This is an accepted manuscript published in *Language Learning*. Please cite as:

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning, Early View*. doi:10.1111/lang.12236

The Optimal Distribution of Practice for the Acquisition of L2 Morphology: A Conceptual Replication and Extension

Yuichi Suzuki

Kanagawa University

Abstract

This study examined optimal learning schedules for second language (L2) acquisition of a morphological structure. Sixty participants studied the simple and complex morphological rules of a novel miniature language system so as to use them for oral production. They engaged in four training sessions in either shorter spaced (3.3-day interval) or longer spaced learning conditions (7-day interval). From the beginning of the third training session, the 3.3-day interval group started to provide more accurate target rules than the 7-day interval group. This superior performance by the 3.3-day interval group was maintained on both 7-day and 28-day delayed posttests with small to medium effect sizes. No significant difference was found between the two groups for utterance speed, nor did linguistic complexity exert an influence on the effectiveness of different distributions of learning conditions.

Keywords lag effects; second language; distribution of practice; linguistic complexity; replication; acquisition of morphology

Introduction

Engaging in repeated practice is an essential component for acquiring second language (L2) skills and further maintaining them for an extended period of time (DeKeyser, 2015; Ellis & Wulff, 2015). An important question regarding repeated practice is when L2 teachers and learners should repeat or recycle the same learning materials for learners' consolidation of knowledge. For instance, is it better to include temporal spacing between multiple learning sessions (distributed learning) than to concentrate learning sessions without spacing (massed learning)? Research in cognitive psychology has shown that distributed learning leads to better retention of knowledge than massed learning (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Janiszewski, Noel, & Sawyer, 2003; Toppino & Gerbier, 2014). This phenomenon is called the spacing effect. The spacing effect is robust. The most comprehensive meta-analysis showed that 259 out of 271 experiments supported the spacing effects (Cepeda et al., 2006). Spacing effects have been found in a number of learning domains such as verbal learning in the first language (e.g., Dempster, 1987), in math (e.g., Rohrer & Taylor, 2006), in physics problems (e.g., Grote, 1995), and in L2 vocabulary learning (e.g., Bloom & Shuell, 1981; see Nakata, 2015, for a recent review on spacing effects in L2 vocabulary learning).

In a related line of investigations, researchers have been interested in identifying the optimal amount of spacing or inter-session intervals (ISIs). Instead of comparing massed and distributed learning, they have focused on the effect of two different distributed learning conditions on the acquisition of knowledge and skills. Comparison between two different distributed learning conditions yield evidence of lag effects. A majority of prior cognitive psychology research on lag effects examined the effects of ISIs manipulated within a single-day study session (Cepeda et al., 2006). Fewer studies have examined the effects of ISIs that are

longer than days or weeks (e.g., Cepeda et al., 2009; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Küpper-Tetzel, Erdfelder, & Dickhäuser, 2014), which should be more applicable to a learning task such as L2 grammar learning because lessons are often distributed and reviewed over weeks. For instance, there are several studies on L2 grammar acquisition that have compared two distributed learning conditions longer than days or weeks (Bird, 2010; Rogers, 2015; Suzuki & DeKeyser, 2015). Following up on this prior research, the current study examined the optimal ISI for L2 morphological learning by comparing shorter and longer ISI conditions.

Research in cognitive psychology has also suggested that the optimal ISI varies depending on the retention interval (RI), which refers to the amount of time between the end of practice and the testing time (Cepeda et al., 2006, 2008, 2009; Rohrer & Pashler, 2007). For instance, Rohrer and Pashler suggested that the optimal ISI is approximately 10–30% of RI. Based on these optimal ISI-RI ratios, previous studies on L2 grammar learning compared a longer ISI condition that fell within the optimal range of ISI-RI ratios (e.g., 20%) and a shorter ISI condition that was outside the optimal range (e.g., 5%). Two studies (Bird, 2010; Rogers, 2015) showed an advantage for the optimal ISI condition in L2 grammar learning which had been predicted by findings in psychology research. In contrast, a recent study by Suzuki and DeKeyser (2015) failed to find an advantage for the optimal ISI condition for acquiring oral production of L2 morphological structures predicted by the previous findings. The findings of this study contradicted prior results and raised an important question as to whether optimal ISI might be influenced by factors other than RIs (e.g., types of learning tasks and L2 skills targeted). Given the inconsistent findings of Suzuki and DeKeyser's original research, the present study aimed to replicate and extend their study. The experiment reported here was deemed a

conceptual replication because it reexamined the previous findings using a different experimental design controlling for confounding factors in the original study (Porte, 2012). This study thus aimed to gain a better understanding of the effects of different learning distributions on L2 grammar acquisition.

ISI × RI Interaction in Cognitive Psychology Research

Cognitive psychology research has shown that the optimal ISI depends on a given RI (Cepeda et al., 2006; Küpper-Tetzel et al., 2014; Verkoeijen, Rikers, & Özsoy, 2008). The general relationship between the ISI and RI on memory performance has been depicted as an inverted U-shaped curve (Glenberg, 1976). Optimal ISI increases as RI increases; however, with a much longer ISI, retention of knowledge declines. Thus, there seems to be a certain level of ISI that can maximize the retention of knowledge for a given RI (Cepeda et al., 2006).

In order to identify the optimal ISI and RI ratios, Cepeda et al. (2008) conducted the most comprehensive web-based experiments to date. Participants learned obscure and trivial facts (e.g., "Norway consumes the most spicy Mexican food") in two study sessions. They were randomly assigned to different learning conditions in which six ISIs (0, 1, 2, 7, 21, and 105 days) and four RIs (7, 35, 70, and 350 days) were systematically manipulated. Both recall and multiple-choice tests were used to assess participants' retention of knowledge. The results from both types of tests showed that the optimal ISI increased as RI increased. Furthermore, the optimal ISIs for given RIs were identified, and both ISIs and RIs were adjusted to fit a mathematical model. For the recall test, a 3-day ISI was optimal for the 7-day RI (ISI-RI ratio of 43%); an 8-day ISI for the 35-day RI (23% ratio); a 12-day ISI for the 70-day RI (17% ratio); and a 27-day ISI for the 350-day RI (8% ratio). A slightly different pattern of results was obtained for the multiple-choice tests. A 1.6-day ISI was optimal for the 7-day RI (24%); a 7-day ISI for the 35-day RI (19%

ratio); a 10-day ISI for the 70-day RI (14% ratio); and a 25-day ISI for the 350-day RI (7% ratio). These results demonstrated that the important factor for selecting the optimal ISI is how long the knowledge should be retained, that is, the RI. Subsequent studies essentially replicated the joint effect of ISI and RI with paired-associate vocabulary learning tasks (Cepeda et al., 2009; Küpper-Tetzel et al., 2014). Whether these findings are generalizable to more complex learning tasks, such as L2 grammar learning, still remains unclear (Wulf & Shea, 2002). The study reported here examined the relationship between ISI and RI for L2 morphological learning by utilizing a set of optimal ISI-RI ratios identified from Cepeda et al.'s (2008) experiment.

