *SSLA* reviewers for their invaluable advice. We would also like to extend our special thanks to Atsushi Mizumoto for his cooperation with data collection.

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

learning distinction were manipulated, and the following three treatments were compared: 20-item whole learning, four-item part learning with short spacing, and four-item part learning with long spacing. Results of the two experiments suggest that, (a) as long as spacing is equivalent, the part-whole distinction has little effect on learning, and (b) spacing has a larger effect on learning than the part-whole distinction.

Research on the frequency effect suggests that the learning of second language (L2) vocabulary (e.g., Pigada & Schmitt, 2006; Zahar, Cobb, & Spada, 2001) as well as of the L2 in general (e.g., Ellis, 2002; Hulstijn, 2002; Larsen-Freeman, 2002) increases as a function of frequency. This raises the question of how the encounters of a given L2 word should be distributed to optimize L2 vocabulary learning. Previous studies have examined the effects of two types of distribution: part learning and whole learning. In whole learning, the materials to be learned are repeated in one large block, whereas, in part learning, the materials are divided into smaller blocks and repeated.

The effects of part and whole learning have been examined in the learning of a number of verbal materials and motor skills (see Woodworth & Schlosberg, 1954, for a review). For instance, suppose that the learner wants to memorize a poem. Would it be more effective to read the whole poem several times or would it be more effective to read part by part for the same number of times? When learning to play music, would it be effective to practice part by part before trying to play the whole tune? Although most previous studies on part and whole learning have been conducted in the field of psychology (e.g., Brown, 1924; Kornell, 2009; McGeoch, 1931; Woodworth & Schlosberg, 1954), the issue of part and whole learning may also be relevant for L2 vocabulary acquisition. For instance, suppose we have 20 words to study. Would it be more effective to learn all 20 words at one time (whole learning) or to divide the words into smaller blocks to be learned block-by-block (part learning)? Hereafter, the number of words to be learned at once will be referred to as *block size* (Crothers & Suppes, 1967; Hulstijn, 2001). For instance, if 20 target words are repeated in one large block of 20 items, the block size is 20. If 20 words are repeated in four blocks of five items, the block size is five.

From a theoretical perspective, the retrieval practice effect (Baddeley, 1997; Ellis, 1995) and the list-length effect (Gillund & Shiffrin, 1984; Van Bussel, 1994) suggest that part learning should be more effective than whole learning (see the "Theoretical Background" section). Learners, teachers, materials developers, and researchers also tend to believe that part learning is more effective (e.g., Joseph, Watanabe, Shiung, Choi, & Robbins, 2009; Kornell, 2009; Salisbury & Klein, 1988; Wissman, Rawson, & Pyc, 2012; Woodworth & Schlosberg, 1954). Wissman and colleagues

(2012) surveyed 374 American college students and found that 72.2% of them considered part learning to be more effective than whole learning, whereas only 16.3% of them responded that whole learning might be superior. Some researchers also claim that studying vocabulary in a relatively small block, such as that of 10 to 12 words, enhances learning (Joseph et al., 2009; Salisbury & Klein, 1988). Part learning is also a common learning method employed by previous empirical studies. For instance, in Webb (2005), 10 target words were encountered three times in a block of one item. In Barcroft and Rott (2010), 24 target words were divided into three blocks of eight items and presented twice. In Pyc and Rawson (2007), 48 target items were studied in a block of six or 24 items. The use of part learning in previous research may partially reflect the belief among researchers that part learning increases vocabulary acquisition.

Contrary to the view that part learning facilitates learning, most empirical studies have shown that whole learning may be more effective (Brown, 1924; Crothers & Suppes, 1967, Experiments 8 & 9; Kornell, 2009, Experiments 1–3; McGeoch, 1931, Experiments 1 & 3; Seibert, 1932). Existing studies, however, are limited in that the part-whole learning distinction and spacing have been confounded. More specifically, in previous studies in which target items were encountered more than once, whole learning always had longer spacing than part learning. Spacing refers to an interval between learning opportunities of a given item. For instance, if encounters of a given item are separated by 5 min, there is spacing of 5 min. Research shows that larger spacing generally leads to better long-term retention than shorter spacing (e.g., Bahrick & Phelps, 1987; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Cepeda et al., 2009; Karpicke & Bauernschmidt, 2011; Kornell, 2009; Metcalfe, Kornell, & Finn, 2009; Pashler, Zarow, & Triplett, 2003; Pyc & Rawson, 2012). The confounding of the part-whole distinction and spacing is problematic considering that spacing affects L2 vocabulary learning. In other words, the results of the earlier studies may be at least partly attributed to spacing rather than the part-whole distinction per se. The present study aimed to investigate the effects of part and whole learning on L2 vocabulary acquisition in two experiments while isolating the effects of the part-whole distinction and spacing.

REVIEW OF LITERATURE

Theoretical Background

Hereafter, we assume that learning involves retrieval, whereby learners are asked to recall information about the L2 word from memory because research shows that retrieval increases learning (e.g., Barcroft, 2007; Karpicke & Roediger, 2008; Pashler, Rohrer, Cepeda, & Carpenter, 2007;

Rohrer & Pashler, 2007). Retrieval can be receptive or productive (e.g., Nation, 2013). For instance, if learners try to remember the meaning of a L2 word in reading or listening, it involves receptive retrieval. If learners try to use a L2 word in speaking or writing, it involves productive retrieval.

The retrieval practice effect and the list-length effect suggest that part learning should be more effective than whole learning. The retrieval practice effect refers to the phenomenon whereby successful retrievals from memory contribute to greater long-term retention than unsuccessful retrievals (Baddeley, 1997; Ellis, 1995). Part learning may lead to a higher level of retrieval success than whole learning because, in part learning, target items are encountered after a shorter interval compared with whole learning. For instance, when 100 items are repeated in a block of 100 items (whole learning), encounters of a given item are separated by 99 other items, whereas, when a block size of four items is used (part learning), only three items intervene between the encounters of a given item. As a result, in part learning, retrieval may take place before forgetting occurs, possibly producing higher retrieval success than whole learning. According to the retrieval practice effect, therefore, part learning should be more effective because learners are more likely to benefit from the positive effects of retrieval success.[1] The list-length effect also predicts an advantage of part over whole learning. According to this effect, memory performance is inversely related to the number of items to be studied (Gillund & Shiffrin, 1984; Van Bussel, 1994). In other words, when 10 words are studied at one time, 60% of them may be learned successfully, whereas when 40 words are studied, only 40% of them may be learned (Gillund & Shiffrin, 1984). The list-length effect also suggests that part learning, in which a smaller number of items are studied at once than whole learning, may lead to superior long-term retention.

Empirical Evidence

Contrary to the view that part learning facilitates learning, most empirical studies have shown that whole learning may be more effective for vocabulary acquisition. Kornell (2009), for instance, compared the effectiveness of part and whole learning on the learning of first language (L1) low-frequency vocabulary in three experiments. In his first experiment, in the whole learning condition, 20 target words were repeated in a block of 20 items and encountered four times throughout the treatment. In the part learning condition, the 20 target items were encountered four times in four blocks of five items. On the posttest conducted 1 day after the treatment, learners in the whole learning condition significantly

outperformed learners in the part learning condition. Kornell also found the advantage of whole over part learning in Experiments 2 and 3. Six experiments have supported Kornell's (2009) findings (Brown, 1924; Crothers & Suppes, 1967, Experiments 8 & 9; McGeoch, 1931, Experiments 1 & 3; Seibert, 1932).

Three experiments, however, failed to find any advantage of whole over part learning (Crothers & Suppes, 1967, Experiments 10 & 11; Van Bussel, 1994). Van Bussel found evidence of the superiority of part learning (block size 20) over whole learning (block size 40). The contradictory results may be ascribed in part to the number of encounters with target items during learning. Whereas the target items were encountered more than once during the treatment in all other earlier studies (Brown, 1924; Crothers & Suppes, 1967, Kornell, 2009; McGeoch, 1931; Seibert, 1932), the target items were encountered only once in Van Bussel (1994). A more detailed explanation of why the inconsistent findings may have stemmed from this difference will be offered later in this section.

