Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 16 (2). 2026. 363-393. Published online: 08.06.2026
https://doi.org/10.14746/ssllt.49487
http://pressto.amu.edu.pl/index.php/ssllt

Effects of spacing on the acquisition of explicit and implicit vocabulary knowledge: An approximate replication of Nakata and Elgort (2021)

Guangliang Zheng ✉
Hiroshima University, Hiroshima, Japan
https://orcid.org/0009-0004-5748-877X
zhengcnu@yeah.net

Tatsuya Nakata
Rikkyo University, Tokyo, Japan
https://orcid.org/0000-0002-1152-653X
nakata@rikkyo.ac.jp

Jon Clenton
Hiroshima University, Hiroshima, Japan
https://orcid.org/0000-0002-3048-8807
jclenton@hiroshima-u.ac.jp

TJ Boutorwick
Kindai University, Hiroshima, Japan
https://orcid.org/0000-0002-5244-512X
tj@hiro.kindai.ac.jp

Abstract

Research suggests that spaced learning, which incorporates intervals between repetitions of material, facilitates second language (L2) learning more than massed learning, where a given item is repeated multiple times without any intervening trials or time. Although prior studies have suggested that spacing enhances the

acquisition of explicit knowledge, it remains unclear whether it aids the acquisition of implicit knowledge. The present study replicated Nakata and Elgort's (2021) study to investigate how massing and spacing impact the acquisition of explicit and implicit vocabulary knowledge in L2. In the present study, 69 Japanese students learning (L2) English studied 48 pseudowords using either massed or spaced repetition. A meaning recall test (immediate and delayed) and a delayed meaning-form matching test assessed participants' explicit knowledge. A semantic priming task (immediate and delayed) measured implicit knowledge. Posttest results showed that spaced learning led to better explicit vocabulary acquisition compared to massed learning. For implicit knowledge development, however, (a) neither schedule was effective, and (b) no significant difference was found between the two schedules.

Keywords: vocabulary learning; spacing effect; spaced/massed practice; explicit/implicit vocabulary knowledge; semantic priming

1. Introduction

Cognitive psychology research demonstrates that practice distribution influences retention (e.g., Wiseheart et al., 2019). Possibly because of its robustness (e.g., Cepeda et al., 2006; Wiseheart et al., 2019), the effects of practice distribution have attracted attention from many second language (L2) researchers (for reviews, see Kim & Webb, 2022; Serrano, 2022; Suzuki et al., 2023), who have attempted to enhance L2 learning by taking advantage of the spacing benefits. Existing studies have distinguished between two types of spacing: between-session and within-session. In the former, temporal spacing between separate treatment sessions is manipulated, whereas in the latter, spacing during a single treatment session is manipulated. Studies on between-session spacing typically use time as a unit of spacing (e.g., three days vs. seven days), while those on within-session spacing often manipulate the number of intervening trials (e.g., zero, two, vs. eight trials) between repetitions of a given item.

There are two related phenomena in practice distribution: the spacing effect and the lag effect (Cepeda et al., 2006). The *spacing effect* suggests that spaced learning, which involves repeated exposure to material after intervals, enhances retention more effectively than massed learning, where a given item is repeated multiple times without any intervening trials or time. A *lag effect*, in contrast, refers to the advantage of longer spacing over shorter spacing. While the results of L2 studies examining the lag effect, especially those manipulating between-session spacing, have been mixed (e.g., Kasprowicz et al., 2019; Li & DeKeyser, 2019; Rogers & Cheung, 2021; Serrano & Pellicer-Sánchez, 2024; Suzuki, 2017; Suzuki & DeKeyser, 2017), most L2 studies on the spacing effect have shown the advantage of spaced learning over massed learning (for a meta-analysis, see Kim & Webb, 2022).

The spacing effect is generally more robust and reliable than the lag effect, but it remains unclear whether it aids the acquisition of not only explicit (declarative) but also procedural knowledge (in this paper, *explicit knowledge* and *declarative knowledge* are used synonymously; DeKeyser & Suzuki, 2025). Explicit knowledge refers to the type of knowledge that is consciously available (e.g., knowing that the Japanese word *ringo* means "apple"), which is typically acquired through intentional efforts, such as memorizing vocabulary or looking up unfamiliar words during reading (Ellis, 2015). Procedural knowledge, in contrast, refers to knowledge that allows for fluent, rapid, and efficient use of linguistic features (e.g., being able to comprehend or produce the Japanese word *ringo* in context). One pathway to procedural knowledge is the automatization of explicit knowledge (Suzuki et al., 2023). Through repeated exposure to and use of linguistic knowledge, learners may develop the ability to use the target language structure fluently (e.g., by studying with flashcards, learners acquire the ability to comprehend or produce *ringo* in context). A second pathway involves implicit knowledge. Implicit knowledge is unconscious and cannot be verbalized or intentionally retrieved (Ellis, 2015). It develops slowly over time as a result of repeated encounters and experience, typically without the learner being consciously aware of it (e.g., through repeated exposure to *ringo*, when learners hear this word, the concept of "apple" is automatically activated in their semantic network without awareness). Hereafter, we will use the term *procedural knowledge* when we refer broadly to the ability to use linguistic knowledge fluently, rapidly, and efficiently regardless of underlying cognitive mechanisms (i.e., automatized explicit knowledge or implicit knowledge). In contrast, we use the term *implicit knowledge* when we refer specifically to knowledge that is unconscious and that supports fluent performance.

Most existing studies on the spacing effect in vocabulary learning have measured only explicit knowledge (e.g., Karpicke & Bauernschmidt, 2011; Nakata, 2015; Nakata & Suzuki, 2019a) and it remains unclear whether the benefits of spacing extend to procedural knowledge development. Notable exceptions include studies conducted by Nakata and Elgort (2021) and Fang et al. (2024) that have yielded somewhat contradictory findings. On the one hand, Nakata and Elgort (2021) found that spaced learning surpasses massed learning for explicit vocabulary knowledge but shows no significant difference for implicit vocabulary knowledge acquisition (although Nakata and Elgort used the term *tacit knowledge*, we use the more common term *implicit knowledge* throughout this paper, as both refer to the same construct). Fang et al. (2024), on the other hand, found that massing equals spacing for explicit knowledge, yet spacing excels for automatized explicit and implicit knowledge.

The present study constitutes an approximate replication of Nakata and Elgort (2021) with two key modifications to the original treatment design (see below

for details). We chose to replicate their study because their findings contradict the widely held view that spaced practice facilitates learning compared to massed practice, which does not involve any intervening trials or time between repetitions of a given item (spacing effect; e.g., Cepeda et al., 2006; Wiseheart et al., 2019), at least for the acquisition of implicit vocabulary knowledge in L2. An approximate replication of their study allows for the evaluation of the robustness of their findings under modified conditions, extending understanding of how different instructional variables influence L2 vocabulary learning (Rogers, 2021).

2. Literature review

2.1. Spacing effect and lag effect in L2 learning

Since the spacing effect is recognized as a reliable and robust finding (e.g., Cepeda et al., 2006; Wiseheart et al., 2019), researchers have attempted to increase L2 learning by taking advantage of the benefits of spacing. Most studies have shown that spacing surpasses massing. Nakata (2015) found that spacing doubled vocabulary gains for first language (L1) Japanese L2 learners of English in posttests conducted one week after the learning phase. The spacing benefits reported in Nakata (2015) are consistent with other studies manipulating within-session spacing, such as those on L2 vocabulary (Karpicke & Bauernschmidt, 2011; Nakata & Elgort, 2021; Nakata & Suzuki, 2019a), as well as grammar (Nakata & Suzuki, 2019b; Pan et al., 2019; Suzuki et al., 2020).[1] Kim and Webb's (2022) meta-analysis of L2 spacing research shows that spacing yields a small to medium effect (g = 0.58) on immediate posttests and a medium to large effect (g = 0.80) on delayed posttests compared to massing (no spacing).

Research on the lag effect (i.e., short vs. long spacing) in L2 learning, however, has yielded inconsistent findings. While some studies have reported benefits of longer spacing over shorter spacing (e.g., vocabulary: Karpicke & Bauernschmidt, 2011; Nakata & Webb, 2016; Nakata et al., 2023; grammar: Bird, 2010; Rogers, 2015), other studies have not (e.g., vocabulary: Nakata, 2015; Rogers & Cheung,

[1] As one anonymous reviewer points out, some grammar studies on the spacing effect (e.g., Nakata & Suzuki, 2019b; Pan et al., 2019; Suzuki et al., 2020) introduce spacing through interleaving. Interleaving refers to a practice schedule where exemplars from multiple concepts or skills are mixed together, as opposed to blocking, where only one skill or concept is practiced at a time (Nakata & Suzuki, 2019b). By definition, interleaved practice entails spaced practice, whereas blocked practice typically involves massed practice, although it is possible to have blocked practice that incorporates spaced practice (Taylor & Rohrer, 2010). Following Kim and Webb (2022), we included studies comparing blocking and interleaving in our review of the literature on the spacing effect.