The interaction between ISI and RI can be explained by a combination of theories proposed to account for spacing and lag effects (for reviews, see Delaney, Verkoeijen, & Spirgel, 2010; Maddox, 2016). In particular, study-phrase retrieval theories (e.g., Toppino & Bloom, 2002) and a recently extended reminding account (Benjamin & Tullis, 2010) can inform the discussion of how optimal ISI is influenced by RI across a wide range of tasks. Essentially, both the study-phase retrieval theory and reminding accounts stipulate that memory performance improves when a learner successfully retrieves (is reminded of) a previously learned item in a relearning session. In other words, when a learner fails to retrieve the previously learned item due to a very long ISI, the repeated practice is less effective or not effective at all. Furthermore, when the retrieval of the previous learned item is too easy due to a very short ISI, the repeated practice is less effective. The benefit of successful retrieval of the previous item is maximized when the retrieval is difficult enough for learners to process the items more effectively, which is based on the idea of desirable difficulty. whereby conditions that present difficulties for learners can enhance long-term learning (Bjork, 1994; Schmidt & Bjork, 1992). In sum, the study-phase

retrieval and reminding accounts can explain why the optimal ISI should not be too short (less desirable difficulty) or too long (more retrieval failure).

Optimal Practice Distribution for L2 Grammar Learning

Motivated by cognitive psychology research, three empirical studies have been conducted to investigate the lag effects on L2 grammar acquisition (Bird, 2010; Rogers, 2015; Suzuki & DeKeyser, 2015).[1] These experiments examined whether the optimal ISI-RI ratios (10–30%) suggested by Rohrer and Pashler (2007) applied to L2 grammar learning. In relation to multiple RIs, they all compared two ISIs: a longer ISI condition, which corresponds to optimal lag (e.g., 23% ratio) and a shorter ISI condition, which is outside the optimal range (e.g., 5% ratio). In these L2 studies, the longer ISI conditions were conventionally labelled as distributed learning condition, whereas the shorter ISI conditions were called massed learning condition. However, this labeling was misleading because massed learning usually refers to a learning condition in which no spacing (e.g., less than 1 second) is available between study sessions (Toppino & Gerbier, 2014). Therefore, throughout this study, distributed and massed learning conditions from the previous L2 studies are referred to as longer ISI and shorter ISI conditions, respectively. More importantly, the terms shorter and longer are used as relative terms. For instance, a 3-day ISI would be considered a shorter spacing condition in comparison to a 7-day ISI condition, whereas a 3-day ISI would be a longer spacing condition compared to a 1-day ISI. Optimal intervals cannot be determined only by the relative lengths of intervals, rather they should always be considered in relation to RIs.

Among these three studies, two showed that longer ISIs, which were optimal based on Rohrer and Pashler's (2007) findings, lead to better learning than shorter ISIs (nonoptimal) for L2 grammar acquisition (Bird, 2010; Rogers, 2015), supporting Rohrer and Pashler's original

findings. Bird's pioneering work compared shorter and longer ISI learning conditions for the acquisition of tense/aspect of the English verb forms (i.e., simple past/present perfect and present/past perfect). Two groups of L2 English learners were engaged in five 1-hour study sessions in an L2 classroom wherein one group studied under a 14-day ISI condition and the other group studied under a 3-day ISI condition. In the training sessions, both groups of learners performed a written error-correction task in which they indicated whether or not the tense or aspect of verb forms was correct. Two delayed posttests with a written error-correction task were administered 7 days and 60 days after the last training session. The results showed that, while no significant difference was found between the two groups on the 7-day delayed posttest (both groups' lags were nonoptimal), the optimal 14-day ISI condition (ISI-RI ratio = 23%) led to better retention of knowledge than the nonoptimal 3-day ISI condition (ISI-RI ratio = 5%) on the 60-day delayed posttest.

A more recent study by Rogers (2015) corroborated Bird's (2010) findings. Rogers compared the acquisition of L2 English syntax under longer ISI (five sessions with an average of 7-day intervals) and shorter ISI (five sessions with 2.25-day intervals) learning conditions. In the 15-minute training sessions, the learners read 40 target complex/cleft sentences (e.g., "Where Noora shops is in London") and answered yes/no comprehension questions (e.g., "Does Noora shop in Paris?"). The results from grammaticality judgment tests showed that the learners in the optimal lag (longer ISI) condition (ISI-RI ratio = 17%) outperformed those in the nonoptimal lag (shorter ISI) condition (ISI-RI ratio = 5%) on the 42-day delayed posttest.

In contrast, a recent study by Suzuki and DeKeyser (2015) found virtually no difference between longer and shorter ISI conditions in acquiring oral production of L2 morphological structures. In this laboratory experiment, 40 beginner-level English speakers learning L2

Japanese engaged in a variety of comprehension and production tasks to learn Japanese verb morphological structures (*–te* forms) expressing the present progressive. The learners were randomly assigned to either a 7-day ISI condition (two training sessions with a 7-day interval) or a 1-day ISI condition (two training sessions with a 1-day interval). They were tested on the retention of grammatical knowledge one week and one month later (7-day and 28-day RIs). For the 28-day RI, the 7-day ISI condition (ISI-RI ratio = 25%) was optimal; it was expected to lead to better learning outcomes than the 1-day ISI condition (ISI-RI ratio = 3%).