Although Crothers and Suppes (1967) found the advantage of whole learning in their Experiments 8 and 9, they failed to do so in Experiments 10 and 11. Nation (2013) points out that the inconsistent results may have been caused because the effects of part and whole learning may interact with task difficulty. According to Nation, the treatments in Crothers and Suppes's Experiments 10 and 11 were more demanding than those in their Experiments 8 and 9 in at least two respects. First, in Experiments 8 and 9, the target words were practiced in a recognition format, whereby participants were asked to choose the correct response from three options. Experiments 10 and 11, in contrast, used a recall format, and participants were asked to produce, rather than to choose, the correct response. Second, although no time limit was imposed in Experiments 8 and 9, participants were required to write down a response within 5 s in Experiments 10 and 11. Due to these two differences, difficulty was probably higher in Experiments 10 and 11 compared with their earlier experiments. This may be partly the reason why Crothers and Suppes found the superiority of whole learning in their earlier experiments but not in Experiments 10 and 11.[2]

Limitations of Previous Research

Even though the findings of these previous studies are very valuable, they may be limited in that the part-whole learning distinction and spacing were confounded. More specifically, in previous studies in which target items were encountered more than once, whole learning always had longer spacing than part learning. One common index of spacing used

in previous studies is the average number of intervening trials between encounters of each target item (e.g., Karpicke & Roediger, 2007; Landauer & Bjork, 1978; Pyc & Rawson, 2007). For instance, in the whole learning condition in Kornell (2009, Experiment 1), 20 target words were repeated in a block of 20 items. As a result, encounters of a given item were separated by 19 trials for other items. In contrast, in Kornell's part learning condition, the 20 target items were repeated in four blocks of five items, and only four trials intervened between the encounters of a given item. Whole learning, hence, had longer spacing (19 trials) than part learning (four trials).

The confounding of the part-whole learning distinction and spacing is problematic because larger spacing generally leads to better long-term retention than shorter spacing, a phenomenon known as the *distributed practice effect* (e.g., Bahrick & Phelps, 1987; Cepeda et al., 2008; Cepeda et al., 2009; Karpicke & Bauernschmidt, 2011; Kornell, 2009; Metcalfe et al., 2009; Pashler et al., 2003; Pyc & Rawson, 2012). The superiority of whole learning in the earlier studies may thus be at least partly attributed to spacing. The part-whole distinction and spacing have been confounded in all existing studies in which target items were encountered more than once (Brown, 1924; Crothers & Suppes, 1967; Kornell, 2009; McGeoch, 1931; Seibert, 1932). Because the target items were not repeated in Van Bussel (1994), there was no spacing, and it was not possible for the part-whole distinction to be confounded with spacing in that study.[3]

Although Van Bussel's (1994) findings are useful, his study may have limited pedagogical value because multiple exposures to target words may be common in a real-life study situation. Thus, it may be beneficial to conduct research in which (a) the part-whole distinction is not confounded with spacing and (b) target items are encountered more than once.

EXPERIMENT 1

With the limitations of the existing studies in mind, Experiment 1 investigated the effects of part and whole learning that were matched in spacing. Specifically, the following three block sizes were compared: block sizes of four, 10, and 20 words. The block size of 20 treatment was whole learning, and the block sizes of four and 10 treatments involved part learning. These three block sizes were chosen for two reasons. First, previous studies have found that whole learning is more effective than part learning when the block sizes being compared are of 20 words or fewer (Brown, 1924; Kornell, 2009, Experiments 1–3; McGeoch, 1931, Experiments 1 & 3; Seibert, 1932). It was judged, therefore, that the use of these three block sizes may provide statistically significant results. Second, using relatively small blocks may increase ecological validity

because researchers, teachers, learners, and materials developers tend to believe that a small block size enhances learning more than a large one (e.g., Joseph et al., 2009; Kornell, 2009; Salisbury & Klein, 1988; Wissman et al., 2012; Woodworth & Schlosberg, 1954).

The treatment in the present study involved learning in a paired-associate format, whereby learners were required to associate the L2 word form with its meaning. A paired-associate learning condition was chosen for three reasons. First, research has indicated that paired-associate learning is an effective and efficient method of L2 vocabulary learning (Elgort, 2011; Fitzpatrick, Al-Qarni, & Meara, 2008; Webb, 2009; see Nation, 2013, for a review). Second, paired-associate learning is a widely used vocabulary learning strategy (Nakata, 2011; Schmitt, 1997; Wissman et al., 2012). Third, earlier studies of part and whole learning on vocabulary acquisition have typically involved paired-associate learning (e.g., Crothers & Suppes, 1967; Kornell, 2009; Seibert, 1932). Using paired-associate learning conditions in the present research may thus provide greater explanatory value than using other approaches.

By isolating the effects of the part-whole learning distinction and spacing, the current study may allow us to determine how part and whole learning influence vocabulary acquisition in a more rigorous manner than do earlier studies. The research question of Experiment 1 is as follows: Is whole learning more effective than part learning for L2 vocabulary acquisition when spacing is equivalent?

Method

There were two independent variables in Experiment 1. The first independent variable was the distribution of encounters (i.e., four-item part, 10-item part, and 20-item whole learning). The second independent variable was the retention interval (interval between the treatment and posttest; i.e., immediate and 1-week delayed posttests). The distribution of encounters was a between-participant variable, and the retention interval was a within-participant variable. The dependent variable was the number of correct responses during the treatment (learning phase performance) and on the posttest (posttest performance). Learning phase performance was analyzed to test the assumption that part learning produces more correct retrievals than whole learning during the treatment (see the "Theoretical Background" section).

Participants. The participants were 95 first-year Japanese students at a university in the Kansai area of Japan. Four students were excluded from analysis because they demonstrated prior knowledge of one or more target words on the pretest (see the "Dependent Measures" subsection for

details about the pretest). The remaining 91 students consisted of 20 engineering, 31 commerce, and 40 law majors. Their average score on the first to the sixth 1,000-word frequency levels of the Vocabulary Size Test (VST; Nation & Beglar, 2007) was 33.92 (*SD* = 6.42) out of 60. The participants were assigned to the four-item part, 10-item part, and 20-item whole learning groups in such a way that there would be no significant difference in the VST scores, $F(2, 90) = 0.22$, $p = .802$, $\eta^2 < .001$. The four-item, 10-item, and 20-item groups consisted of 28, 30, and 33 participants, respectively. The difference in the number of participants was due to the absence of participants. The three groups had a roughly equal number of participants from each of the three majors: engineering, commerce, and law.

Target and Filler Items. Twenty low-frequency English words were used as target items: *apparition*, *billow*, *cadge*, *citadel*, *dally*, *fawn*, *fracas*, *gouge*, *grig*, *levee*, *loach*, *mane*, *mirth*, *nadir*, *pique*, *quail*, *rue*, *scowl*, *toupee*, and *warble*. Items that were beyond the most frequent 9,000 word families in Nation's (2006) British National Corpus lists were chosen because the target items needed to be unfamiliar to the participants.[4] Three filler items (*husk*, *polemic*, and *smudge*) were also studied and tested like target words but were not included in the analysis. The filler items were used for two reasons. First, they were used to match spacing in the three groups (see Appendix S1 in the online supplementary material for details). Second, filler items were used as primacy and recency buffers and studied at the beginning and end of the treatment (e.g., Karpicke & Roediger, 2007).

Procedure. The experiment was conducted during regular class hours using a computer program that was developed by one of the authors. Participants first received instruction about the computer software and practiced using it with three sample word pairs excluded from the treatment (*apple* - りんご, *orange* - オレンジ, and *banana* - バナナ). After the practice, the pretest was administered to determine whether the participants had any prior knowledge of the target words. The participants then completed the treatment. In the treatment, the participants studied 23 English words (including three filler items). Target items were studied using a different condition (four-item part, 10-item part, and 20-item whole learning) depending on the group to which participants were assigned. Following the treatment, the participants completed a distractor task that involved answering 10 two-digit addition (i.e., math) problems (e.g., 53 + 49 = ?). The immediate posttest was administered after the distractor task. A delayed posttest was administered 1 week after the treatment to measure retention.

Treatment. Each target item was encountered five times throughout the treatment in all three groups. In the first encounter with each item,

each English target word and its Japanese translation were presented together for 8 s (e.g., *mane* - たてがみ). In the second and third encounters, target items were practiced in a receptive recall format, which required participants to translate target English words into Japanese (e.g., *mane* = ____?). In the fourth and fifth encounters, target items were practiced in a productive recall format. In this format, participants were presented with the Japanese meanings and asked to type the corresponding English translations (e.g., たてがみ = ____?). Participants were given as much time as they needed to type responses in both formats. After each response, the target word, L1 meaning, and learners' response were shown for 5 s as feedback.