2021; grammar: Kasprowicz et al., 2019; Suzuki, 2017; Suzuki & DeKeyser, 2017; pronunciation: Carpenter & Mueller, 2013; Li & DeKeyser, 2019). Kim and Webb's (2022) meta-analysis reports no significant difference between shorter and longer spacing on immediate posttests (g = -0.15), whereas longer spacing slightly improves delayed posttest retention (g = 0.40).

The spacing and lag effects in L2 learning also depend on several variables, including learning target, posttest timing, and practice type. Kim and Webb (2022) note stronger spacing effects for L2 vocabulary (g = 0.76 to 1.15) than grammar (g = 0.11 to 0.14). Posttest timing may also influence the lag effect; the benefits of longer spacing tend to be more pronounced on long-delay posttests than on immediate or short-delay posttests (e.g., Bird, 2010; Cepeda et al., 2009; Kim & Webb, 2022; Nakata & Suzuki, 2019a; Nakata & Webb, 2016). A further factor that may moderate the lag effect is practice type: decontextualized or contextual vocabulary learning. In decontextualized learning, vocabulary is presented outside of meaningful context, and learners are explicitly instructed to memorize vocabulary. Examples include paired-associate learning. In this type of learning, learners are presented with pairs of L1 and L2 words at the outset (e.g., *ringo* [L2 Japanese] – *apple* [L1 English]). This is followed by retrieval practice, where learners are asked to translate from L2 to L1 (e.g., *ringo* = ____) or vice versa (e.g., *apple* = ____). In contextual vocabulary learning, in contrast, the primary focus is on meaning, and learners are not explicitly asked to learn L2 vocabulary. Examples include incidental vocabulary learning from context, where learners read a passage with the primary purpose of text comprehension and vocabulary may be picked up as a by-product. A review of the literature suggests that while studies examining decontextualized learning report benefits of longer over shorter spacing (e.g., Karpicke & Bauernschmidt, 2011; Nakata & Webb, 2016; Nakata et al., 2023), those examining contextual learning often do not (e.g., Elgort & Warren, 2014; Serrano & Huang, 2021; Serrano & Pellicer-Sánchez, 2024; Webb & Chang, 2015; for a discussion of why this might be the case; see the section below presenting the motivation for the study). Furthermore, the research setting may also moderate the lag effect, as the benefits reported for longer spacing in laboratory studies (e.g., Karpicke & Bauernschmidt, 2011; Nakata & Webb, 2016; Nakata et al., 2023) are sometimes not replicated in quasi-experimental classroom research (e.g., Rogers & Cheung, 2021; Serrano & Huang, 2021; Webb & Chang, 2015).

2.2. Effects of spacing on the acquisition of explicit and procedural knowledge

Another factor relevant to the present study is that spacing may have differential effects depending on whether explicit or procedural knowledge is measured.

Suzuki et al. (2023) observed that the lag effect in L2 learning is moderated by knowledge type: explicit or procedural. Studies have shown that longer spacing enhances the acquisition of explicit knowledge (e.g., Karpicke & Bauernschmidt, 2011; Nakata & Suzuki, 2019a; Nakata & Webb, 2016; Nakata et al., 2023; Rogers, 2015), but does not benefit procedural knowledge development (e.g., Li & DeKeyser, 2019; Suzuki, 2017; Suzuki & DeKeyser, 2017). The results may be explained by skill acquisition theory (DeKeyser & Suzuki, 2025). Specifically, unlike explicit knowledge, which can be acquired within a short time, sometimes only from a single learning event, procedural knowledge develops slowly through repeated practice, exposure, and experience. As such, shorter spacing, which provides more intensive practice within a short time frame, may be more effective for the acquisition of procedural knowledge (Suzuki et al., 2023).

Although research indicates that the lag effect in L2 learning is influenced by knowledge type, it remains unclear whether the spacing effect varies across different types of L2 knowledge. This lack of clarity might be due to many L2 spacing effect studies measuring only explicit knowledge (e.g., Karpicke & Bauernschmidt, 2011; Nakata, 2015; Nakata & Suzuki, 2019a), overlooking procedural knowledge. Studies conducted by Nakata and Elgort (2021) and Fang et al. (2024) are, however, notable exceptions.

Nakata and Elgort (2021) investigated the effects of massing and spacing on the acquisition of explicit and implicit knowledge. Sixty-six L1 Japanese learners of L2 English read 144 sentences, each containing one of 48 pseudowords. Half of the pseudowords were assigned to a massed condition, and the other half to a spaced condition. In the massed condition, three sentences for each pseudoword were presented simultaneously, and participants inferred the meaning of the pseudoword. In the spaced condition, participants viewed one sentence per pseudoword, with repetitions after 47 intervening sentences (Figure 1, left). In both massed and spaced conditions, after inferring pseudoword meanings, participants were shown the correct meaning as feedback. Explicit knowledge was measured by a meaning recall (L2 to L1 translation) test given immediately and two days after the treatment, as well as a meaning-form matching test given two days after the treatment. Implicit knowledge was measured using a semantic priming task conducted immediately and two days after the treatment. Nakata and Elgort (2021) found that, although spacing was more effective than massing for acquiring explicit vocabulary knowledge, no significant difference existed between the two for acquiring implicit knowledge.

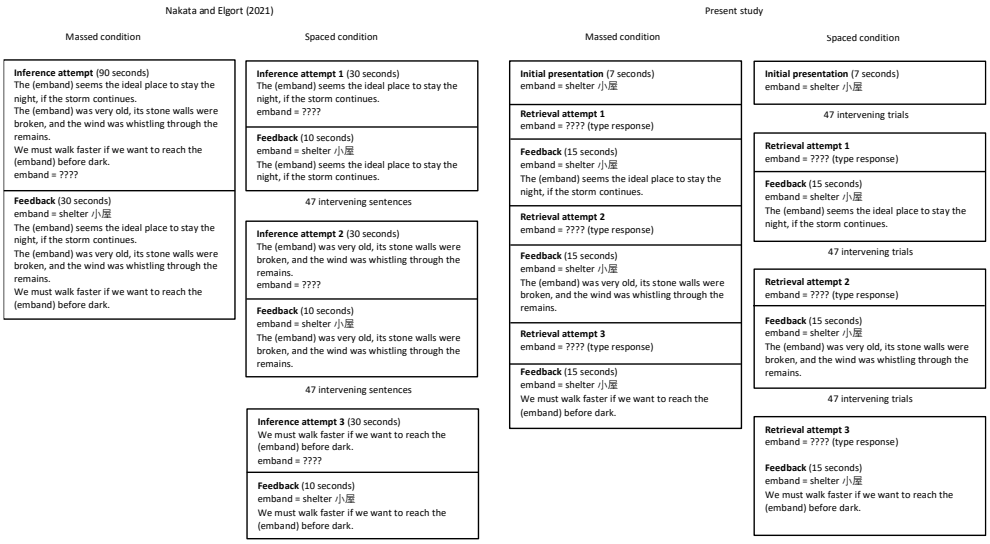

Figure 1 Differences in treatment between Nakata and Elgort (2021) and the present study.

Fang et al. (2024) examined the effects of massing and spacing on the acquisition of explicit, automatized explicit, and implicit knowledge of English collocations. In their study, 29 Chinese EFL learners studied 36 verb-noun collocations (e.g., *cast doubt*, *drop a hint*) in the form of flashcards and a matching exercise. Learning was measured by three types of dependent measures: a form recall posttest to measure explicit knowledge, an acceptability judgment task to measure automatized explicit knowledge, and a primed lexical decision task to measure implicit knowledge. Unlike Nakata and Elgort (2021), Fang et al. (2024) found that massing was as effective as spacing for explicit knowledge acquisition, but spacing was superior for both automatized explicit and implicit knowledge. The inconsistency may arise from different materials: Nakata and Elgort (2021) examined the learning of novel pseudowords, whereas Fang et al. (2024) investigated the learning of collocations comprising familiar words.

3. The present study

3.1. Motivation for the study

The present study is an approximate replication of Nakata and Elgort (2021). Although the method used in this study closely followed that of the original study, we introduced two key modifications to the treatment design. First, Nakata and Elgort (2021) examined contextual vocabulary learning, where participants were

asked to infer the meaning of pseudowords from context and were not explicitly instructed to memorize the target pseudowords (Figure 1, left). The treatment in this study, in contrast, involved decontextualized paired-associate learning, where participants were presented with the target pseudowords outside of meaningful context, and explicitly instructed to learn the pairs of target pseudowords and their meaning (i.e., L1 translations and L2 definitions; Figure 1, right). This change was made to examine whether the results reported by Nakata and Elgort (2021) also extend to decontextualized paired-associate learning.