The target grammatical structure (i.e., *–te*) involved six allomorphic changes depending on the verb stem. For instance, the uninflected verb *nobor-u* ("to climb," r-stem verb), should be converted to *nobot-te*, and *migak-u* ("to polish," k-stem verb) should be converted to *migai-te*. In the training, the learners practiced these six morphological rules with 18 unfamiliar verbs (three verbs for each verb-stem category). In contrast to the prior research (Bird, 2010; Rogers, 2015), the practice materials involved aural comprehension and oral production of the present progressive. Two types of oral production tests were employed. A picture description test was used to assess whether the learners were able to use the correct *–te* form of the verbs that they had practiced, and a rule application test presented pseudo verbs that the learners were unfamiliar with and asked them to apply *–te* form rules. Learners' responses were measured in terms of accuracy (whether the present progressive forms were used correctly) and speed (how fast the accurate sentence was uttered). The measure of utterance speed indexed cognitive fluency or procedural knowledge and the degree of automatization (Segalowitz, 2010), while accuracy scores likely tapped declarative and procedural knowledge (DeKeyser, 2015).[2]

The results showed that both 1-day and 7-day ISI conditions resulted in no difference in any accuracy measure and most of the speed measures on both delayed posttests (7-day and 28-

day RIs). The 1-day ISI group, however, significantly outperformed the 7-day ISI only on the speed measure in the picture description test (28-day RI). This contradicted the predictions based on the optimal ISI-RI ratios and previous findings because the nonoptimal and much smaller ISI-RI ratio in the 1-day ISI condition (3%) led to better performance. This finding might be attributable to the cognitive processes assessed in Suzuki and DeKeyser's (2015) study. These researchers interpreted their results as indicating that faster utterance production (a more sensitive measure of procedural/automatized knowledge, compared to accuracy) could be attained more effectively through more concentrated practice (with a shorter ISI) and that procedural knowledge was less susceptible to memory decay (Kim, Ritter, & Koubek, 2013).

The Need for a Conceptual Replication and Extension of Suzuki and DeKeyser's (2015) Study

The findings obtained by Suzuki and DeKeyser (2015) contradicted those reported in previous research (Bird, 2010; Rogers, 2015), raising questions as to what extent Suzuki and DeKeyser's results are generalizable. Their experiment also needed to be revisited because the issue of identifying optimal practice distribution for L2 grammar learning is important for both practical and theoretical reasons (e.g., How should language teachers schedule or repeat grammar practice? How do different levels of learning distribution influence L2 learning processes and outcomes?). Therefore, the present study was designed to improve the original study in three aspects. First, Suzuki and DeKeyser did not control prior knowledge about the target structure (–*te* form) before the training because the participants were beginner-level learners of L2 Japanese who had (partially) learned the target structure in the classroom. The participants in the present study were trained on morphological structures in a novel miniature language, loosely based on Spanish, which had six allomorphic morphological rules depending on the verb ending

(described in detail below). This methodological change helped control for extraneous factors (i.e., no prior knowledge or no exposure outside of the laboratory), improving the study's internal validity (Mackey, 2012).

Second, the original study included a small number of participants (N = 40). For the present study, more participants were recruited based on priori power analysis (Cohen, 1988), which should increase the generalizability of findings. Third, compared to Bird's (2010) and Rogers' (2015) research, there were fewer training sessions and ISIs in Suzuki and DeKeyser's (2015) study, with only two training sessions and one ISI targeted. In contrast, Bird and Rogers employed five training sessions with four ISIs. Lag effects might be more likely to occur when training sessions are repeated more frequently, facilitating consolidation of knowledge through a larger number of practice sessions and ISIs (Bird). In the present study, the number of training sessions was doubled, compared to those used by Suzuki and DeKeyser, for a total of four 1-hour training sessions followed by two delayed posttest sessions.

In addition to focusing on the conceptual replication component, the present study also extended the original study to target the role of potential moderating factors. Suzuki and DeKeyser interpreted the discrepancy in findings between their study and prior research (Bird, 2010; Rogers, 2015) as being caused by the complexity of the training/assessment tasks, which moderated lag effects (For other potential moderating factors such as declarative/procedural knowledge, see Suzuki & DeKeyser, 2015). In support of this explanation, Donovan and Radosevich (1999), who reported a meta-analysis of 63 studies on distributed and massed practice, demonstrated that the advantage of distributed learning over massed learning decreased as task complexity increased from low (e.g., typing) to high (e.g., music memorization and performance). While benefits of distributed learning were previously reported for receptive,

error-correction tasks (Bird; Rogers), no advantages for distributed learning were found for the training and assessment tasks involved in sentence production, which could have been due to the higher complexity level of these learning tasks (Suzuki & DeKeyser). However, Donovan and Radosevich's findings should be interpreted with caution for two reasons. First, their classification of simple and complex tasks may not directly apply to the complexity involved in learning L2 grammar because many of the tasks in their meta-analysis (e.g., gymnastic skills, air traffic controller tasks) were different from L2 learning tasks. Second, Donovan and Radosevich examined spacing effects (i.e., comparing massed learning vs. distributed learning groups), whereas the present research investigated lag effects (i.e., comparing two distributed learning groups). Their findings on spacing effects may not be applicable to lag effects.

With these caveats, the current study examined linguistic complexity as a potential moderating factor of lag effects. There are a number of ways to operationalize complexity. In this study, the complexity of linguistic structure was manipulated because it has previously been found to influence L2 grammar acquisition (DeKeyser, 2005; Robinson, 1996). The complexity of morphological structure was defined as the number of transformations required to arrive at the correctly inflected form (Hulstijn & de Graaff, 1994; Spada & Tomita, 2010). Specifically, six types of allomorphic verb categories were used. Half of them required only a single transformation at the end of the verb, whereas the other half were more complex in that the transformation had to be applied twice, in the middle and at the end of the verb (see below). Although this operationalization of complexity may not be fully compatible with the definition of overall complexity by Donovan and Radosevich (1999), this operationalization was objective in that it was relevant to L2 morphological learning.