Experiment 1 attempted to compare part and whole learning with equivalent spacing. There are two ways to control for spacing (Nakata, 2015). One is to match the average number of intervening trials between treatments (e.g., Karpicke & Roediger, 2007; Logan & Balota, 2008; Pyc & Rawson, 2007). In this approach, encounters of a given item are separated by the same number of trials on average between treatments so that they have equivalent spacing. The other approach is to control for the average amount of time between repetitions (e.g., Kang, Lindsey, Mozer, & Pashler, 2014; Storm, Bjork, & Storm, 2010). In this method, if encounters of a given item are separated by the same amount of time (on average) in two treatments, they are considered to be equivalent in spacing. Note that to use the second method, the treatment needs to be paced by the experimenter or computer. Otherwise, it would not be possible to ensure that a given item is encountered every 3 min, for instance. In the present study, spacing was controlled using the first method, and the number of trials was used as an index of spacing. This method was chosen because a self-paced treatment may be more desirable than a computer-paced treatment in terms of effectiveness and ecological validity (Nakata, 2013). At the same time, the average amount of time between repetitions was analyzed after the experiment to investigate whether spacing in the part and whole learning groups was equivalent when time is used as a unit of spacing rather than trial (see the "Results" section).

To ensure that part and whole learning would have equivalent spacing, trials in the four-item part, 10-item part, and 20-item whole learning groups were arranged as in Table 1. The table shows the item order and the practice format in the three groups. For instance, for Cycle 2 in the four-item part learning group, the items are listed as 1–4 and the practice format as receptive. This means that, in this cycle, items 1 to 4 were practiced in a receptive recall format once. Because encounters of a given item were separated by 19 trials on average in all three groups, they are regarded as being matched in spacing (see Appendix S1 in the online supplementary material for details). To ensure that the order of items would not offer inappropriate help in remembering (e.g., Nation,

Table 1. Item order and spacing in Experiment 1

20-item whole learning group

Cycle	Items	Practice format	Spacing
Primacy buffers	11 fillers		-
Cycle 1	1–20	Initial presentation	19
Cycle 2	1–20	Receptive recall	19
Cycle 3	1–20	Receptive recall	19
Cycle 4	1–20	*Productive recall	19
Cycle 5 (Final review)	1–20	Productive recall	-
Recency buffers	3 fillers		-
Average			19

10-item part learning group

Cycle	Items	Format	Spacing	Cycle	Items	Format	Spacing
Primacy	6 fillers		-	Cycle 5	1–10	Receptive	9
Cycle 1	1–10	Presentation	9	Cycle 6	1–10	*Productive	29
Cycle 2	1–10	Receptive	34	Cycle 7	11–20	Receptive	9
Cycle 3	11–20	Presentation	9	Cycle 8	11–20	Productive	19
Cycle 4	11–20	Receptive	34	Review	1–20	Productive	-
Filler	5 fillers		-	Recency	3 fillers		
Average							19

Four-item part learning group

Cycle	Items	Practice format	Spacing	Cycle	Items	Practice format	Spacing
Primacy	3 fillers		-	Cycle 11	1–4	Receptive	3
Cycle 1	1–4	Presentation	3	Cycle 12	1–4	*Productive	35
Cycle 2	1–4	Receptive	43	Cycle 13	5–8	Receptive	3
Cycle 3	5–8	Presentation	3	Cycle 14	5–8	Productive	31
Cycle 4	5–8	Receptive	43	Cycle 15	9–12	Receptive	3
Cycle 5	9–12	Presentation	3	Cycle 16	9–12	Productive	27
Cycle 6	9–12	Receptive	43	Cycle 17	13–16	Receptive	3
Cycle 7	13–16	Presentation	3	Cycle 18	13–16	Productive	23
Cycle 8	13–16	Receptive	43	Cycle 19	17–20	Receptive	3
Cycle 9	17–20	Presentation	3	Cycle 20	17–20	Productive	19
Cycle 10	17–20	Receptive	43	Review	1–20	Productive	-
Filler	8 fillers		-	Recency	3 fillers		
Average							19

Note. Average refers to the average spacing (mean intervening trials) for a given target word pair when collapsed across all cycles.

2013), the item order was randomized for each repetition. Therefore, 1–4 does not mean that items were studied in a fixed order such as items 1, 2, 3, and 4. In the actual treatment, items were studied in a random order such as items 2, 1, 3, 4 or 1, 2, 4, and 3. The "3 fillers" indicates that there were three trials for filler items. In all three groups, there were three filler trials at the end of the treatment, which served as recency buffers (e.g., Karpicke & Roediger, 2007). Three, six, and 11 primacy buffers were included at the beginning of the treatment in the four-item part, 10-item part, and 20-item whole learning groups, respectively. The number of primacy buffers differed among the three groups to match the total number of filler trials (i.e., 14). The asterisk in Table 1 indicates the position where a target word was practiced in a productive recall format for the first time in each group (see the "Procedure and Materials" subsection of the "Experiment 2" section).

Immediately before the recency buffers, there was a final review in all groups, during which the target items were studied once in a block of 20 items. The final review is based on Brown (1924), McGeoch (1931), and Kornell (2009, Experiment 3) and was included for four reasons. First, it was used to control spacing in the three groups (see Appendix S1 in the online supplementary material for details). Second, it was used to control the *lag to test* in the three groups. Lag to test refers to the interval between the last encounters with items and the test; it has been shown to affect memory performance (e.g., Cepeda et al., 2008; Metcalfe et al., 2009; Seibert, 1932). For example, if tests are given either 2 or 24 hr after the last encounter with items (lag to test is 2 or 24 hours), memory performance will naturally be worse with the longer lag to test. Without the final review, the first several words in the four-item part and 10-item part groups would have a rather long interval to the posttest and may be forgotten. The final review ensures that all three treatments would be controlled for lag to test. Third, the inclusion of the final review may also increase ecological validity because most students review what they will be tested on shortly before a test (Kornell, 2009). Fourth, previous studies on part and whole learning using the final review still found the advantage of whole over part learning (Brown, 1924; Kornell, 2009, Experiment 3; McGeoch, 1931, Experiment 3). These results suggest that the use of the final review may not necessarily overshadow differences in part and whole learning.

Dependent Measures

Pretest. Immediately before the treatment, a receptive recall test was given as the pretest; on this test, participants translated the target English words into Japanese (e.g., *mane* = ____?). The order of the test items

was randomized for each participant to reduce the possibility of an order effect.

Posttest. Productive and receptive tests were administered in that order as the posttest. In the productive test, participants were presented with the Japanese forms and needed to type the corresponding English translation (e.g., たてがみ = ____?). In the receptive test, they needed to provide the Japanese meanings when presented with the target words (e.g., *mane* = ____?). Learning was measured by both productive and receptive posttests because previous studies have shown that measuring both receptive and productive knowledge can provide a more accurate assessment of vocabulary learning than measuring only receptive or productive knowledge (e.g., Chen & Truscott, 2010; Nakata, 2013; Webb, 2005, 2009). The posttest was administered immediately and 1 week after the treatment. The participants were given no prior notice of the delayed posttest. The order of items in the delayed posttest was randomized again for each participant. The delayed and immediate posttests were exactly the same except for the item order. The same 23 items (20 target and three filler) were tested on the immediate and delayed posttests.

Scoring. Two procedures (strict and sensitive) were used to score responses on the tests. Scoring responses at two levels of sensitivity can provide a more accurate assessment of vocabulary learning than scoring at one level (Nakata, 2013; Webb, 2008). In the strict scoring method for the productive test, only correctly spelled responses were scored as correct. In the sensitive scoring method, which is partially based on the lexical production scoring protocol (LPSP-written; e.g., Barcroft, 2007; Barcroft & Rott, 2010), responses that were spelled correctly and those that would be awarded .75 using LPSP-written were scored as correct (e.g., *apparation*, *appartion*, and *applition* for *apparition*). For the receptive test, using the strict scoring procedure, responses were scored as incorrect if (a) they were the wrong part of speech (e.g., 後悔 [noun] for *rue*) or (b) an intransitive verb was provided for a transitive verb (e.g., 怒る [intransitive] for *pique*) and vice versa. Using the sensitive scoring system, these two kinds of responses were both scored as correct.