As stated in the previous section, while studies examining decontextualized learning tend to find benefits of longer spacing, those examining contextual learning often fail to do so. One possible explanation for these findings is that in contextual vocabulary learning, novel lexical items need to be repeated relatively soon after the first exposure to prevent the decay of the initial memory trace (Laufer, 2003; Nation, 2022). Otherwise, the second encounter may essentially function as an initial encounter, instead of repetitions of (partly) familiar words. This suggests that massing or short spacing, where novel L2 words are more likely to be repeated before forgetting occurs, may facilitate contextual vocabulary learning more than longer spacing, where new words are repeated after memory decays. Another reason why the benefits of spacing are more pronounced in decontextualized learning than contextual learning is that contextual learning typically involves induction (i.e., guessing meanings of unknown words from context; Figure 1, left). Decontextualized vocabulary learning, such as paired-associate learning, in contrast, does not involve induction, because learners are typically provided with novel L2 words together with their meanings (e.g., L1 translation or L2 definition) at the beginning of the learning phase (Figure 1, right). Some psychology studies suggest that although spacing may facilitate recall of previously stored information, it might inhibit inductive learning because, as Kornell and Bjork (2008) write, it is possible that "spacing is the friend of recall, but the enemy of induction" (p. 585). Because spacing is not necessarily beneficial for inductive learning, existing studies on contextual vocabulary learning might have found smaller benefits of spacing compared with those on decontextualized learning.

By changing the treatment task from contextual learning to decontextualized paired-associate learning, this study allows us to examine whether the results reported by Nakata and Elgort (2021) regarding implicit knowledge were due to the nature of the treatment task used (i.e., contextual vocabulary learning with induction), or a more fundamental limitation in the ability of spaced practice to enhance the acquisition of implicit vocabulary knowledge. The use of paired-associate learning as the treatment task also has ecological validity. This is because paired-associate learning, such as learning from flashcards, is not only effective but also a common and popular strategy in countries such as the USA and Japan (Nakata, 2020; Zung et al., 2022).

Another change introduced into the treatment was the number of sentences presented to the learners. As described in the previous section, in the massed condition in Nakata and Elgort (2021), participants were presented with three sentences for a pseudoword simultaneously, whereas in the spaced condition, they were presented with only one sentence at a time (Figure 1, left). One drawback of this design feature is that the number of sentences presented simultaneously (i.e., one vs. three) was confounded with the practice schedule (massed vs. spaced). Consequently, it is unclear to what extent their results were because of the spacing schedule versus the number of sentences presented at once. Notably, Nakata and Elgort (2021) argue that simultaneous exposure to three example sentences for a given pseudoword in the massed condition might have helped the integration of the pseudowords into the existing networks of already familiar words in the mental lexicon, diminishing any potential benefits of spacing. Another possibility is that processing three sentences simultaneously in the massed condition in the Nakata and Elgort (2021) study imposed a higher cognitive load on the learner compared to processing a single sentence at a time in the spaced condition (Figure 1, left). This increased difficulty might have led to more effortful processing in the massed condition, resulting in more robust mental representations of target pseudowords (desirable difficulties framework; Suzuki et al., 2019), which might have reduced spacing benefits. In this study, in both massed and spaced conditions, participants were presented with only one sentence for a pseudoword at a time (Figure 1, right). This change was made to ensure that the potential effects of spacing or massing were not confounded by the number of sentences provided simultaneously, allowing a clearer investigation of how the practice distribution might influence the acquisition of explicit and implicit knowledge.

3.2. Research questions and hypotheses

Based on the rationale presented above, in this study, we examined the effects of spacing, relative to massing, on vocabulary learning in a decontextualized paired-associate learning paradigm. Specifically, two research questions (RQs) were addressed in the present study:

RQ1: To what extent does spacing facilitate the acquisition of explicit vocabulary knowledge compared to massing?
RQ2: To what extent does spacing facilitate the acquisition of implicit vocabulary knowledge compared to massing?

Based on previous studies which were reviewed above, the following two hypotheses (Hs) were formulated:

H1: Spacing facilitates the acquisition of explicit vocabulary knowledge more than massing for decontextualized paired-associate learning.
H2: Spacing facilitates the acquisition of implicit vocabulary knowledge more than massing for decontextualized paired-associate learning.

4. Method

The method used in this study closely followed that used in Nakata and Elgort (2021), except for the treatment (Figure 1). This was done to ensure that any difference in results between the present and original studies could be attributed to the difference in the treatments (i.e., paired-associate learning vs. contextual learning; number of sentences provided simultaneously for the massed and spaced conditions). Table 1 illustrates the comparison of methodological features of Nakata and Elgort (2021) and the present study.

Table 1 Comparison of methodological features of Nakata and Elgort (2021) and the present study

	Nakata and Elgort (2021)	Present study
Location	Laboratory	Identical
Participants	66 Japanese university students with an average English vocabulary size of 8,698.5 word families	69 Japanese university students with an average English vocabulary size of 8,355.8 word families
Materials	48 pseudowords	Identical
Procedure	Session 1 1) informed consent 2) treatment 3) background questionnaire 4) filler task 5) immediate posttest Session 2 1) delayed posttest 2) Vocabulary Size Test (VST) 3) operation span (O-Span) task 4) questionnaire	Identical
Treatment	Massed condition (Figure 1, left): – An inference attempt for the meaning of a pseudoword based on three sentences containing a pseudoword followed by feedback (presentation of the pseudoword, its meaning, and three context sentences) Spaced condition (Figure 1, left): – 1st to 3rd trials: An inference attempt for the meaning of a pseudoword based on a sentence containing a pseudoword followed by feedback (presentation of the pseudoword, its meaning, and one context sentence)	Massed and spaced conditions (Figure 1, right): – 1st trial: Presentation of a pseudoword and its meaning – 2nd to 4th trials: A retrieval attempt for the meaning of a pseudoword followed by feedback (presentation of the pseudoword, its meaning, and one context sentence)
Treatment duration	Approximately 96 minutes	Approximately 62.3 minutes on average (see Section 5. reporting results)
Spacing	Massed condition: 0 intervening sentences Spaced condition: 47 intervening sentences	Massed condition: 0 intervening trials Spaced condition: 47 intervening trials

Retention interval (interval between the treatment and posttest)	Immediate posttest: 0 days Delayed posttest: 2 days	Identical
Immediate posttest	1) semantic priming 2) meaning recall	Identical
Delayed posttest	1) semantic priming 2) meaning recall 3) meaning-form matching	Identical

4.1. Participants

Sixty-nine L1 Japanese, L2 English undergraduate learners, with an average age of 19.2 (*SD* = 1.3), participated in the study (37 female, 32 male). This sample size was determined based on that of the original Nakata and Elgort (2021) study (*N* = 66). The participants included 21 English, 17 education, 12 mathematics, 10 economics, 5 engineering, and 4 law majors. Their English vocabulary size, assessed via the bilingual Japanese Vocabulary Size Test (VST; Nation & Beglar, 2007), averaged 8,355.8 word families (*SD* = 908.8), comparable to the 8,698.5 word families (*SD* = 1,136.6) reported by Nakata and Elgort (2021). The participants' average TOEIC score was 746.8 (*SD* = 97.8), reflecting higher-intermediate to advanced English proficiency. Each participant received 6,000 yen for their participation.

4.2. Materials

The study materials, sourced from Nakata and Elgort (2021), comprised 48 pseudowords (e.g., *askent*, *creptor*). Twenty-four pseudowords were building/household-themed (e.g., *shelter*, *bathroom*), and the other half were cooking/food-themed (e.g., *peanut*, *tasting*). The 48 pseudowords were grouped into two 24-word sets (A and B), with 12 pseudowords per theme. In a counterbalanced within-participant design, half the participants studied Set A under the massed condition and Set B under the spaced condition, and vice versa for the other half.

4.3. Measures

The present study adopted Nakata and Elgort's (2021) three dependent measures: meaning recall, meaning-form matching, and semantic priming tasks. Explicit vocabulary knowledge was assessed through meaning recall and meaning-form matching tests, with response accuracy as the primary variable. Implicit vocabulary

knowledge was evaluated using the semantic priming task, where both response accuracy and response time served as dependent variables.

The meaning recall test, designed to measure explicit vocabulary knowledge, was conducted immediately after the treatment (Session 1) and two days later (Session 2; see Section 4.4. describing the procedure). During the test, 48 pseudowords were presented individually on a computer screen, and participants were asked to type their meanings in L1 Japanese or L2 English. The meaning-form matching test, administered only in Session 2, provided a pseudoword's meaning as an English synonym and Japanese translation (e.g., *menu-メニュー*) and asked participants to select the correct pseudoword from four options (e.g., *warshim*, *brophy*, *dapson*, *ganse*).