The Present Study

The present study was motivated by the inconsistent results for the effectiveness of distributed practice in L2 grammar learning. This study's main objective was to replicate the results of Suzuki and DeKeyser (2015) and to further explore the role of linguistic complexity for different levels of practice distribution. Sixty learners were trained on an element of a miniature language system (i.e., present progressive morphological markers) consisting of simple and complex rules. They all completed four 1-hour training sessions to learn morphological rules accompanied by new vocabulary and grammatical explanations, but half of the learners participated in the training sessions at an interval of seven days (7-day ISI learning group), whereas the other half received the same training sessions twice a week (3.3-day ISI learning group). The following two research questions were addressed:

1. Does the 3.3-day ISI group outperform the 7-day ISI group in the acquisition of L2 morphology?
2. Does the linguistic complexity of the morphological structure moderate lag effects?

One of the major changes in this research design was that the number of training sessions was increased from two to four sessions. It is possible that Suzuki and DeKeyer (2015) found diminished lag effect precisely because there were not enough repetitions of training sessions and ISIs. Therefore, because this study employed four training sessions, in line with previous research by Bird (2010) and Rogers (2015), one could expect the findings to support the results reported by these researchers. By contrast, if no difference were to be found between the two groups, even with more repeated sessions for oral grammar practice, this finding would support the original results of Suzuki and DeKeyser. Furthermore, it is also possible that, compared to a larger ISI-RI ratio, a smaller ratio might be more advantageous to speakers' faster production of

utterances, which would be consistent with Suzuki and DeKeyser's original findings. Procedural, automatized knowledge (captured by the utterance speed measure) could be attained more effectively through more condensed practice, and it might be less susceptible to memory decay (Kim et al., 2013). Finally, the present study included a manipulation of the rule complexity for the target morphological structures in order to examine the moderating role of complexity. Because the spacing effect had previously been found to decrease in learning which involves more complexity (Donovan & Radosevich, 1999), it was anticipated that lag effects might also decrease for the more complex rules compared with the simpler rules.

Methods

Participants

Sixty students at a private Japanese university (20 male, 40 female) participated in the study. Their mean age was 19.63 (*SD* = 0.96). A priori power analysis was conducted to compute the minimum number of participants required by using G* Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009). Because multiple statistical tests were conducted with different assessment tasks across different time intervals, the most interesting result was chosen as the base for the power analysis (Prajapati, Dunne, & Armstrong, 2010). This was the significantly superior performance of the 3.3-day ISI practice group, compared to the 7-day ISI practice group, on the speed measure in the picture description task on the most delayed test, that is, for the 28-day RI (see Suzuki & DeKeyser, 2015). To sufficiently detect a significant group difference with a power of .95 and an alpha level of .05, the estimated number was 54, with 24 and 30 participants in the 3.3- and 7-day ISI groups, respectively. In order to keep an equal number of participants for both groups, six participants were added to the 3.3-day ISI group.

None of the participants had learned Spanish before. Twenty-eight participants had studied a foreign language other than English at university (15 and 13 participants in the 3.3- and 7-day ISI groups, respectively). Overall, languages were equally distributed within the 3.3- and 7-day ISI groups: French (2 and 3), German (1 and 1), Russian (1 and 0), Chinese (5 and 2), Korean (5 and 6), and Hindi (0 and 1). One participant in the 3.3-day ISI learning group had studied both French and Korean. No significant difference existed for the mean length of study (in months) between the 3.3-day ISI group (M = 23.00, SD = 20.70) and the 7-day ISI group (M = 21.04; SD = 13.15), t (26) = 0.29, p = .77.

Target Structures

For this study, a miniature language called Supurango was created, loosely based on Spanish. Spanish was chosen because the words can be pronounced easily by Japanese speakers. The target grammatical structure was the present progressive, which was expressed by a morphological marking on a verb. As shown in Table 1, the language had six morphological rules depending on the verb ending, which corresponded to the same number of rules learned in Suzuki and DeKeyser's (2015) study. While three simple verb types (*–ar*, *–er*, *–ir*) required a change only in a verb ending, the other three complex verb types (*–as*, *–es*, *–is*) involved two transformations, one in the first vowel as well as one in the verb ending.

<COMP: Place Table 1 near here>

Four action verbs were chosen for each verb type, yielding a total of 24 verbs for training (see Appendix S1 in the Supporting Information online). These uninflected verbs were real Spanish verbs, but meaning was arbitrarily assigned to them.[3] For instance, *lavar* means "to wash" in Spanish, but it means "to laugh" in Supurango. Suzuki and DeKeyser presented a set of

18 verbs (three verbs for each of the six categories) to their participants, but the number of verbs was increased here because the length of the training was longer.[4]

Outcome Tests

The outcome tests were computerized and administered through DMDX (Forster & Forster, 2003), and the responses were audio recorded. Three types of tests were used: (a) a vocabulary test, (b) a rule application test, and (c) a present progressive test. No feedback was provided. The rule application and present progressive test procedures corresponded to the ones in Suzuki and DeKeyser's study. However, that study did not use a separate outcome test for assessing vocabulary knowledge. This study included the vocabulary test to assess vocabulary knowledge independently. The order of the test items was randomized for each test at each different testing time (Sessions 1 through 6), and the participants could not anticipate which verb item would be presented, which was different from the learning phase.

Vocabulary Test

The vocabulary test assessed the participants' knowledge of the 24 verbs in the training sessions. The participants were presented with a fixation cross on the computer screen followed by a Japanese word (e.g., *warau*, "to laugh"). They were asked to say the corresponding word in Supurango (e.g., *lavar*) as quickly as possible. Each test item was presented for 6 seconds after which the next item was automatically presented. Before the test, the participants were reminded that they had to use an uninflected form instead of a present progressive form. It took approximately 2.5 minutes to complete the test.

Rule Application Test

The rule application test assessed the extent to which the participants had learned the morphological rules independently from knowledge of vocabulary. For this test, new verbs were created based on the verbs in the training sessions, replacing the phonemes of the stem but keeping the number of syllables (e.g., the practiced verb *lavar* was changed to nonce verbs such as *nopar*). The task objective was to convert these unknown, uninflected verbs (e.g., *nopar*) to a present progressive form (e.g., *nopiando*) as quickly as possible. The participants heard a new uninflected verb through headphones and saw the spelling on the screen. They were then asked to change it to the present progressive form within 8 seconds. After the time limit, the next item was automatically presented. Twenty-four items (four verbs per category) were created and used for the posttests (see the list of all target verbs in Appendix S2 in the Supporting Information online). It took approximately 3 minutes to complete the test.