Results

Learning Phase Data. Because the treatment was self-paced by participants, the study times of the three groups may not have been comparable. The study time was analyzed to examine whether it was roughly

equivalent among the three groups. The participants spent 19.29 (3.30), 18.34 (2.34), and 18.31 (2.82) min on average (*SD*s in parentheses) studying the target items in the four-item part, 10-item part, and 20-item whole learning groups, respectively. No statistically significant difference was found among the three groups in study time, $F(2, 90) = 1.13$, $p = .328$, and a very small effect size was found ($\eta^2 < .001$). On the basis of the results, it may be possible to assume that the average study time was roughly equivalent among the three groups.

In the present study, the number of intervening trials was used as an index of spacing (see the "Treatment" section). At the same time, because matching the average amount of time between repetitions is another common method of controlling spacing (e.g., Kang et al., 2014; Storm et al., 2010), the average amount of time between encounters was also analyzed. The analysis showed that repetitions of a given target item were separated by 233.60 (36.63), 226.28 (28.52), and 224.14 (36.16) s on average (*SD*s in parentheses) in the four-item part, 10-item part, and 20-item whole learning groups, respectively. The difference was not statistically significant, $F(2, 90) = 0.63$, $p = .535$, and a very small effect size was found ($\eta^2 < .001$). Hence, it may be possible to assume that the three groups had roughly equivalent spacing whether time or trial was used as the index of spacing.

Table 2 (top) summarizes the number of correct responses for the four retrieval attempts during the treatment. To test the assumption that part learning produces more correct retrievals than whole learning during the treatment (see the "Theoretical Background" section), the number of correct responses on receptive and productive retrieval was submitted to two separate two-way 3 (treatment: four-item part, 10-item part, 20-item whole) × 2 (retrieval attempt: 1st or 2nd for receptive retrieval and 3rd or 4th for productive retrieval) ANOVAs. The ANOVA for receptive retrieval showed a significant main effect of treatment, $F(2, 88) = 4.44$, $p = .015$, $\eta_p^2 = .09$, and a significant interaction between the treatment and retrieval attempt, $F(1, 88) = 44.02$, $p < .001$, $\eta_p^2 = .50$. The ANOVA for productive retrieval detected a significant interaction between the treatment and retrieval attempt, $F(2, 88) = 6.96$, $p = .002$, $\eta_p^2 = .14$. The main effect of treatment was not significant on productive retrieval, $F(2, 88) = 1.97$, $p = .146$, $\eta_p^2 = .04$.

As the interaction between the treatment and retrieval attempt proved significant on both receptive and productive retrieval, the simple main effect of treatment was tested. The simple main effect of treatment was significant on the first, $F(2, 88) = 25.21$, $p < .001$, and third retrievals, $F(2, 88) = 4.46$, $p = .014$, but not on the second, $F(2, 88) = 0.31$, $p = .734$, and fourth retrievals, $F(2, 88) = 0.52$, $p = .598$. To follow up the significant simple main effect on the first and third retrievals, the Bonferroni method of multiple comparisons was used. The multiple comparisons showed that (a) on the first retrieval attempt, the four-item part group significantly

Table 2. Average number of correct responses during the learning phase

Experiment 1				
	Retrieval attempts			
Group	1st	2nd	3rd	4th
Four-item part	11.93	9.14	10.32	10.86
(n = 28)	*4.14*	*5.15*	*5.13*	*5.18*
10-item part	8.40	8.83	7.80	9.60
(n = 30)	*3.54*	*4.35*	*4.89*	*5.54*
Whole	5.58	9.73	6.70	9.64
(n = 33)	*2.75*	*4.27*	*4.41*	*5.34*
Experiment 2				
	Retrieval attempts			
Group	1st	2nd	3rd	4th
Control	10.62	14.92	17.19	17.88
(n = 26)	*4.45*	*4.13*	*3.10*	*2.83*
Four-item part	7.62	5.23	11.96	10.73
(n = 26)	*4.18*	*3.48*	*5.00*	*4.92*
Whole	4.31	7.23	10.04	12.77
(n = 26)	*2.40*	*3.81*	*4.10*	*4.35*

Note. Standard deviations in italics. The maximum score is 20 for each cell. Responses were scored with the strict scoring procedure. See the "Scoring" subsection of the Experiment 1 "Method" section.

outperformed the 10-item part (p = .001, d = 0.94) and 20-item whole groups (p < .001, d = 1.87), producing large effect sizes; (b) on the first retrieval attempt, the 10-item part group significantly outperformed the 20-item whole group (p = .005), producing a large effect size (d = 0.91); (c) the four-item part group significantly outperformed the 20-item whole group on the third retrieval attempt (p = .013), yielding a large effect size (d = 0.80); and (d) no statistically significant difference existed between the 10-item part group and the four-item part (p = .145, d = 0.53) or 20-item whole groups (p = 1.000, d = 0.25) on the third retrieval attempt, and medium or small effect sizes were found. Overall, the findings suggest that the four-item part group produced the largest number of correct responses during learning followed by the 10-item part group (four-item part > 10-item part > 20-item whole). The results are consistent with the assumption that part learning produces more correct retrievals than whole learning during the learning phase.

Posttest Performance. Table 3 provides the immediate and delayed posttest results for the three groups. Cronbach's alpha was .85 or higher

Table 3. Average number of correct responses on the posttests (Experiment 1)

	Immediate posttest				Delayed posttest			
	Productive		Receptive		Productive		Receptive	
Group	Strict	Sensitive	Strict	Sensitive	Strict	Sensitive	Strict	Sensitive
Four-item part	12.82	14.86	14.54	14.93	3.07	5.57	11.50	11.71
(n = 28)	*5.03*	*4.67*	*4.71*	*4.74*	*3.10*	*3.92*	*5.15*	*5.26*
10-item part	11.83	13.50	13.87	14.27	3.03	5.13	10.27	10.53
(n = 30)	*5.15*	*4.93*	*4.45*	*4.53*	*3.62*	*4.39*	*5.21*	*5.19*
Whole	11.88	13.30	14.39	14.70	3.06	4.58	9.61	9.82
(n = 33)	*6.24*	*6.08*	*5.12*	*5.23*	*3.67*	*4.18*	*5.63*	*5.65*

Note. Standard deviations in italics. The maximum score is 20 for each cell. Strict = strict scoring; Sensitive = sensitive scoring.

(.85–.91) for all tests, indicating good reliability. The productive and receptive test scores were analyzed by four separate two-way 3 (treatment: four-item part, 10-item part, 20-item whole) × 2 (retention interval: immediate, 1-week delayed) ANOVAs. To test whether the results could be generalized beyond the specific participants and items used in this study, both F_1 (participants) and F_2 (items) analyses were performed, and it was assumed that an effect was statistically significant only when both F_1 and F_2 produced statistically significant results (e.g., Raaijmakers, Schrijnemakers, & Gremmen, 1999). Table 4 shows the results of the ANOVAs. Although the F_2 analysis revealed several statistically significant results, the F_1 analysis found that, regardless of the posttest (productive or receptive) or scoring system (strict or sensitive), neither the main effect of treatment nor the interaction between the treatment and retention interval reached statistical significance. Thus, the results indicate that the part-whole learning distinction had little effect on posttest performance.

Discussion

Results of Experiment 1 indicated that, although part learning produced more correct retrievals than whole learning during the learning phase, there was little difference in their posttest scores. The findings are at odds with most previous studies on part and whole learning, which have found whole learning to contribute to greater learning than part learning (Brown, 1924; Crothers & Suppes, 1967; Kornell, 2009; McGeoch, 1931; Seibert, 1932). There are three explanations for the contradictory results. First, the inconsistent findings may stem from a methodological difference.

Table 4. Results of two-way ANOVAs for the posttest scores (Experiment 1)

			Strict scoring				Sensitive scoring			
Posttest	Effect		*df*	*F*	*p*	η_p^2	*df*	*F*	*p*	η_p^2
Productive	Treatment	F_1	2, 88	0.14	.867	.01	2, 88	0.69	.505	.02
		F_2	2, 38	2.61	.087	.12	2, 38	11.47	< .001	.38
	Treatment × Retention interval	F_1	2, 88	0.45	.638	.01	2, 88	0.36	.698	.01
		F_2	2, 38	1.73	.190	.08	2, 38	1.32	.279	.06
Receptive	Treatment	F_1	2, 88	0.41	.664	.01	2, 88	0.42	.660	.01
		F_2	2, 38	5.12	.011	.21	2, 38	5.98	.006	.24
	Treatment × Retention interval	F_1	2, 88	2.24	.113	.05	2, 88	1.97	.145	.04
		F_2	2, 38	5.27	.010	.22	2, 38	5.82	.006	.23

Note. F_1 = ANOVA by participants; F_2 = ANOVA by items.