To measure participants' implicit vocabulary knowledge, the semantic priming task was administered in both Sessions 1 and 2 using E-Prime® software on an ASUS personal computer with a Chronos® response box. For this task, participants were presented with letter sequences (e.g., *nation*, *hankest*) and quickly indicated whether they were real English words by pressing *Yes* or *No* on the response box. We adopted counterbalanced item lists (A and B) from Nakata and Elgort (2021), each containing 192 prime-target pairs: 24 related, 24 unrelated, and 144 filler pairs. In related pairs, one of the 48 pseudowords studied during treatment served as the prime for a semantically related target word (e.g., *narage*, meaning *mural*, paired with *artist*). In unrelated pairs, a semantically unrelated word was the prime for the same target (e.g., *twenty – artist*). Prime relatedness was evenly distributed between lists; for instance, a related prime in list A (e.g., *narage – artist*) corresponded to an unrelated prime in list B (e.g., *twenty – artist*), and vice versa. The two lists were counterbalanced across participants to ensure balanced exposure to related and unrelated primes, minimizing bias in the semantic priming task. Half the participants were assigned list A in Session 1 and list B in Session 2, with the other half receiving list B then list A. The 144 filler pairs were excluded from analysis. Each list began with 50 practice trials to familiarize participants with the semantic priming task.

4.4. Procedure

We replicated Nakata and Elgort's (2021) experimental procedure, modifying only the treatment (Figure 1). The study spanned two sessions, with Session 2 conducted two days after Session 1 (see the details of the procedure in Table 1). Data collection took place individually in a secluded, soundproof office. At the beginning of Session 1, participants were informed about the study and signed consent forms if agreeing to participate. Next, the treatment was conducted with a computer program originally developed for Nakata and Elgort (2021) but modified for the present study. During the treatment, participants encountered

each target pseudoword four times. The first encounter was a presentation trial, displaying each pseudoword alongside its L1 (Japanese) translation and L2 (English) synonym for seven seconds. The subsequent three encounters were self-paced retrieval trials, where participants were presented with a pseudoword and asked to type its meaning in either L1 (Japanese) or L2 (English), without a time limit. To address variations in retrieval durations, time on task was treated as a covariate in analyses (see information on scoring and data analysis below). Each retrieval trial was followed by 15 seconds of feedback, showing the target pseudoword, its L1 (Japanese) translation, L2 (English) synonym, and an example sentence. We used the same sentences from Nakata and Elgort (2021) but presented them one at a time for both the massed and spaced conditions (Figure 1, right).

Following Nakata and Elgort (2021), the 48 pseudowords were split evenly into massed and spaced schedules. Massed pseudowords were repeated four times consecutively (one presentation trial + three retrieval trials, each followed by feedback). Spaced pseudowords, however, were repeated after every 47 intervening trials (approximately 10.5 minutes). After the treatment, participants completed the same English learning background questionnaire, followed by a filler task of 50 addition and subtraction problems (e.g., 92 + 55 = ?). Following the filler task, participants took two immediate posttests: a semantic priming task and a meaning recall test (see Section 4.3. for details).

Session 2 occurred two days after Session 1, using the same procedure and computer programs as Nakata and Elgort (2021). Specifically, Session 2 began with the semantic priming task, identical to Session 1 but using the alternate item list (list A in Session 1 switched to list B in Session 2, or vice versa). The semantic priming task was followed by the meaning recall test, identical to Session 1 except for a randomized item order. Lastly, a meaning-form matching test was administered to assess explicit vocabulary knowledge (see Section 4.3. for details). After the tests, participants completed levels 1 to 14 of the bilingual Japanese version of the VST (Nation & Beglar, 2007). Next, participants completed the Japanese version of the operation span (O-Span) task (Kobayashi & Okubo, 2014) to assess working memory capacity. At the end of Session 2, a questionnaire was administered to assess their perceptions of the experiment.[2]

[2] As described in the section dealing with scoring and data analysis, the score on a working memory test (O-span) was included as a secondary variable for all dependent measures. O-span was included in the final model for the meaning-form matching posttest (Table 5). However, the effect of O-span was not statistically significant (p = .118). Furthermore, the final models for all other dependent measures did not include O-span (see Tables 3-6 and 8-10 in the results section), as keeping it in the models did not lead to a significantly better fit. These findings suggest that working memory capacity did not have a large effect on the results of this study. Results from the perception questionnaires were not included in the analysis, as Nakata and Elgort (2021) also did not include them in their analysis.

4.5. Scoring and data analysis

Following Nakata and Elgort (2021), participants' responses during the treatment and meaning recall posttests were automatically categorized by software into four types: (1) correct, (2) blank, (3) cross-association errors (correct responses for other pseudowords), and (4) other responses. Category 4 was independently scored by one author and a research assistant, achieving 99.5% inter-rater agreement. Any disagreements were resolved through discussion. Correct responses were coded as 1, while all other responses were coded as 0 (incorrect).

Responses on the semantic priming and meaning-form matching tests were scored following Nakata and Elgort (2021). In the semantic priming task, E-Prime software recorded accuracy (correct: 1, incorrect: 0) and participants' response time from stimulus appearance to their *Yes/No* button press on the response box. For the meaning-form matching test, responses were automatically scored as correct (1) or incorrect (0) using the same custom computer program as in Nakata and Elgort (2021).

All statistical analyses were conducted using R software. Mixed-effects regression analyses were conducted to investigate massing and spacing effects on explicit and implicit word knowledge acquisition. Mixed logit models analyzed binary response accuracy (1: correct, 0: incorrect) for treatment, meaning recall, meaning-form matching, and semantic priming tasks. Linear mixed-effects models were used to analyze response time in the semantic priming task.

Results were analyzed using R software's lme4 package, following Nakata and Elgort (2021), with two key differences. First, unlike Nakata and Elgort's (2021) single inference trial in their massed condition versus three in their spaced condition, our study featured three retrieval trials for each pseudoword in both massed and spaced conditions (Figure 1). To examine whether retrieval accuracy increased as a function of the number of retrieval trials, Retrieval number (1/2/3) was added as a covariate when analyzing retrieval accuracy during the treatment. Second, unlike Nakata and Elgort's (2021) computer-paced treatment, the present study's treatment was self-paced. To mitigate potential differences in retrieval practice durations, we used time on task during retrieval trials as a covariate when analyzing performance during the learning phase, meaning recall posttests, and the meaning-form matching posttest.

Crossed random effects were applied for participants and pseudowords. A minimally adequate statistical model was fitted using stepwise variable selection and analyses of variance (ANOVAs) for model comparisons. For explicit knowledge analyses, the primary predictors were schedule (massed/spaced), accuracy, and their interaction (H1). In the semantic priming task, relatedness (related/unrelated) was the primary predictor, expecting faster responses to related pairs

(priming effect). Schedule (spaced/massed) was the secondary predictor. We hypothesized that spacing and massing would differentially affect semantic knowledge development (H2), predicting an interaction between relatedness and schedule. Specifically, the spaced condition was expected to produce higher response accuracy and faster response times compared to the massed condition, reflecting stronger facilitation of implicit knowledge acquisition through spaced practice.

For the analyses of both explicit and implicit knowledge, schedule, accuracy, and their two-way interactions were tested. Secondary variables included participants' English vocabulary size (VST), O-span, age, and theme (building/cooking). We repeated the posttest, factoring in Session (1 or 2) and its relationship with other variables. For the semantic priming task, we tested the following variables: item type (pseudoword/word), prime length (number of letters in a prime), prime accuracy (1/0), prime response time, and target length (number of letters in a target). Effect sizes *d* were calculated for the main effects and interactions. For interpreting the effect sizes, field-specific guidelines proposed by Plonsky and Oswald (2014) were used. R code used for data analysis is provided in the Appendix.

5. Results

5.1. Learning phase performance

The mean total retrieval latency was 9.3 [8.6, 10.0] minutes (*SD* = 4.1) for the massed condition and 11.4 [10.3, 13.0] minutes (*SD* = 1.4) for the spaced condition (95% confidence intervals are given inside brackets). Overall, the treatment took approximately 62.3 minutes on average (initial presentation: 5.6 minutes; retrieval practice: 20.7 minutes; feedback: 36.0 minutes; Figure 1, right). Response accuracy during the treatment is summarized in Table 2 and Figure 2. The mixed logit model (Table 3) revealed a significant main effect of schedule ($z = 22.46$, $p < .001$), with a large effect size ($d = 4.26$). This suggests that the massed condition produced more correct responses during the treatment than the spaced condition. The main effect of retrieval number was also significant ($z = 22.09$, $p < .001$) with a small effect size ($d = 0.85$), suggesting that response accuracy increased as learning progressed. There was also a significant, but small, main effect of theme ($z = 2.69$, $p = .007$, $d = 0.33$); the accuracy of the pseudowords in the cooking theme was significantly higher than the pseudowords in the building theme. There was a significant interaction between schedule (massed) and retrieval number ($z = -7.39$, $p < .001$) with a small effect size ($d = -0.65$), suggesting that the benefit of repeated retrieval in the massed condition was smaller compared to the spaced condition.