Present Progressive Test

The present progressive test assessed the extent to which the participants could use the correct present progressive form of the 24 verbs that they had practiced. They were presented with pictures in which a man was shown performing various activities. These pictures were still images adapted from the learning sessions, ensuring that the participants knew what the picture was depicting. As in the rule application test, 8 seconds were given for each test item. It took approximately 3 minutes to complete the test.

Research Design

Participants individually engaged in four training sessions and two delayed test sessions. They were randomly assigned to the following two groups that received identical treatments while their training intervals were manipulated: a 3.3-day ISI learning group or a 7-day ISI group. As shown in Figure 1, participants in the 3.3-day ISI learning group engaged in the four training

sessions on Mondays and Thursdays or on Tuesdays and Fridays (Day 1, 4, 8, and 11), whereas those in the 7-day ISI learning group engaged in the session once a week (Day 1, 8, 15, and 22). Posttests 1 and 2 were administered seven days and 28 days after Session 4, respectively. Six participants (four learners in the 3.3-day ISI group and two learners in the 7-day ISI group) could not follow the same schedule due to unexpected events (e.g., absence due to sickness). They were excluded from the analyses in order to adhere strictly to the specified learning schedules. The pattern of results did not change when these six participants were included.

<COMP: Place Figure 1 near here>

The ISIs and RIs were determined based on Cepeda et al. (2008), one of the most comprehensive studies conducted to date.[5] The optimal ratios identified by Cepeda et al. ranged from 8% to 43%. Combinations of ISI and RI that were most relevant to the current study were chosen as benchmarks. A 3-day ISI was deemed optimal for the7-day RI (43% ratio), and a 8-day ISI was deemed optimal for the 35-day RI (23% ratio) on the recall test. The optimal ISI-RI ratios for the recall test were selected as a reference (rather than those based on the multiple-choice test) because the recall test format was more similar to the current test format.

The ISI and RI were manipulated so that the ratio of only one of the groups was in close proximity to the optimal ratio (Suzuki & DeKeyser, 2015). As shown in Table 2, on Posttest 1 (i.e., 7-day RI), the ISI-RI ratio for the 3.3-day ISI learning group was closer to the optimal ratio of 43% than that for the 7-day ISI learning group. In contrast, the ratio of the 7-day ISI learning group was closer to the optimal ratio of 23% on Posttest 2 (i.e., 28-day RI) than that of the 3.3-day ISI learning group. Based on these ISI-RI ratios, it was predicted that the 3.3-day ISI learning group would outperform the 7-day ISI learning group on Posttest 1, but that the 7-day ISI learning group would outperform the 3.3-day ISI learning group on Posttest 2. These ISIs and

RIs were identical to those employed in Suzuki and DeKeyser (2015), except for the 3.3-day ISI. This change from the original study was made for a practical reason. Because this study used four training sessions rather than two, it would have been impossible to recruit participants able to attend four consecutive days in the space of one week. Instead of testing the exact same ISI-RI ratios used in the original study, this conceptual replication project examined the same underlying mechanisms, that is, it compared different ISI-RI ratios in L2 morphological learning.

<COMP: Place Table 2 near here>

Training Procedure

All the training sessions were computerized and conducted using DMDX (Forster & Forster, 2003). Table 3 provides an overview of the training procedures. Following Suzuki and DeKeyser (2015), the target structure was taught in an explicit step-by-step manner: (a) vocabulary learning, (b) providing explicit grammatical explanations about the present progressive, and (c) oral practice using the present progressive. The learning procedure closely corresponded to Suzuki and DeKeyser's, but one modification was made to the grammar practice task. Suzuki and DeKeyser included an interactive task where the experimenter provided a recast as a form of feedback to incorrect responses, and the number of recasts could vary within each group. In this study, all the tasks were computerized and administered to each participant in the same way, ensuring the equal instances of feedback that all participants received.

<COMP: Place Table 3 near here>

Vocabulary Practice Phase

In the vocabulary practice phase, the participants saw a picture depicting an action verb with its Japanese translation placed in the top right corner and were prompted to say the Supurango word. The picture was presented for 4 seconds, followed by the feedback. Feedback was always given

in correct form both orally and visually and remained on the screen for 4 seconds. The set of 24 verbs was repeated three times in Training Session 1 and four times in Training Sessions 2 through 4. Only in the beginning of Session 1 did the participants study two cycles of 24 verbs by seeing and repeating each Supurango verb presented both orally and visually, along with its Japanese translation, for 5 seconds. This procedure was followed only in Session 1 so that participants could first learn the association before the retrieval practice. The presentation order of the vocabulary items was fixed throughout the entire experiment equally for both groups; a verb from each category was presented in sequence (i.e., *–ar*$_1$, *–er*$_1$, *–ir*$_1$, *–as*$_1$, *–es*$_1$, *–is*$_1$, *–ar*$_2$, *–er*$_2$, *–ir*$_2$, *–as*$_2$, *–es*$_2$, *–is*$_2$…).

Verb Conjugation Sheet

After the vocabulary learning phase, the participants were provided with a sheet that explained verb conjugations (see Appendix S3 in the Supporting Information online). The participants were encouraged to refer to it any time during the grammar practice in all sessions. In Session 1 only, participants also read a series of slides that explained the conjugations for each category after receiving the explicit information sheet.

Grammar Practice Phase

In the grammar practice phase, participants saw an animation video in which a man performed the action of the verbs. Each video clip lasted 8 seconds, and participants had to orally describe the animation using the present progressive form of the verb. The training tasks were designed to gradually increase in difficulty in three steps. First, each video clip showed an uninflected verb in the top right corner for the entire duration of the video clip (i.e., 8 seconds). The uninflected verbs appeared on the screen so that participants were able to practice using morphological rules while seeing the lexical items.[6] This set of 24 verbs was presented in the animation once. Second,

each video clip started the animation without presenting an uninflected verb for the first 4 seconds, and the verb appeared in the right top corner as a hint for the last 4 seconds. The first 4 seconds encouraged participants to retrieve a lexical item without any aid, making the task more engaging. This second training set was repeated twice. In the first and second steps, the presentation order of the verbs was blocked, meaning that all the verbs from each category were presented in the sequence such as *ar*$_1$, *–ar*$_2$, *–ar*$_3$, *–ar*$_4$, *–er*$_1$, *–er*$_2$, *–er*$_3$, *–er*$_4$. This allowed participants to apply morphological rules to uninflected verbs more easily. In the third stage, the animation was identical to the second (i.e., an uninflected verb appeared after 4 seconds), but the order of the presentation was changed; a verb from each category was presented as a set in the sequence *–ar*, *–er*, *–ir*, *–as*, *–es*, *–is*, followed by a second, third, and fourth set in the same order.