In all earlier studies where target words were encountered more than once, the part-whole learning distinction and spacing were confounded. In this experiment, however, part and whole learning had equivalent spacing. The findings of the present experiment suggest that, as long as spacing is equivalent, the part-whole distinction has little effect on learning.

Another possible cause for the lack of statistical significance may be the relatively limited range of block sizes used. In this experiment, the block sizes of four, 10, and 20 words were chosen based on previous studies (Brown, 1924; Kornell, 2009; McGeoch, 1931; Seibert, 1932). However, considering that the present and previous studies differed in several factors such as the participants, materials, or posttest format, the findings of the earlier research may not necessarily be applicable to the present experiment. As a result, a wider range of block sizes (e.g., four, 20, and 60 words) may have been necessary to yield statistical significance.

A third explanation is that the task difficulty was too high in this experiment. Previous studies have suggested that whole learning may be superior to part learning only when difficulty is low (Crothers & Suppes, 1967; Nation, 2013). Because the present experiment used a recall format rather than a recognition (multiple-choice) format, the task difficulty may have been relatively high. This could be, in part, responsible for the lack of significant differences among the three treatments.

Limitations

The findings of Experiment 1 have pedagogical significance because they imply that, as long as spacing is equivalent, the part-whole learning

distinction has little effect on learning. However, a possible limitation of the experiment is that the lack of statistical significance cannot necessarily be attributed to the fact that part and whole learning were controlled for spacing. To argue that equivalent spacing was responsible for the inconsistent results between Experiment 1 and earlier studies, the following three hypotheses need to be supported: (1) When spacing is equivalent, whole learning does not outperform part learning, (2) when whole learning has longer spacing than part learning, whole learning outperforms part learning, and (3) part learning with longer spacing outperforms part learning with shorter spacing. Although the results of Experiment 1 were consistent with the first hypothesis, the latter two were not investigated. None of the previous studies have examined these three hypotheses either. Unless the second and third hypotheses are also verified, we cannot necessarily rule out the possibility that some factors other than spacing, such as the limited range of block sizes used or high task difficulty, were responsible for the lack of a significant effect.

EXPERIMENT 2

The purpose of Experiment 2 was to test the three hypotheses put forward in the previous section. If all three hypotheses are supported, it would suggest that (a) spacing has a larger effect on learning than the part-whole distinction and (b) the lack of statistical significance in Experiment 1 was because part and whole learning had equivalent spacing. By testing the three hypotheses, Experiment 2 may allow us to examine the relative importance of the part-whole distinction and spacing on L2 vocabulary learning.

Method

The following three treatments were compared in this experiment: control, four-item part learning, and whole learning. The four-item part learning and whole learning treatments were exactly the same as in Experiment 1. They had different block sizes (four and 20), but had equivalent spacing (19 trials). Hereafter, the four-item part learning and whole learning treatments will be collectively referred to as the *experimental treatments*. The control treatment also used a block size of four items but had shorter spacing (three trials) than the other two. The 10-item part learning treatment was not used in this experiment because (a) the purpose of the current experiment can be achieved by comparing only the control, four-item part learning, and whole learning treatments and

(b) Experiment 1 found no significant difference among the four-item part, 10-item part, and 20-item whole learning in their effectiveness.

Participants. The participants were 78 first-year Japanese students at the same university as in Experiment 1. The participants consisted of 39 engineering and 39 economics majors. Their average score on the first to the sixth 1,000-word frequency levels of the VST (Nation & Beglar, 2007) was 29.83 (*SD* = 6.27) out of 60. The participants were assigned to the control, four-item part learning, and whole learning groups so that there would be no significant difference in the VST scores, $F(2, 77) = 0.03$, $p = .972$, $\eta^2 < .001$. Each group consisted of 26 participants. The three groups also had a roughly equal number of engineering and economics majors. None of the participants exhibited prior knowledge of any of the target words on the productive pretest (see the "Procedure and Materials" section for details about the pretest). The average scores on the receptive pretest (*SD*s in parentheses) were 0.27 (.45), 0.31 (.55), and 0.19 (.57) out of 20 with strict scoring and 0.27 (.45), 0.35 (.56), and 0.19 (.57) out of 20 with sensitive scoring in the control, four-item, and whole learning groups, respectively.

Procedure and Materials. The methodology of Experiment 2 differed from that of Experiment 1 in three respects. First, although the target words were practiced in both receptive and productive recall formats during the treatment in Experiment 1, Experiment 2 involved only productive retrieval. This was done to control the position of initial productive retrieval during learning. In Experiment 1, the first and second retrievals (second and third encounters) used the receptive recall format, and the third and fourth retrievals (fourth and fifth encounters) used the productive recall format (see the "Treatment" subsection in the "Experiment 1" section). If the same procedure were employed in Experiment 2, the three treatments would differ greatly in the position of initial productive retrieval. Specifically, although target words would be practiced productively for the first time at the beginning of Cycle 4 (24th trial out of 114; see Table 5) in the control group, the productive format would not be used until the beginning of Cycle 12 (56th trial) and Cycle 4 (72nd trial) in the four-item part and whole groups, respectively (marked with an asterisk in Table 1). The difference in the position of initial productive retrieval may be problematic because it may affect how much attention participants pay to the spelling of target words during the treatment. More specifically, exposure to productive retrieval earlier in the control treatment may encourage the control group to pay close attention to the spelling of target words earlier than the other two groups, potentially leading to higher gains in productive knowledge.

Table 5. Item order and spacing in Experiment 2

20-item whole learning group

Cycle	Items	Practice format	Spacing	Cycle	Items	Practice format	Spacing
Primacy	11 fillers		-	Cycle 4	1–20	Productive	19
Cycle 1	1–20	Presentation	19	Cycle 5	1–20	Productive	-
Cycle 2	1–20	Productive	19	Recency	3 fillers		-
Cycle 3	1–20	Productive	19	Average			19

Four-item part learning group

Cycle	Items	Format	Spacing	Cycle	Items	Format	Spacing
Primacy	3 fillers		-	Cycle 11	1–4	Productive	3
Cycle 1	1–4	Presentation	3	Cycle 12	1–4	Productive	35
Cycle 2	1–4	Productive	43	Cycle 13	5–8	Productive	3
Cycle 3	5–8	Presentation	3	Cycle 14	5–8	Productive	31
Cycle 4	5–8	Productive	43	Cycle 15	9–12	Productive	3
Cycle 5	9–12	Presentation	3	Cycle 16	9–12	Productive	27
Cycle 6	9–12	Productive	43	Cycle 17	13–16	Productive	3
Cycle 7	13–16	Presentation	3	Cycle 18	13–16	Productive	23
Cycle 8	13–16	Productive	43	Cycle 19	17–20	Productive	3
Cycle 9	17–20	Presentation	3	Cycle 20	17–20	Productive	19
Cycle 10	17–20	Productive	43	Review	1–20	Productive	-
Filler	8 fillers		-	Recency	3 fillers		-
				Average			19

Control group

Cycle	Items	Practice format	Spacing	Cycle	Items	Practice format	Spacing
Primacy	11 fillers		-	Cycle 14	9–12	Productive	3
Cycle 1	1–4	Presentation	3	Cycle 15	9–12	Productive	-
Cycle 2	1–4	Productive	3	Cycle 16	13–16	Presentation	3
Cycle 3	1–4	Productive	3	Cycle 17	13–16	Productive	3
Cycle 4	1–4	Productive	3	Cycle 18	13–16	Productive	3
Cycle 5	1–4	Productive	-	Cycle 19	13–16	Productive	3
Cycle 6	5–8	Presentation	3	Cycle 20	13–16	Productive	-
Cycle 7	5–8	Productive	3	Cycle 21	17–20	Presentation	3
Cycle 8	5–8	Productive	3	Cycle 22	17–20	Productive	3
Cycle 9	5–8	Productive	3	Cycle 23	17–20	Productive	3
Cycle 10	5–8	Productive	-	Cycle 24	17–20	Productive	3
Cycle 11	9–12	Presentation	3	Cycle 25	17–20	Productive	-
Cycle 12	9–12	Productive	3	Recency	3 fillers		-
Cycle 13	9–12	Productive	3	Average			3

Note. Average refers to the average spacing (mean intervening trials) for a given target word pair when collapsed across all cycles.