Table 2 Response accuracy by retrieval attempt during the treatment

Schedule	Retrieval number	*M*	95% CI	*SD*
Massed	1	96%	[95%, 97%]	20%
	2	99%	[98%, 99%]	12%
	3	98%	[97%, 99%]	15%
Spaced	1	13%	[11%, 15%]	34%
	2	36%	[33%, 39%]	48%
	3	59%	[56%, 62%]	49%

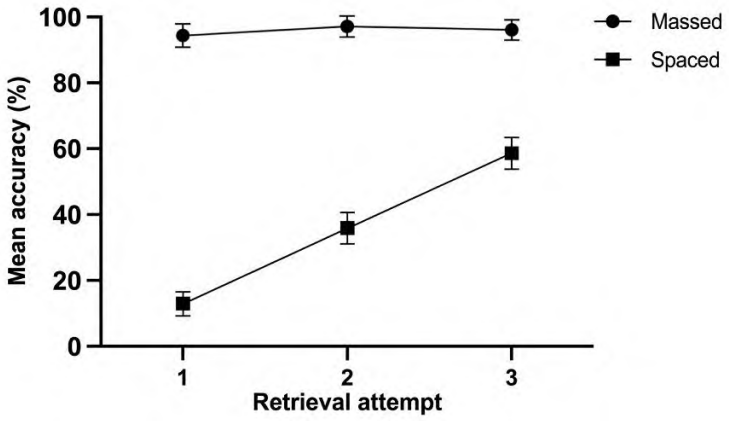

Figure 2 Effect plot for response accuracy by retrieval attempt (fit and 95% CIs).

Table 3 Retrieval accuracy during the treatment (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d* [95% CI]
Intercept[a]	-4.28	0.30	-14.19	< .001	-2.36 [-2.69, -2.03]
Schedule = massed	7.73	0.34	22.46	< .001	4.26 [3.89, 4.64]
Retrieval number	1.54	0.07	22.09	< .001	0.85 [0.77, 0.92]
Theme = cooking	0.59	0.22	2.69	.007	0.33 [0.09, 0.56]
Schedule = massed: Retrieval number	-1.18	0.16	-7.39	< .001	-0.65 [-0.82, -0.48]

Note. [a]Intercept levels: schedule = spaced, retrieval number = 1, theme = building

5.2. Meaning recall posttest

On the immediate posttest, the mean response accuracy was 18% [12%, 24%] (*SD* = 39%) and 65% [58%, 72%] (*SD* = 48%) for the massed and spaced conditions, respectively. On the delayed posttest, mean response accuracy was 14% [9%, 19%] (*SD* = 35%) for the massed condition and 52% [44%, 60%] (*SD* = 50%) for the spaced condition (Figure 3). The mixed logit model (Table 4) revealed a significant main effect of schedule ($z = 26.05$, $p < .001$), with a large effect size ($d = 1.63$). The results suggest that the spaced condition led to better meaning recall

than the massed condition on both immediate and delayed posttests. A significant main effect of session (z = -7.82, p < .001) with a very small effect size (d = -0.38) indicates that scores on the immediate posttest were significantly higher than those on the delayed posttest. Theme was also a significant factor, with the words in the cooking theme significantly easier than in the building theme (z = 2.02, p = .044), despite a very small effect size (d = 0.31). The main effect of time on task during the treatment was also significant (z = 4.22, p < .001), although the effect size was very small (d = 0.12). There were no significant interactions in the model.

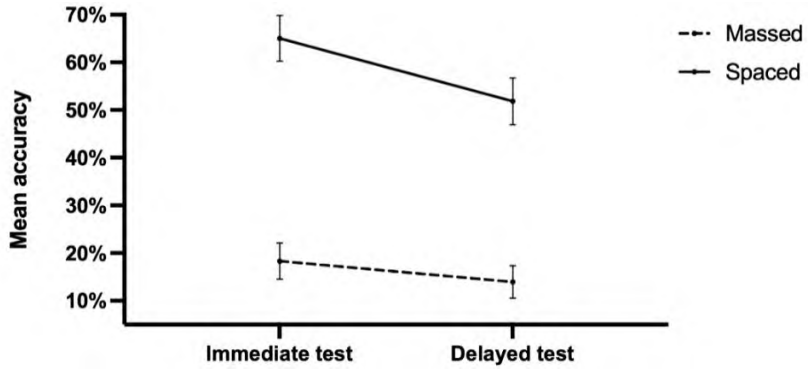

Figure 3 Mean accuracy of massed vs. spaced schedules on meaning recall posttests (fit and 95% CIs).

Table 4 Accuracy on recall posttests (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d* [95% CI]
Intercept[a]	-2.81	0.33	-8.53	< .001	-1.55 [-1.91, -1.19]
Schedule = spaced	2.97	0.11	26.05	< .001	1.63 [1.51, 1.76]
Session = delayed	-0.70	0.09	-7.82	< .001	-0.38 [-0.48, -0.29]
Theme = cooking	0.57	0.28	2.02	.044	0.31 [0.01, 0.62]
Time on task	0.01	0.01	4.22	< .001	0.12 [0.06, 0.18]

Note. [a]Intercept levels: schedule = massed, session = immediate, theme = building

5.3. Meaning-form matching posttest

The mean accuracy on the meaning-form matching posttest was 61% [55%, 67%] (*SD* = 49%) for the massed and 86% [82%, 91%] (*SD* = 35%) for the spaced condition (Figure 4). The mixed logit model (Table 5) showed a significant main effect of schedule (z = 12.98, p < .001) with a small effect size (d = 0.91), suggesting that the spaced condition yielded a significantly higher score than the massed schedule. Time on task was also significant (z = 2.56, p = .010), although the effect size was very small (d = 0.10). There were no significant interactions in the model.

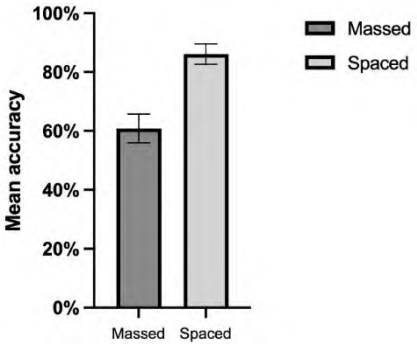

Figure 4 Mean accuracy of massed vs. spaced schedules on meaning-form matching posttest (fit and 95% CIs).

Table 5 Accuracy on meaning-form matching posttest (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d* [95% CI]
Intercept[a]	0.11	0.26	0.42	.673	0.06 [-0.22, 0.35]
Schedule = spaced	1.65	0.13	12.98	< .001	0.91 [0.77, 1.05]
Vocabulary size	2.40	1.61	1.49	.136	1.32 [-0.41, 3.06]
Theme = cooking	0.46	0.24	1.91	.056	0.25 [-0.01, 0.51]
Ospan	0.03	0.02	1.56	.118	0.01 [0.00, 0.03]
Time on task	0.01	0.01	2.56	.010	0.10 [0.02, 0.17]

Note. [a]Intercept levels: schedule = massed, theme = building

5.4. Semantic priming task

In the analyses of the semantic priming task, accuracy and response time were examined for semantic priming and lexical decision, respectively.

5.4.1. Semantic priming: Accuracy analysis

The mixed logit model (Table 6) showed a significant main effect of schedule, with the spaced condition being significantly more accurate than the massed condition (z = 3.46, p < .001), despite a very small effect size (d = 0.26). This indicates that response accuracy for targets was significantly higher when they were preceded by a prime studied in the spaced condition than in the massed condition (Figure 5). Target length was also a significant factor (z = -2.59, p = .010) with a very small effect size (d = -0.16), suggesting that targets with more letters had reduced accuracy. Session was also a significant main effect, with significantly reduced accuracy in the delayed test compared to the immediate test (z = -3.36, p = .001), although the effect size was very small (d = -0.18). There were no significant interactions in the model.

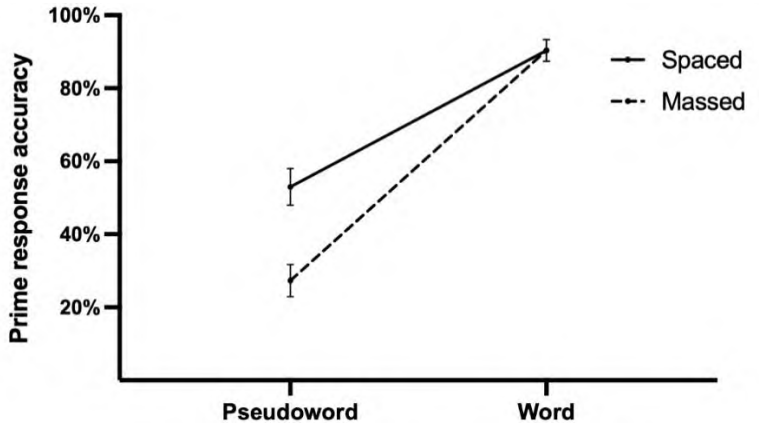

Figure 5 Effect plot for the interaction between schedule and item type in semantic priming (fit and 95% CIs).