Learning Monitoring Tests

In the present study, learning monitoring tests were added as another component different from the procedure in the original study. These monitoring tests were the same three tests used in Posttests 1 and 2 (i.e., vocabulary, rule application,[7] and present progressive). These monitoring tests were given in the beginning of Training Sessions 2–4 (Monitoring Tests 2A, 3A, and 4A) and at the end of each training session (Monitoring Tests 2B, 3B, and 4B). The objective of these tests was to document the acquisition, retention, and forgetting of learning materials during the training sessions. No feedback was provided during the tests.

Data Analysis

Four independent trained raters coded the outcome tests using the sound analysis software Praat (Boersma & Weenink, 2016). Two raters coded the same set of data (16% of each test) until their coding matched. For all three tests, accuracy and speed were coded following the procedures used in Suzuki and DeKeyser (2015). Accuracy was scored as all or nothing for each test item,

and utterances with repairs (i.e., multiple attempts) were scored based on the final utterance. For example, when a response started with a wrong word but the repair was correct (e.g., *avenzo… avanzar*), credit was given. For the speed measures, response time (RT) was measured from the onset of the prompt to the end of the utterance. No RT measures were coded when (a) the response was incorrect or (b) the response included repairs and/or rephrasing. For a reliable computation of the average RT for each participant, participants were excluded from the speed analysis if their correct response rate (accuracy) was below 33% (as in Suzuki and DeKeyser). In addition, the outlying RT responses were excluded following Suzuki and DeKeyser: RTs below a minimum of 500 milliseconds and RTs higher than three standard deviations above the grand mean for each participant (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013). The outlying responses accounted for 8.5% of the vocabulary tests, 1% for the rule application tests, and 0.6% for present progressive tests.

Internal consistency (Cronbach's alpha) was computed for each posttest. For Posttest 1, Cronbach's alpha was .87, .93, and .89 for the accuracy measure and .77, .64, and .91 for the speed measures in the vocabulary, rule application, and present progressive tests, respectively. For Posttest 2, Cronbach's alpha was .81, .90, and .84 for accuracy and .94, .70, and .92 for speed in the vocabulary, rule application, and present progressive tests, respectively.

Results

Accuracy Measures

*Performance Change During Training Phase (Learning Monitoring Tests***)**

Figure 2 presents a graph of the mean accuracy scores and 95% confidence intervals (CIs) for the 3.3-day ISI and 7-day ISI groups in the three tests used throughout the experiment. The accuracy performance was first examined during the training phase (from Monitoring Test 1 to Monitoring

Test 4B). The 95% CIs indicate that the participants in the 3.3-day ISI group started to outperform those in the 7-day ISI group about the same time in the three tests (i.e., on Monitoring Test 3A, administered in the beginning of the third training session). From that point on, the 3.3-day ISI group consistently obtained higher accuracy scores until the end of Training Session 4 (see Appendix S4 in the Supporting Information online for the results during the training sessions).

<COMP: Place Figure 2 near here>

In order to examine whether the level of linguistic complexity influenced performance on the learning monitoring tests, accuracy was graphed separately for the two groups for simple and complex structures in the rule application and present progressive tests (see Figure 3 and Appendix S4). Accuracy scores during the training phase were first examined for the rule application tests. The CI error bars for the simple structures overlapped between the two groups. In contrast, there were gaps between the CI bars of the two groups for the complex structures, particularly on Monitoring Tests 3B and 4A. This suggests that the advantage of the 3.3-day ISI group was more pronounced for the complex than simple structures. In the present progressive test, on the other hand, there was no difference between the simple and complex structures during the training sessions.

<COMP: Place Figure 3 near here>

Posttest Performance and Logistic Mixed Effects Model

In order to statistically test the effects of learning conditions, the accuracy scores on Posttests 1 and 2 were analyzed for each performance test using a logistic mixed effects model (mixed logit model), implemented through the lme4 software package in R (Bates, Mächler, Bolker, & Walker, 2014).[8] The dependent variable was a binary response (correct/incorrect). Group (i.e.,

3.3- and 7-day ISI) and time (i.e., 7- and 28-day RIs) were modeled as fixed effect variables (for full results, see Appendix S5 in the Supporting Information online). For the rule application and present progressive tests, complexity (i.e., simple and complex) was added as a fixed effect variable. These fixed effects factors were centered using deviation coding (–.5, .5) in order to match the inferences drawn from ANOVA (Linck & Cunnings, 2015). Participants and items were treated as random effects. The models included the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013). The effect sizes were interpreted using the following criteria (Plonsky & Oswald, 2014): small (d = 0.4), medium (d = 0.7), and large (d = 1.0).

Vocabulary Test

The mixed effects logit model for the vocabulary test revealed a significant fixed effect for group, z = –2.37, p = .02, as well as for time, z = –4.23, p < .001 (for full results of mixed effects models, see Appendix S6 in the Supporting Information online). No significant interaction was found. Post hoc comparison showed that the 3.3-day ISI group scored significantly higher than the 7-day ISI group, with small effect sizes on Posttest 1, z = 2.37, p = .02, d = 0.54, and on Posttest 2, z = 2.24, p = .03, d = 0.48.

Rule Application Test

The mixed effects logit model for the rule application test showed a significant fixed effect for group, z = –2.74, p = .01, and time, z = –3.18, p < .001. No significant interaction was found. Post hoc comparison showed that the 3.3-day ISI group scored significantly higher than the 7-day ISI group with small to medium effect sizes on Posttest 1, z = 2.37, p = .02, d = 0.60, and on Posttest 2, z = 1.99, p = .047, d = 0.49. With regard to the effect of linguistic complexity, the

interaction between complexity and group was not significant, $z = 1.28$, $p = .20$, indicating that the complexity of linguistic structure did not moderate the lag effects.