In Experiment 2, therefore, the initial position of productive retrieval was controlled by taking out receptive retrieval and using only productive retrieval. By so doing, the position of initial productive retrieval was roughly equivalent in all three groups: the eighth trial in the four-item part group (beginning of Cycle 2 in Table 5) and the fourth trial in the other two groups (fourth filler trial; Table 5). Productive, not receptive, retrieval was used because earlier studies indicate that productive retrieval is more effective than receptive retrieval because it results in adequate gains in receptive knowledge as well as large gains in productive knowledge (Griffin & Harley, 1996; Mondria & Wiersma, 2004; Steinel, Hulstijn, & Steinel, 2007; Webb, 2009).

Second, whereas only a receptive pretest was given in Experiment 1, a productive pretest was given prior to the receptive pretest in Experiment 2. In the productive pretest, participants were presented with Japanese (L1) meanings and needed to type the corresponding English target words. The productive pretest was also conducted because scores on the productive posttest were the main dependent variables in Experiment 2: As target words were practiced only in productive recall in Experiment 2, scores on the productive posttest, which used exactly the same format as the productive recall format during learning, may be a more direct and reliable measure of learning outcomes than those on the receptive posttest. Hence, it was decided to measure productive as well as receptive knowledge in the pretest. In the productive pretest, it was important to prevent participants from providing synonyms for target words because if, for instance, participants typed *hair* for the target word *mane*, it would not be clear whether or not they knew *mane* (Barcroft & Rott, 2010). To prevent learners from responding with synonyms, the number of letters and one letter in the target words (e.g., _ _ *n* _ for *mane*) were provided together with the Japanese translation in the pretest (see Nakata, 2013, for the protocol to determine the hints).

Third, the filler items used in Experiment 1 (*husk*, *polemic*, and *smudge*) were replaced with *promontory*, *urn*, and *vestige*. This is because *rue*, *citadel*, and *apparition* were the only three-, seven-, and 10-letter target items, respectively, and adding a filler item of the same length could minimize effects that the productive pretest would have on performance on the receptive pretest, which was given immediately after the productive pretest. More specifically, when learners are given 後悔する (_ *u* _) as a cue for *rue* in the productive pretest, they may be able to infer that a three-letter item used in the experiment means 後悔する (*rue*). Without any other three-letter item except *rue*, participants may have a relatively high chance of answering correctly on this item on the receptive pretest without having any prior knowledge. Another three-letter word (*urn*), therefore, was added as a filler item in Experiment 2. With the addition of *urn*, learners may not know whether *rue* or *urn* means 後悔する (*rue*), which may minimize effects on receptive pretest performance.

Similarly, as *citadel* and *apparition* were the only seven- and 10-letter target items, seven- and 10-letter filler items (*vestige* and *promontory*) were added in Experiment 2. Although *polemic*, a filler item used in Experiment 1, also consists of seven letters, it was not used in Experiment 2 because using two filler items beginning with *p* (*promontory* and *polemic*) may affect learning. Other than these three differences, the methodology of Experiment 2 was exactly the same as in Experiment 1.

Treatment. Table 5 summarizes the item order in the three treatments in Experiment 2. Table 5 should be read in the same way as Table 1. In the control treatment, target words were repeated in five blocks of four items, and encounters of a given item were separated by three trials on average throughout the treatment. As a result, the control treatment had shorter spacing (three trials) than the two experimental treatments (19 trials). As in the whole learning treatment, there were 11 primacy and three recency buffers in the control treatment. Unlike in the four-item and whole learning treatments, the final review was not used in the control treatment. This is because adding the final review would increase spacing in the control treatment from three to 13 trials, which is not very different from the spacing in the experimental treatments (19 trials). Because adding the final review might have made it difficult to test Hypotheses 2 and 3 by increasing spacing in the control treatment, the final review was not used in the control treatment.

At the same time, because the final review was not used, the first four blocks of items (items 1–16) in the control treatment had a greater lag to test time than in the experimental treatments (see Table 5). To ascertain to what extent differential lag to test times affected learning, a possible relationship between the posttest performance and lag to test was analyzed by a three-way 2 (treatment: control, four-item) × 5 (block during learning: 1, 2, 3, 4, 5) × 2 (retention interval: immediate, delayed) ANOVA. This analysis showed that the advantage of the four-item group was not significantly larger for items initially studied in earlier blocks, suggesting that lag to test had little effect on the posttest scores (see Nakata, 2013, for details).

Results

Learning Phase Data. The participants spent 17.87 (3.02), 17.56 (2.13), and 17.34 (2.42) min on average (*SD*s in parentheses) studying the target items in the control, four-item part, and whole learning groups, respectively. No statistically significant difference was found among the three groups in study time, $F(2, 77) = 0.28$, $p = .755$, and a very small effect size was found ($\eta^2 < .001$). On the basis of these results, it may

be possible to assume that the average study time was roughly equivalent among the three groups.

The amount of time that intervened between the repetitions of target words was also investigated to measure spacing as a function of time. On average, repetitions of a given item were separated by 32.42 (5.38), 198.47 (27.15), and 195.65 (32.76) s in the control, four-item, and whole learning groups, respectively (*SD*s in parentheses). The difference in the time between the repetitions reflects the difference in the number of intervening trials (control: three trials; four-item and whole learning: 19 trials; see Table 5). A one-way ANOVA found a statistically significant difference among the three groups, $F(2, 77) = 383.32$, $p < .001$, $\eta^2 = .83$. The Bonferroni method of multiple comparisons showed that even when time was used as an index of spacing, (a) the control group had significantly shorter spacing than the four-item ($p < .001$, $d = 8.48$) and whole learning groups ($p < .001$, $d = 6.95$) and (b) the two experimental groups could be considered as having roughly equivalent spacing ($p = 1.000$, $d = 0.09$).

Table 2 (bottom) summarizes the number of correct responses for the four retrieval attempts during the treatment. To test the assumption that part learning produces more correct retrievals than whole learning during the treatment, the number of correct responses during learning was submitted to a two-way 3 (treatment: control, four-item, whole) × 4 (retrieval attempt: 1st, 2nd, 3rd, 4th) ANOVA. The ANOVA detected a significant main effect of treatment, $F(2, 76) = 26.47$, $p < .001$, $\eta_p^2 = .41$. The interaction between the treatment and retrieval attempt was also significant, $F(4.76, 178.63) = 24.18$, $p < .001$, $\eta_p^2 = .39$.[5]

Due to the significant interaction between the treatment and retrieval attempt, the simple main effect of treatment was tested to investigate where the significant differences lay. The simple main effect of treatment was significant on all four retrieval attempts, first: $F(2, 75) = 18.04$, $p < .001$; second: $F(2, 75) = 46.80$, $p < .001$; third: $F(2, 75) = 20.78$, $p < .001$; fourth: $F(2, 75) = 20.72$, $p < .001$. To follow up the significant simple main effect, the Bonferroni method of multiple comparisons was used. The multiple comparisons indicated the following three things. First, the control group significantly outperformed the four-item (first retrieval: $p = .017$, $d = 0.69$; second: $p < .001$, $d = 2.54$; third: $p < .001$, $d = 1.26$; fourth: $p < .001$, $d = 1.78$) and whole groups ($p < .001$ for all retrievals; first: $d = 1.76$; second: $d = 1.94$; third: $d = 1.97$; fourth: $d = 1.39$) on all four retrieval attempts, and medium to large effect sizes were found. Second, on the first retrieval, the four-item group fared significantly better than the whole group, showing a large effect size ($p = .007$, $d = 0.97$). Third, the differences between the four-item and whole groups were not statistically significant on all other retrievals (second: $p = .188$, $d = 0.55$; third: $p = .295$, $d = 0.42$; fourth: $p = .237$, $d = 0.44$), and only small to medium effect sizes were observed. Overall, the findings suggest that the control

group produced the largest number of correct responses during learning followed by the four-item group (control > four-item > whole).