Table 6 Semantic priming, accuracy analysis (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d* [95% CI]
Intercept[a]	2.75	0.32	8.55	< .001	1.51 [1.17, 1.86]
Relatedness = unrelated	0.19	0.11	1.65	.100	0.10 [-0.02, 0.23]
Schedule = spaced	0.47	0.14	3.46	< .001	0.26 [0.11, 0.41]
Vocabulary size	0.09	0.14	0.61	.540	0.05 [-0.11, 0.20]
Target length	-0.30	0.12	-2.59	.010	-0.16 [-0.29, -0.04]
Session = delayed	-0.32	0.09	-3.36	.001	-0.18 [-0.28, -0.07]

Note. [a]Intercept levels: relatedness = related, schedule = massed, session = immediate

5.4.2. Semantic priming: Response time analysis

Following Nakata and Elgort (2021), incorrect responses to targets (and their corresponding primes) were removed before the data analysis (18% of the data points; cf. 21% in Nakata & Elgort, 2021). Descriptive statistics for response times are provided in Table 7. The final model showed no significant main effects or interactions. There was also no effect of schedule, and the simplest model (see Table 8) did not include schedule. The results suggest that no significant difference existed between the massed and spaced schedules regarding the development of implicit vocabulary knowledge. Furthermore, neither the main effect of relatedness nor the interaction between relatedness and schedule was statistically significant, indicating a lack of significant semantic priming effect. The results suggest that neither massing nor spacing facilitated the acquisition of implicit knowledge.

Table 7 Descriptive statistics for response times (ms) by condition and time

	Immediate posttest			Delayed posttest		
	M	*SD*	95% CI	*M*	*SD*	95% CI
Massed	1,232	556	[1,179, 1,285]	1,030	513	[979, 1,080]
Spaced	1,151	581	[1,097, 1,205]	1,025	424	[985, 1,066]
Unrelated	868	454	[837, 898]	809	353	[785, 833]

Table 8 Semantic priming, response time analysis (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d* [95% CI]
Intercept[a]	-0.01	0.07	-0.23	1.000	-0.01 [-0.08, 0.06]
Relatedness = unrelated	-0.06	0.05	-1.37	.608	-0.03 [-0.08, 0.01]
Prime response time	-0.02	0.02	-1.12	.781	-0.01 [-0.03, 0.01]
Prime accuracy	0.09	0.05	2.04	.198	0.05 [0.00, 0.10]
Session = delayed	-0.01	0.04	-0.15	1.000	0.00 [-0.04, 0.04]
Target length	-0.01	0.01	-0.68	.969	0.00 [-0.01, 0.01]

Note. [a]Intercept levels: relatedness = related, session = immediate; prime response time is inverted and centered

5.4.3. Lexical decisions to primes: Response accuracy analysis

On the immediate posttest, the mean response accuracy for lexical decisions to the primes was 35% [28%, 41%] (*SD* = 48%) for the massed condition and 62% [54%, 71%] (*SD* = 49%) for the spaced condition. On the delayed posttest, it was 20% [15%, 25%] (*SD* = 40%) for the massed condition and 44% [36%, 52%] (*SD* = 50%) for the spaced condition. The fixed effects are presented in Table 9, with all main effects reaching statistical significance. Critically, accuracy in the spaced condition was significantly higher than in the massed condition ($z = 4.68$, $p < .001$) with a small effect size ($d = 0.89$). Accuracy on the immediate posttest (session = immediate) was significantly higher than on the delayed posttest ($z = 3.50$, $p < .001$), although the effect size was very small ($d = 0.14$).

Table 9 Lexical decisions to primes, accuracy analysis (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d* [95% CI]
Intercept[a]	2.03	0.20	9.99	< .001	1.12 [0.90, 1.34]
Item type = pseudoword	-1.95	0.42	-4.64	< .001	-1.07 [-1.53, -0.62]
Item type = word	1.65	0.32	5.23	< .001	0.91 [0.57, 1.25]
Schedule = spaced	1.61	0.34	4.68	< .001	0.89 [0.52, 1.26]
Session = immediate	0.25	0.07	3.50	< .001	0.14 [0.06, 0.21]
Prime response time	-0.01	0.03	-0.24	.809	0.00 [-0.03, 0.03]
Item type = pseudoword: Session = delayed	-1.22	0.18	-6.85	< .001	-0.67 [-0.87, -0.48]
Item type = word: Session = delayed	-0.52	0.13	-4.13	< .001	-0.29 [-0.43, -0.15]
Schedule = spaced: Session = delayed	-0.07	0.22	-0.31	.759	-0.04 [-0.28, 0.20]

Note. [a]Intercept levels: schedule = massed, session = immediate, item type = pseudoword

5.4.4. Lexical decisions to primes: Response time analysis

As in Nakata and Elgort (2021), instead of removing incorrect responses to primes, we included prime accuracy in the model for this analysis. In the immediate posttest, the mean response time for lexical decisions to the primes was 1,220 ms [1,112, 1,328] (*SD* = 543) and 1,141 ms [1,048, 1,234] (*SD* = 555) for the massed and spaced conditions, respectively. In the delayed posttest, the mean response time was 1,005 ms [907, 1,105] (*SD* = 491) for the massed condition and 1,017 ms [940, 1,094] (*SD* = 420) for the spaced condition. As shown in Table 10, none of the fixed effects reached statistical significance.

Table 10 Lexical decisions to primes, response time analysis (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d* [95% CI]
Intercept[a]	0.04	0.07	0.55	.987	0.02 [-0.06, 0.10]
Item type = pseudoword	-0.03	0.03	-1.08	.805	-0.02 [-0.05, 0.01]
Item type = word	0.00	0.02	0.12	1.000	0.00 [-0.02, 0.02]
Session = delayed	0.01	0.02	0.44	.995	0.00 [-0.02, 0.02]
Prime accuracy = 1	-0.01	0.02	-0.60	.981	-0.01 [-0.03, 0.02]
Prime length	0.00	0.01	0.14	1.000	0.00 [-0.01, 0.01]

Note. [a]Intercept levels: item type = pseudoword, session = immediate

5.5. Comparing Nakata and Elgort (2021) with this study

Table 11 and Figure 6 present a comparison of descriptive statistics between Nakata and Elgort (2021) and the present study. Descriptive statistics are provided only for the meaning recall and meaning-form matching posttests because those for the semantic priming task are not reported in the original Nakata and Elgort study. Table 11 shows that the posttest scores are 1.1 to 2.0 times higher in the present study than in the original study. The comparison of the effect sizes between the two studies is summarized in Table 12. The table indicates that the main effect of schedule led to larger effect sizes in this study (meaning recall: *d* = 1.63; meaning-form matching: *d* = 0.91) than those reported by Nakata and Elgort (2021; meaning recall: *d* = 0.86; meaning-form matching: *d* = 0.29). Overall, the results suggest that, for the acquisition of explicit knowledge, (a) the treatments in this study led to larger gains than those in the original study, and (b) spacing facilitated learning more in this study than in the original study.

Table 11 Comparing descriptive statistics: Nakata and Elgort (2021) vs. the present study

Measure	Timing	Nakata and Elgort (2021)				Present study			
		Massed		Spaced		Massed		Spaced	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Meaning recall	Immediate	12%	12%	42%	24%	18%	39%	65%	48%
Meaning-form matching	Delayed	7%	8%	32%	21%	14%	35%	52%	50%
	Delayed	55%	17%	75%	17%	61%	49%	86%	35%

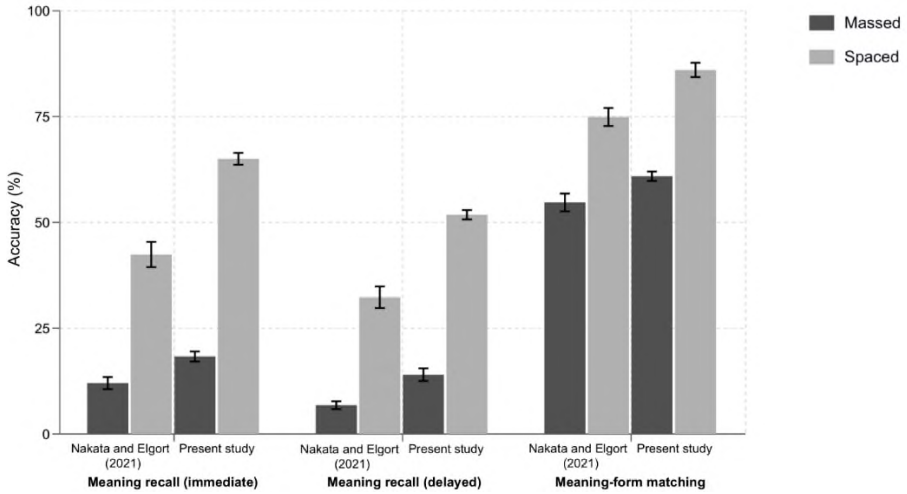

Figure 6 Comparing average accuracy on meaning recall and meaning-form matching posttests: Nakata and Elgort (2021) vs. the present study.