Present Progressive Test

The mixed effects logit model showed a significant fixed effect for group, $z = –2.25$, $p = .02$, and time, $z = –2.28$, $p = .02$. No significant interaction was found. Post hoc comparison showed that the 3.3-day ISI group scored significantly higher than the 7-day ISI group with small to medium effect sizes on Posttest 1, $z = 2.50$, $p = .01$, $d = 0.60$, and on Posttest 2, $z = 2.37$, $p = .02$, $d = 0.52$. With regard to the effect of linguistic complexity, the interaction between complexity and group was not significant, $z = 0.12$, $p = .90$, indicating that the complexity of the linguistic structure did not moderate the lag effects.

Speed Measures

Performance Change During Training Phase (Learning Monitoring Tests)

Figure 4 presents a graph of the mean RTs and associated 95% CIs for the three tests across time. In contrast to the pattern for the accuracy scores, the two groups of learners showed a similar level of speed in all the tests. Although the learners in the 3.3-day ISI group achieved faster performance starting in the third session (Monitoring Test 3A) at the descriptive level, all the error bars overlapped between the two groups both during the training phase (see Appendix S4).

<COMP: Place Figure 4 near here>

The performance on the speed measures showed a similar pattern for the simple and complex structures in both tests (see Figure 5). No reliable group difference was found for any test or time point with all the error bars overlapping, except for present progressive Test 3A for the complex structures (see Appendix S4).

<COMP: Place Figure 5 near here>

Posttest Performance and Linear Mixed Effects Model

A similar statistical approach to the accuracy measures was employed to examine the effects of learning conditions for the speed measures on Posttests 1 and 2. The data were analyzed for each test using a linear mixed effects model, which was used for continuous dependent variables (RT). All fixed and random effect variables were identical to those discussed for the logistic mixed effects models.

Vocabulary Test

Except for the fixed effect of time, none of the factors or interactions were significant in the model (see Appendix S6). Critically, no significant fixed effect was found for group, $t = 1.17$, $p = .25$, indicating that both groups showed a similar level of performance in terms of speed.

Rule Application Test

The linear mixed effects model for the rule application test yielded a similar pattern to that of the vocabulary test. No significant fixed effect for group was found, $t = 1.25$, $p = .22$. Furthermore, no significant interaction was found between group and complexity, $t = –0.63$, $p = .53$. Except for the fixed effect of time, none of the factors or interactions were significant in the model.

Present Progressive Test

The same results were obtained in the linear mixed effects model for the present progressive test. No significant fixed effect for group was found, $t = 1.48$, $p = .15$. Furthermore, no significant interaction was found between group and complexity, $t = –0.51$, $p = .61$. Except for the fixed effect of time, $t = 2.40$, $p = .02$, none of the factors or interactions were significant in the model.

Discussion

The present study was designed as a conceptual replication and extension of Suzuki and DeKeyser's (2015) experiment. In order to facilitate the comparison between their study and the present one, two relevant findings from the original study are summarized. First, Suzuki and DeKeyser found no difference between the 1-day ISI and the 7-day ISI groups in terms of the accuracy scores in the rule application test and the present progressive test (Cohen's $d = –0.07$ and 0.12, respectively). Second, Suzuki and DeKeyser reported that the learners in the 1-day ISI group, which corresponded to a nonoptimal condition (based on Rohrer & Pashler, 2007), significantly outperformed those in the 7-day ISI group (optimal) only on the speed measure in the present progressive test (Cohen's $d = –1.01$). In the present study, the 3.3-day ISI group (which was equivalent to the 1-day ISI group in Suzuki and DeKeyser's study) also outperformed the 7-day ISI group in accuracy scores ($0.48 < d < 0.60$) but not in speed measures. This study further attempted to investigate whether linguistic complexity moderated the effects of different levels of distributed practice. However, the influence of linguistic complexity, as operationalized here, was found to be statistically nonsignificant.

In addition to the increased sample size, two factors were controlled more rigorously in this replication study: (a) prior knowledge of grammatical rules and (b) number of training sessions and ISIs. First, the amount of prior knowledge for the target structure may influence outcomes. Using a novel miniature language, the participants started with the same level of knowledge. In addition, because the miniature language was not available to participants outside the laboratory, they were less likely to study the target linguistic items outside of the experiment. The present findings thus provide more reliable evidence, compared to those reported by Suzuki and DeKeyser (2015). Second, increasing the number of training sessions could favor the 7-day

ISI learning condition for the 28-day RI because more frequent spacing might facilitate the consolidation of linguistic knowledge (cf., Bird, 2010; Rogers, 2015). However, the present findings seemed to indicate the opposite pattern, namely, that the 3.3-day ISI learning condition led to better performance than the 7-day ISI learning condition.

The effect sizes in this study were interpreted using Plonsky and Oswald's (2014) field-general benchmarks (as discussed above), but they were also compared with the effect sizes of lag effects found in similar cognitive psychology and L2 acquisition research. First, although there is little cognitive psychological research that has directly compared learning conditions with nonzero ISIs that are longer than a day, a study by Küpper-Tetzel and Erdfelder (2012) showed that the optimal ISI condition yielded a superior performance in cued recall (Cohen's $d \geq 0.66$) (see also Küpper-Tetzel et al., 2014).[9] This value is closer to the effect sizes for accuracy scores on Posttest 2 ($0.48 < d < 0.52$). Second, the two previous L2 studies found a larger effect size for the optimal ISI (i.e., larger ISI) condition: Cohen's *d* was 2.02 in Bird (2010) and 0.94 in Rogers (2015). Given these effect sizes for the optimal ISI condition, the magnitude of advantage for the 3.3-day ISI learning condition seems modest. Because little research is available at this moment, more empirical evidence needs to be accumulated so that the overall lag effects can be estimated from meta-analyses (see Hattie, 2009, for a synthesis of meta-analyses on spacing effects).