Posttest Performance. Table 6 provides the immediate and delayed posttest results for the three groups. Cronbach's alpha was .73 or higher (.73–.90) for all tests, indicating good reliability. The productive and receptive test scores were analyzed by a two-way 3 (treatment: control, four-item, whole) × 2 (retention interval: immediate, delayed) ANOVA. Because some items were answered correctly on the receptive pretest (see the "Participants" section), gains (pretest scores subtracted from the posttest scores) were analyzed when examining the receptive test results. Table 7 shows the results of the ANOVAs. The F_1 analysis revealed that the main effect of treatment was statistically significant with strict scoring on the productive posttest and approached significance with sensitive scoring on the productive posttest and with both scoring procedures on the receptive posttest. According to the F_1 analysis, the interaction between the treatment and retention interval also approached statistical significance with sensitive scoring on the productive posttest but was not significant with strict scoring on the productive posttest or with strict or sensitive scoring on the receptive posttest. The F_2 analysis found (a) a significant main effect of treatment on all dependent variables and (b) a significant interaction between the treatment and retention interval on the productive posttest but not on the receptive posttest regardless of the scoring procedure.

The Bonferroni method of multiple comparisons was used to investigate where the significant differences lay. The multiple comparisons indicated the following three things (see Appendix S2 in the online supplementary material for detailed results of the multiple comparisons). First, on the delayed productive posttest, the experimental

Table 6. Average number of correct responses on the posttests (Experiment 2)

	Immediate posttest				Delayed posttest			
	Productive		Receptive		Productive		Receptive	
Group	Strict	Sensitive	Strict	Sensitive	Strict	Sensitive	Strict	Sensitive
Control	11.69	13.77	11.81	12.38	1.62	3.00	6.81	7.08
	4.33	*4.07*	*4.78*	*4.95*	*1.92*	*2.65*	*4.72*	*4.86*
Four-item part	12.31	14.12	13.58	14.12	4.04	5.81	9.54	9.85
	5.27	*4.57*	*4.15*	*4.05*	*3.42*	*3.90*	*5.27*	*5.38*
Whole	13.92	15.27	14.46	15.12	3.73	6.04	9.62	10.04
	4.82	*4.40*	*4.53*	*4.60*	*2.20*	*3.61*	*5.59*	*5.77*

Note. Standard deviations in italics. n = 26 for each group. The maximum score is 20 for each cell.

Table 7. Results of two-way ANOVAs for the posttest scores (Experiment 2)

			Strict scoring				Sensitive scoring			
Posttest	Effect		*df*	*F*	*p*	η_p^2	*df*	*F*	*p*	η_p^2
Productive	Treatment	F_1	2, 75	3.13	.049	.08	2, 75	3.10	.051	.08
		F_2	2, 38	20.01	< .001	.51	2, 38	18.74	< .001	.50
	Treatment × Retention interval	F_1	2, 75	1.61	.207	.04	2, 75	2.52	.088	.06
		F_2	2, 38	6.46	.004	.25	2, 38	7.23	.002	.28
Receptive	Treatment	F_1	2, 75	2.94	.059	.07	2, 75	3.00	.056	.07
		F_2	2, 38	38.25	< .001	.67	2, 38	37.73	< .001	.67
	Treatment × Retention interval	F_1	2, 75	0.57	.569	.01	2, 75	0.55	.578	.01
		F_2	2, 38	1.50	.236	.07	2, 38	1.81	.177	.09

Note. F_1 = ANOVA by participants; F_2 = ANOVA by items.

groups significantly outperformed the control group with both scoring methods, and large effect sizes were observed ($.84 \leq d \leq 1.02$). Second, no significant difference existed among the three groups for all other comparisons, showing medium or smaller effect sizes ($.04 \leq d \leq 0.60$). Third, the difference between the experimental groups was rather small on both productive and receptive tests regardless of the scoring procedure as indicated by the lack of statistical significance as well as the no more than small effect sizes ($.04 \leq d \leq 0.32$). In summary, the posttest scores indicate that (a) the two experimental groups significantly outperformed the control group on the delayed productive posttest but not on the other posttests, and (b) no significant difference existed between the experimental groups, mirroring the findings of Experiment 1.

Discussion

Experiment 2 demonstrated the superiority of the four-item part and whole learning groups over the control group. The advantage was particularly large on the delayed productive posttest, on which the experimental groups significantly outperformed the control group, producing large effect sizes. The difference between the experimental groups was relatively small regardless of the posttest, scoring system, or retention interval, mirroring the results of Experiment 1. The findings of the current experiment seem to support all three hypotheses put forward at the end of Experiment 1. Hypothesis 1 was supported because when spacing was equivalent, whole learning did not outperform (four-item) part learning. Hypothesis 2 was also confirmed because

whole learning outperformed part learning with shorter spacing (i.e., the control treatment). The results were also congruent with Hypothesis 3 as part learning with longer spacing (i.e., four-item part learning) fared significantly better than part learning with shorter spacing (i.e., the control).

Although the present experiment demonstrated the advantage of the experimental groups, the results were not consistent across the retention intervals or posttests. The experimental groups fared significantly better than the control group on the delayed productive posttest but not on the immediate productive test. The results may be explained in part by the *spacing by retention interval interaction*, which refers to a phenomenon whereby shorter spacing is typically effective at short retention intervals, but longer spacing is typically effective at longer retention intervals (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Cepeda et al., 2008; Cepeda et al., 2009; Pashler et al., 2007; Rohrer & Pashler, 2007). Due to this interaction, the treatments with longer spacing (i.e., the experimental treatments) may have been particularly effective on the delayed posttest.[6]

The effects of the treatments were also conditional upon the type of posttest. Although the superiority of the experimental groups was found on the productive posttest, no significant difference existed among the three groups on the receptive test. The results could be partially due to the test order. At each retention interval, the productive test was administered prior to the receptive test. Because correct responses on the receptive test were used as cues in the productive test, performance on the receptive test may have been affected by first completing the productive test. This may have possibly reduced a potential difference among the three groups on the receptive posttest.[7] Alternatively, the direction of learning could be partially responsible for the results. In Experiment 2, target words were practiced only in a productive format. Because productive learning may have a greater effect on productive tests than receptive tests, significant differences were perhaps found only on the productive posttest.

PEDAGOGICAL IMPLICATIONS

The results of the present study indicate that (a) learners may study with either part or whole learning without significantly affecting learning outcomes and (b) it is useful to pay more attention to spacing rather than the part-whole learning distinction. The pedagogical implications of the findings are useful because, although spacing has a large effect on vocabulary learning, its value has not been exploited fully in typical instructional settings (e.g., Cepeda et al., 2009; Ellis, 1995). Moreover, learners may not be aware that spacing facilitates learning

(e.g., Kornell, 2009; Wissman et al., 2012). The results of this study suggest that it may be useful to raise awareness of the value of spacing.

Although the part-whole distinction was found to have little effect on posttest performance, the present study nonetheless suggested possible advantages and disadvantages of part and whole learning. One benefit of part learning may be that it increases learning phase performance and, thus, motivates learners. In both experiments, part learning produced significantly more correct responses during learning than whole learning (Experiment 1: four-item part > 10-item part > 20-item whole; Experiment 2: four-item part > whole). As incorrect responses during learning may potentially demotivate learners (e.g., Logan & Balota, 2008), the use of part learning may be more desirable. A disadvantage of part learning, however, is that it may possibly lead to underlearning. That is, a high probability of retrieval success caused by part learning may create what Kornell (2009) refers to as "an illusion of effective learning" (p. 1302), and learners may stop studying before lexical items are actually acquired, resulting in underlearning.

Another implication of this study is that learning phase performance is not necessarily a good measure of long-term retention (e.g., Bjork, 1994; Ellis, 1995). In Experiment 1, although the 4-item group led to the best learning phase performance, no statistically significant difference was found among the three groups in their posttest scores. Similarly, in Experiment 2, the control treatment, which produced the largest number of correct responses during learning, turned out to be the least effective 1 week after the treatment. The findings contradict the retrieval practice effect (Baddeley, 1997; Ellis, 1995) but are supported by the desirable difficulty framework (e.g., Bjork, 1994), which suggests that a treatment that increases the rate of acquisition initially does not always facilitate long-term retention. Pedagogically, the results indicate that it is important to raise awareness that mistakes made during learning are not necessarily a sign of ineffective learning (e.g., Karpicke & Roediger, 2007; Logan & Balota, 2008; Nakata, 2015).