Table 12 Comparing the main effect of schedule: Nakata and Elgort (2021) vs. the present study

	Nakata and Elgort (2021)			Present study		
	z	*p*	*d*	*z*	*p*	*d*
Meaning recall	5.81	< .001	0.86	26.05	< .001	1.63
Meaning-form matching	2.79	.005	0.29	12.98	< .001	0.91
Semantic priming (accuracy)	2.97	.003	0.12	3.46	< .001	0.26
Lexical decisions to primes (accuracy)	9.19	< .001	0.47	4.68	< .001	0.89

Note. For semantic priming and lexical decisions to primes, only accuracy analysis is included. This is because the main effect of schedule did not reach statistical significance in the analysis of response time for these measures in either the previous or present study, and thus it was not included in the final model

6. Discussion

The purpose of this study was to replicate a study conducted by Nakata and Elgort (2021) to examine the effects of massing and spacing on the acquisition of

explicit and implicit vocabulary knowledge. H1 predicted that spacing would surpass massing for explicit knowledge acquisition. The results showed the advantage of spacing over massing across all explicit knowledge measures: immediate meaning recall, delayed meaning recall, and meaning-form matching. Our results support H1, confirming the benefits of spacing for explicit vocabulary knowledge acquisition, aligning with existing literature (e.g., Karpicke & Bauernschmidt, 2011; Nakata, 2015; Nakata & Elgort, 2021; Nakata & Suzuki, 2019a). The effect sizes in the present study (meaning recall: $d = 1.63$; meaning-form matching: $d = 0.91$) exceeded those reported by Kim and Webb (2022) in their meta-analysis ($0.58 \leq g \leq 0.80$), indicating that spaced practice in this study effectively enhanced explicit vocabulary knowledge acquisition. The advantage of spacing over massing in this study is also consistent with research in cognitive psychology (e.g., Cepeda et al., 2006; Wiseheart et al., 2019).

According to the descriptive statistics detailed in Table 11 and Figure 6, post-test scores for meaning recall and meaning-form matching in this study were 1.1 to 2.0 times higher than those found in the original Nakata and Elgort (2021) study. The results support the finding that paired-associate learning, such as the one used in this study, leads to larger vocabulary gains than contextual learning (Nakata, 2020). The present study also led to stronger spacing effects (meaning recall: $d = 1.63$; meaning-form matching: $d = 0.91$) than the Nakata and Elgort study (meaning recall: $d = 0.86$; meaning-form matching: $d = 0.29$; Table 12). The results are consistent with the finding that for the acquisition of explicit knowledge, the benefits of spacing are more pronounced in decontextualized learning than in contextual learning (see Section 3.1. on motivation for the study).

H2 predicted that the spacing effect would also be observed for the acquisition of implicit vocabulary knowledge. Results on the semantic priming task, however, showed no significant difference between massing and spacing for the acquisition of implicit knowledge. The results are consistent with Nakata and Elgort (2021), failing to support H2. The findings suggest that Nakata and Elgort's (2021) implicit knowledge results may not stem from the design features of their study (i.e., contextual learning with induction vs. paired-associate learning without induction; number of sentences provided simultaneously for the massed and spaced conditions). Instead, they may indicate that spaced practice may have fundamental limitations in enhancing implicit vocabulary knowledge acquisition. No spacing effect emerged for implicit knowledge, possibly because it develops slowly through repeated exposure and usage (Suzuki et al., 2023), likely requiring more extensive and varied input than this study provided.

Although consistent with Nakata and Elgort's (2021) results, our findings were at odds with those reported by Fang et al. (2024), who found that whereas massing is as effective as spacing for acquiring explicit knowledge, spacing may be more effective for acquiring automatized explicit and implicit knowledge. The inconsistency between Fang

et al. (2024), on the one hand, and Nakata and Elgort (2021) and the present study, on the other, may partly stem from the target items, as we speculated earlier in this paper: While Nakata and Elgort (2021) examined novel single words (pseudowords), Fang et al. (2024) investigated the learning of collocations comprising familiar words.

Importantly, although Nakata and Elgort (2021) observed a significant semantic priming effect for the pseudowords regardless of the schedules, no such effect was found for pseudowords studied in the massed or spaced condition in this study. The findings suggest that neither massing nor spacing facilitated implicit knowledge acquisition. The inconsistency may partly stem from three factors: treatment type, treatment duration, and L2 proficiency. First, the treatment in the present study involved decontextualized paired-associate learning, whereas Nakata and Elgort (2021) examined contextual vocabulary learning. The inconsistency suggests that contextual learning may facilitate implicit knowledge acquisition more than decontextualized learning (Chun et al., 2012). Second, the lack of the semantic priming effect in this study may also partly be attributed to the relatively short duration of the treatment. Elgort (2011) observed a significant semantic priming effect for pseudowords studied in a decontextualized format. However, the treatment in her study was implemented over one week, whereas it was concentrated into one session in this study. From the perspective of skill acquisition theory (DeKeyser & Suzuki, 2025), L2 learners progress from an initial reliance on explicit knowledge to automatized knowledge through repeated practice. Because the development of such fluent knowledge, whether it takes the form of automatized explicit knowledge or implicit knowledge, is a gradual process requiring extensive exposure and usage (Suzuki et al., 2023), neither massing nor spacing may have been sufficient to facilitate the acquisition of implicit knowledge in the present study. Notably, Nakata and Elgort (2021) observed a significant semantic priming effect in both massed and spaced conditions, although their treatment involved only one session. However, it is worth noting that their treatment lasted longer (approximately 96 minutes) compared to the treatment in the present study (approximately 62.3 minutes; Table 1). The findings suggest that, depending on the nature of the treatment, target stimuli, and participants, it is possible for implicit knowledge to develop after only a single treatment session. Third, another possible reason for the lack of a significant semantic priming effect relates to the proficiency level of the participants. Although participants in the present study had a relatively large English vocabulary size (M = 8,355.8 word families, SD = 908.8, range = 5,400–11,800), this was still smaller than that of the participants in Elgort and Piasecki (2014; M = 9,444 word families, SD = 1,689, range = 5,100–13,800). Given that Elgort and Piasecki (2014) found that decontextualized paired-associate learning may lead to a significant semantic priming effect only for those with larger vocabulary sizes, if the vocabulary size of our participants had been larger, we might have observed a significant priming effect. Therefore, it may be useful to recruit higher-proficiency learners in future research.

7. Pedagogical implications

Pedagogically, the findings of this study suggest that when learning novel words, it may be more beneficial to distribute practice opportunities over a certain period instead of repeating an item multiple times with no intervals, at least for the acquisition of explicit knowledge. This can be achieved by using flashcard software with spaced retrieval algorithms (Nakata, 2020), giving quizzes on previously introduced words, or using the same reading or listening materials after certain intervals. Although the spaced schedule outperformed the massed schedule on the meaning recall test, delayed posttest scores in the spaced condition were significantly lower than those on the immediate posttest. These findings suggest that, in order to promote long-term retention, distributing practice within a given treatment session, as in the present study, may not be sufficient; practice may need to be distributed over multiple days. We also observed larger gains in explicit vocabulary knowledge and a more pronounced spacing effect than those reported by Nakata and Elgort (2021). These enhanced outcomes are likely attributed to a key methodological modification: a shift from contextual to decontextualized paired-associate learning. Overall, the findings show the value of incorporating deliberate, spaced practice for the learning of explicit vocabulary knowledge.

The present study showed no spacing advantage for the acquisition of implicit knowledge. Therefore, if the goal is to develop learners' implicit vocabulary knowledge, relying solely on decontextualized paired-associate learning may not be sufficient. Because the brief treatment duration in our study did not help to form robust implicit knowledge, instructors are advised to extend learning sessions and maximize learners' repeated exposure to, and active use of, target words across various contexts (Suzuki et al., 2023). However, considering that explicit knowledge serves as a foundation for developing procedural knowledge (DeKeyser & Suzuki, 2025; Li & DeKeyser, 2019), explicit spaced learning may still have some value for classroom instruction. Instructors can utilize spaced practice to help learners establish initial form-meaning mappings of new words. By ensuring that learners have an initial, explicit understanding of new vocabulary items, instructors can create opportunities for learners to encounter and use these words in meaningful and communicative contexts.

8. Conclusion

In the present study, we replicated Nakata and Elgort's (2021) study to investigate the impact of massing and spacing on the acquisition of explicit and implicit knowledge in paired-associate learning of L2 vocabulary. The acquisition of explicit,

but not implicit, knowledge showed a spacing advantage in the results. The findings of the present study are useful because, although the benefits of spacing over massing are well-documented (e.g., Cepeda et al., 2006; Wiseheart et al., 2019), it remains unclear if spacing advantages apply to the acquisition of implicit knowledge.