SUMMARY AND CONCLUDING DISCUSSION

The purpose of this study was to examine the effects of part and whole learning on L2 vocabulary acquisition. Experiment 1 found little difference between part and whole learning in their effectiveness. Experiment 2 demonstrated that whole learning is more effective than part learning only when the former has larger spacing. Taken together, the two experiments indicate that, (a) as long as spacing is equivalent, the part-whole distinction has little effect on learning (hence, four-item part = 10-item part = 20-item whole in Experiment 1 and four-item = whole in Experiment 2), and (b) spacing has a larger effect on learning than

the part-whole learning distinction (hence, four-item part = whole > control in Experiment 2). These findings have value because they suggest that the results of the earlier studies may be attributed to spacing rather than the part-whole distinction. Pedagogically, the findings indicate that introducing a large amount of spacing is more important than choosing between part or whole learning.

The present study may be methodologically significant in that it isolated the effects of the part-whole distinction and spacing, which have been confounded in previous research. However, it is important to note that there are several limitations to the research. One limitation is the rather short duration of the treatments. Although study opportunities tend to be distributed over multiple sessions in a real-life study situation (Cepeda et al., 2008), they were massed into a single treatment session in this study. In future research, it may be useful to investigate the effects of part and whole learning over a longer period of time. Another limitation of the present study is that learning was restricted to a paired-associate learning condition. Further research investigating the effects of part and whole learning in other learning conditions would be a useful follow-up to this study.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit http://dx.doi.org/S0272263115000236.

Received 19 March 2014
Accepted 21 January 2015
Final Version Received 27 January 2015

NOTES

1. Note that some earlier studies have yielded inconsistent results regarding the retrieval practice effect. For instance, some studies have shown that a treatment that produces a large number of unsuccessful retrievals during learning can sometimes lead to superior long-term retention, contradicting the retrieval practice effect (e.g., Karpicke & Roediger, 2007; Logan & Balota, 2008; Nakata, 2015). This may be especially true if the correct response is provided as feedback after retrieval presumably because feedback may allow learners to correct errors on subsequent retrievals (e.g., Pashler et al., 2003).

2. One anonymous reviewer has pointed out that a U-shaped relationship between block size and retention may exist. Crothers and Suppes (1967) may offer support for this view: Although they found the advantage of a block size of 216 over that of 108 in their Experiment 9, they failed to show any significant difference between block sizes of 100 and 300 in Experiment 10. Their findings suggest that (a) there may be a threshold beyond which the benefits of increasing a block size diminish, and (b) the threshold may lie somewhere between 216 and 300. At the same time, it is also possible that the lack of significant difference between block sizes of 100 and 300 in Crothers and Suppes's Experiment 10 was because part and whole learning may have interacted with task difficulty (Nation, 2013; see the "Review of Literature" section). Although the issue of a possible U-shaped relationship between block size and retention is interesting, it was not addressed in this study because

using block sizes of around 300 words may be neither ecologically valid nor practical (e.g., Kornell, 2009; Nakata, 2011; Salisbury & Klein, 1988; Wissman et al., 2012; Woodworth & Schlosberg, 1954).

3. Several earlier studies are aware of and explicitly mention this confound (Kornell, 2009; McGeoch, 1931; Van Bussel, 1994; Woodworth & Schlosberg, 1954). Kornell (2009), for instance, argues that whole learning is an effective strategy because it helps to introduce a large amount of spacing between encounters.

4. As one anonymous reviewer points out, some of the target words have cognates in other languages such as French (e.g., *citadel* and *fracas*). However, because participants in this study had little or no prior knowledge of languages other than Japanese (their L1) and English (their L2), the use of these words probably did not have much effect on the results of this study. Please note also that participants who demonstrated prior knowledge of one or more target words on the pretest were excluded from analysis (see the "Participants" subsection of the "Experiment 1" section).

5. As Mauchly's test showed that sphericity assumptions were violated, the Greenhouse-Geiser correction was used. As a result, the degrees of freedom for the interaction between the treatment and retrieval attempt contain decimal values.

6. One anonymous reviewer has pointed out that we may expect little or no difference between the effects of the control and experimental treatments on the delayed posttest because some studies failed to find any significant difference between short and long spacing when the spacing to retention interval ratios are small (e.g., Crothers & Suppes, 1967; Hausman & Kornell, 2014; Logan & Balota, 2008). However, previous L2 vocabulary studies have demonstrated the superiority of long over short spacing with relatively small spacing to retention interval ratios (e.g., Bahrick & Phelps, 1987; Karpicke & Bauernschmidt, 2011; Pashler et al., 2003; Pyc & Rawson, 2012). Please note also that the findings of this study are consistent with the results of earlier studies showing that when the spacing to retention interval ratios are smaller than 10 to 30%, increasing spacing increases retention (e.g., Cepeda et al., 2006; Cepeda et al., 2008; Cepeda et al., 2009; Pashler et al., 2007; Rohrer & Pashler, 2007).

7. The authors are grateful to an anonymous reviewer for pointing this out.

REFERENCES

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

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

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

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

Brown, W. (1924). Whole and part methods in learning. *Journal of Educational Psychology*, *15*, 229–233.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236–246.

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*, 693–713.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Crothers, E., & Suppes, P. (1967). *Experiments in second language learning*. New York, NY: Academic Press.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*, 367–413.
Ellis, N. C. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, *8*, 103–128.
Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188.
Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: A case study. *Language Learning Journal*, *36*, 239–248.
Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67.
Griffin, G. F., & Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, *17*, 443–460.
Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in Memory and Cognition*, *3*, 153–160.
Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal, and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge, UK: Cambridge University Press.
Hulstijn, J. H. (2002). What does the impact of frequency tell us about the language acquisition device? *Studies in Second Language Acquisition*, *24*, 269–273.
Joseph, S., Watanabe, Y., Shiung, Y.-J., Choi, B., & Robbins, C. (2009). Key aspects of computer assisted vocabulary learning (CAVL): Combined effects of media, sequencing and task type. *Research and Practice in Technology Enhanced Learning*, *4*, 1–36.
Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, *21*, 1544–1550.
Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1250–1257.
Karpicke, J. D., & Roediger, H. L. (2007). 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. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968.
Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297–1317.
Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London, UK: Academic Press.
Larsen-Freeman, D. (2002). Making sense of frequency. *Studies in Second Language Acquisition*, *24*, 275–285.
Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, *15*, 257–280.
McGeoch, G. O. (1931). The intelligence quotient as a factor in the whole-part problem. *Journal of Experimental Psychology*, *14*, 333–358.
Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*, 1077–1087.
Mondria, J.-A., & Wiersma, B. (2004). Receptive, productive and receptive + productive L2 vocabulary learning: What difference does it make? In B. Laufer (Ed.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 79–100). Amsterdam, The Netherlands: Benjamins.
Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*, 17–38.
Nakata, T. (2013). *Optimising second language vocabulary learning from flashcards* (Unpublished doctoral dissertation). Victoria University of Wellington, New Zealand.
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. Advance online publication. doi: 10.1017/S0272263114000825.
Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*, 59–82.
Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge, UK: Cambridge University Press.
Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.
Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin and Review*, *14*, 187–193.
Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1051–1057.
Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*, 1–28.
Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*, 1917–1927.
Pyc, M. A., & Rawson, K. A. (2012). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition*, *40*, 976–988.
Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416–426.
Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, *16*, 183–186.
Salisbury, D. F., & Klein, J. D. (1988). A comparison of a microcomputer progressive state drill and flashcards for learning paired associates. *Journal of Computer-Based Instruction*, *15*, 136–143.
Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 199–227). Cambridge, UK: Cambridge University Press.
Seibert, L. C. (1932). *A series of experiments on the learning of French vocabulary*. Baltimore, MD: The Johns Hopkins Press.
Steinel, M. P., Hulstijn, J. H., & Steinel, W. (2007). Second language idiom learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency. *Studies in Second Language Acquisition*, *29*, 449–484.
Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, *38*, 244–253.
Van Bussel, F. J. J. (1994). Design rules for computer-aided learning of vocabulary items in a second language. *Computers in Human Behavior*, *10*, 63–76.
Webb, S. A. (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. A. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79–95.
Webb, S. A. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review*, *65*, 441–470.
Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*, 568–579.
Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology* (3rd ed.). London, UK: Methuen & Co.
Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, *57*, 541–572.