Despite showing the robustness of the spacing effect on explicit vocabulary knowledge, the present study has several limitations. First, unlike Nakata and Elgort (2021), this study showed no significant semantic priming effect, possibly because of the relatively short treatment duration or the L2 proficiency levels of the participants. Future research should consider extending the treatment duration or recruiting higher-proficiency learners. Another limitation lies in the use of pseudowords. Although using pseudowords helps control participants' prior knowledge and ensures that any vocabulary gains are solely due to the treatment, it may limit generalizability. Future research might explore whether these pseudoword findings extend to real words. Since the development of implicit knowledge is critical for fluent and effortless language use, further research examining the acquisition of not only explicit but also implicit knowledge will be a useful follow-up to this study.

Acknowledgements

This research was supported in part by a Japanese Society for the Promotion of Science Grant-in-Aid for Research (#24K04110). We are deeply grateful to Irina Elgort for providing us with materials and instruments used in her study.

The authors used ChatGPT (GPT-4) for language improvement. No generative AI was used for other purposes, such as idea generation, content generation, or analysis.

References

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *32* (2), 635-650. https://doi.org/10.1017/S0142716410000470

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671-682. https://doi.org/10.3758/s13421-012-0291-4

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*(4), 236-246. https://doi.org/10.1027/1618-3169.56.4.236

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*(3), 354-380. https://doi.org/10.1037/0033-2909.132.3.354

Chun, E., Choi, S., & Kim, J. (2012). The effect of extensive reading and paired-associate learning on long-term vocabulary retention: An event-related potential study. *Neuroscience Letters*, *521*(2), 125-129. https://doi.org/10.1016/j.neulet.2012.05.069

DeKeyser, R., & Suzuki, Y. (2025). Skill acquisition theory. In B. VanPatten, G. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (4th ed., pp. 157-182). Routledge.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367-413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Elgort, I., & Piasecki, A. E. (2014). The effect of a bilingual learning mode on the establishment of lexical semantic representations in the L2. *Bilingualism: Language and Cognition*, *17*(3), 572-588. https://doi.org/10.1017/S1366728913000588

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, *64*(2), 365-414. https://doi.org/10.1111/lang.12052

Ellis, N. (2015). Implicit and explicit language learning: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 3-23). John Benjamins.

Fang, N., Elgort, I., & Chen, Z. (2024). Effects of retrieval schedules on the acquisition of explicit, automatized-explicit, and implicit knowledge of L2 collocations. *Studies in Second Language Acquisition*, *46*(3), 663-685. https://doi.org/10.1017/s0272263124000184

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*(5), 1250-1257. https://doi.org/10.1037/a0023436

Kasprowicz, R. E., Marsden, E., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. *Modern Language Journal*, *103*(3), 580-606. https://doi.org/10.1111/modl.12586

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, *72*(1), 269-319. https://doi.org/10.1111/lang.12479

Kobayashi, A., & Okubo, M. (2014). Assessment of working memory capacity with a Japanese version of the Operation Span Test. *The Japanese Journal of Psychology*, *85*(1), 60-68. https://doi.org/10.4992/jjpsy.85.60

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction?" *Psychological Science, 19*(6), 585-592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, *59*(4), 567-587. https://doi.org/10.3138/cmlr.59.4.567

Li, M., & DeKeyser, R. (2019). Distribution of practice effects in the acquisition and retention of L2 mandarin tonal word production. *Modern Language Journal*, *103*(3), 607-628. https://doi.org/10.1111/modl.12580

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning. *Studies in Second Language Acquisition*, *37*(4), 677-711. https://doi.org/10.1017/s0272263114000825

Nakata, T. (2020). Learning words with flash cards and word cards. In S. Webb (Ed.), *Routledge handbook of vocabulary studies* (pp. 304-319). Routledge.

Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, *37*(2), 233-260. https://doi.org/10.1177/0267658320927764

Nakata, T., & Suzuki, Y. (2019a). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, *41*(2), 287-311. https://doi.org/10.1017/s0272263118000219

Nakata, T., & Suzuki, Y. (2019b). Mixing grammar exercises facilitates long-term retention: Effects of blocking, interleaving, and increasing practice. *Modern Language Journal*, *103*(3), 629-647. https://doi.org/10.1111/modl.12581

Nakata, T., Suzuki, Y., & He, X. (2023). Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing? *Language Learning*, *73*(3), 799-834. https://doi.org/10.1111/lang.12553

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition*, *38*(3), 523-552. https://doi.org/10.1017/S0272263115000236

Nation, P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13. https://doi.org/10.26686/wgtn.12552197.v1

Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., & Rickard, T. C. (2019). Does interleaved practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills. *Journal of Educational Psychology*, *111*(7), 1172-1188. https://doi.org/10.1037/edu0000336

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912. https://doi.org/10.1111/lang.12079

Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*(4), 857-866. https://doi.org/10.1002/tesq.252

Rogers, J. (2021). Input spacing in second language classroom settings: Replications of Bird (2010) and Serrano (2011). *Language Teaching*, *54*(3), 424-433. https://doi.org/10.1017/S0261444820000439

Rogers, J., & Cheung, A. (2021). Does it matter when you review? Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition*, *43*(5), 1138-1156. https://doi.org/10.1017/S0272263120000236

Serrano, R. (2022). A state-of-the-art review of distribution-of-practice effects on L2 learning. *Studies in Second Language Learning and Teaching*, *12*(3), 355-379. https://doi.org/10.14746/ssllt.2022.12.3.2

Serrano, R., & Huang, H.-Y. (2021). Time distribution and intentional vocabulary learning through repeated reading: A partial replication and extension. *Language Awareness*, *32*(1), 1-18. https://doi.org/10.1080/09658416.2021.1894162

Serrano, R., & Pellicer-Sánchez, A. (2024). Online processing and vocabulary learning in massed versus spaced repeated reading. *Vigo International Journal of Applied Linguistics*, *21*, 129-164. https://doi.org/10.35869/vial.v0i21.4402

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning*, *67*(3), 512-545. https://doi.org/10.1111/lang.12236

Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, *21*(2), 166-188. https://doi.org/10.1177/1362168815617334

Suzuki, Y., Nakata, T., & DeKeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *Modern Language Journal*, *103*(3), 713-720. https://doi.org/10.1111/modl.12585

Suzuki, Y., Nakata, T., & Rogers, J. (2023). Optimizing input and intake processing: A role for practice and explicit learning. In Y. Suzuki (Ed.), *Practice and automatization in second language research: Theory, methods, and pedagogical implications* (pp. 39-62). Routledge.

Suzuki, Y., Yokosawa, S., & Aline, D. (2020). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research*, *26*(4), 671-695. https://doi.org/10.1177/1362168820913985

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837-848. https://doi.org/10.1002/acp.1598

Webb, S., & Chang, A. C.-S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, *19*(6), 667-686. https://doi.org/10.1177/1362168814559800

Wiseheart, M., Kim, A. S. N., Kapler, I. V., Foot-Seymour, V., & Kupper-Tetzel, C. E. (2019). Enhancing the quality of student learning using distributed practice. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 550-584). Cambridge University Press.

Zung, I., Imundo, M. N., & Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? *Memory, 30*(8), 923-941. https://doi.org/10.1080/09658211.2022.2058553

APPENDIX

R code used for data analysis

Table 3:
Model: glmer(Correctness ~ Schedule * Retrieval_number + Theme + (1 | id) + (1 | item_id), dataset, family= "binomial")

Table 4:
Model: glmer(Correctness ~ Schedule + Session + Theme + Time_On_Task + (1 | id) + (1 | item_id), dataset, family= "binomial")

Table 5:
Model: glmer(Correctness ~ Schedule + Vocabulary_size + Theme + Ospan + Time_On_Task + (1 | id) + (1 | Correct.Response), dataset, family= "binomial")

Table 6:
Model: glmer(Target_Accuracy ~ Relatedness + Schedule + Vocabulary_size + Target_length + Session + (1 | id) + (1 + Vocabulary_size | Target), dataset, family = "binomial")

Table 8:
Model: lmer(Target_Response_Time ~ Relatedness + Prime_Response_Time + Prime.Accuracy + Session + Target_length + (1 + Relatedness + Target_length + Prime.Response_Time + Session | id) + (1 | Target), dataset)

Table 9:
Model: glmer(Prime.Accuracy ~ Item_type * Schedule * Session + Prime.Response_Time + (1 + Item_type | id)+ (1 | Prime), dataset, family="binomial", control = glmerControl(optimizer = "bobyqa"))

Table 10:
Model: lmer(Prime.Response_Time ~ Item_type + (Session + Prime.Accuracy) + Prime_length + (1 + Item_type + Session + Prime.Accuracy | id) + (1 + Prime.Accuracy | Prime), dataset)