Optimal Learning Distribution for L2 Grammar Learning

This study showed that the 3.3-day ISI group outperformed the 7-day ISI group for accurate oral production of vocabulary and morphological structures. The present findings essentially replicated Suzuki and DeKeyser's (2015) findings, although the results for accuracy and speed were different one from the other. One possible reason for why no group difference was found in

the speed measures in this study may be due to the amount of practice required for proceduralization and automatization (DeKeyser, 1997). In this study, the learners from both groups had sufficient practice for proceduralizing their knowledge to a larger extent (with four 1-hour sessions) than those in the previous study (with two 1-hour sessions). With an insufficient amount of practice and longer spacing, the learners in the longer ISI group (7-day ISI) in the previous study might not have been able to proceduralize their knowledge to the extent that they could retain their speed of access to their knowledge. This interpretation is speculative, but it suggests that there may be a complex interplay among lag effects, amount of practice, and types of knowledge (tapped by accuracy and speed measures).

In contrast to the speed measures, the 3.3-day ISI group achieved significantly higher accuracy scores than the 7-day ISI group on Posttest 2, where ISI-RI ratios were 13% and 25%, which is inconsistent with findings from previous L2 research (Bird, 2010; Rogers, 2015). One of the most plausible explanations likely relates to the complexity of the skill to be learned (Suzuki & DeKeyser, 2015). In the experiments by Bird (2010) and Rogers (2015), the learners only engaged in receptive written practice, whereas the learners in the present study performed oral production tasks in which multiple processes had to be executed in parallel to produce accurate utterances, such as conceptualization of the meaning, retrieval of lexical/grammatical information, and articulation (Levelt, 1989). The higher complexity of L2 oral production thus increases the difficulty of the learning task, and learners can be more susceptible to skill loss in the longer spacing condition. This could be explained by study-phase retrieval or reminding theories (Benjamin & Tullis, 2010; Toppino & Bloom, 2002). As shown in Figure 2, the 7-day ISI group suffered greater memory loss between the training sessions (particularly from Session 2 to Session 3 and from Session 3 to Session 4) than did the 3.3-day ISI group. This suggests that

the learners in the 7-day ISI group found it difficult to retrieve the lexicon and grammatical rules in subsequent learning sessions. This retrieval failure might have decreased the benefit of more widely spaced practice. More successful retrieval, which still imposed a certain level of difficulty, might have facilitated L2 morphological learning (Bjork, 1994; Schmidt & Bjork, 1992).

One of the anonymous reviewers suggested reanalyzing the current data by applying a learning criterion threshold in the training phase so that participants who did not show a certain level of accuracy at the end of the first training session would be excluded from further analysis (e.g., Cepeda et al., 2006). From a pedagogical viewpoint, the analysis including all the participants' data provided results that are more relevant to instructional settings (because not all learners can reach a certain criterion threshold of learning in the beginning). However, this additional reanalysis can ensure that a certain amount of learning actually took place in the very beginning stage. Because participants with higher accuracy in the first learning session are more likely to succeed in retrieving previously learned items in subsequent training sessions (Benjamin & Tullis, 2010; Toppino & Bloom, 2002), different learning curves may be found. The learning criterion was set at 25% in accuracy scores on Monitoring Test 1 to ensure a sufficient number of participants for statistical analysis (n = 29, 37, 23 for the vocabulary, rule application, and present progressive tests, respectively). The results showed a similar pattern to those from the original analysis in terms of effect sizes (see Appendixes S7 and S8 in the Supporting Information online).

Relationship between ISIs and RIs: Potential Confounding Factor

Although Bird (2010) and Rogers (2015) found optimal ISI-RI ratios (i.e., 17% and 23%) that were consistent with those in psychology research (Cepeda et al., 2008, 2009; Rohrer & Pashler, 2007), the smaller ISI-RI ratios (3% and 12%) yielded better performance in this study and in

Suzuki and DeKeyser's (2015) research. These contradictory findings may be explained as a methodological issue.[10] In L2 studies of distributed learning, multiple posttests have usually been administered to assess the retention of knowledge within individuals, the same learners taking two delayed posttests after completing all the training sessions (Bird, 2010; Suzuki & DeKeyser, 2015). However, as several anonymous reviewers noted, this can confound the current research design. Posttest 1 could represent an opportunity to practice retrieval of learned items, and it may have influenced the retention of knowledge assessed by Posttest 2. To address this concern, only one of the posttests should have been administered to the same participant. RI should have been manipulated as a between-subjects factor (e.g., Cepeda et al., 2008). It is thus worth taking into account this additional opportunity for retrieval practice. ISI and RI ratios were recalculated for the previous and current studies (see Table 4). In this study, where Posttest 1 was treated as another training session, the average ISI increased to a 4.25-day ISI (3 + 4 + 3 + 7/4) for the shorter ISI learning group, and no change was needed for the longer ISI learning group (i.e., 7-day ISI). The recalculated RI was 21 days for both groups, and the recalculated ISI-RI ratios were 20% and 33% for the shorter and longer ISI groups, respectively. For ease of comparison, ISI-RI ratios were recomputed for the two previous studies in the same way (Bird, 2010; Suzuki & DeKeyser, 2015).

<COMP: Place Table 4 near here>

As a result of this post hoc analysis, superior performance was observed in learning conditions where the ISI-RI ratios were close to 20% across studies. This value is similar to the ones obtained by Cepeda et al. (2008), who demonstrated that optimal ISI-RI ratios were 17–23% for the recall test and 14–19% for the recognition test for 35-day and 70-day RIs.[11] This suggests the possibility that the optimal ISI-RI ratios for L2 grammar learning are close to the

ones found in Cepeda et al.'s study and even in broader cognitive psychology research. As Posttest 1 can never be considered a full learning session, the pattern found from these recalculations is preliminary but provides a promising direction for further research. Future L2 research should employ multiple RIs as a between-subjects factor to confirm and extend the current findings. Additionally, it is recommended that a wider range of ISIs and RIs be manipulated, as was the case in the comprehensive study by Cepeda et al. (2008), so as to provide a better understanding of the extent to which optimal ISIs and RIs can change for L2 grammar learning.

Limited Effects of Linguistic Complexity on Different Levels of Learning Distribution

In addition to the conceptual replication component, the present study explored the relationships between the lag effect and complexity level. Overall, linguistic complexity appeared to have no influence on the lag effect. The level of linguistic complexity seemed to moderate the effects of the two distributed learning conditions more strongly during the learning phase. The advantage of the 3.3-day ISI condition was more evident for the complex structures in Monitoring Tests 3B and 4A (see the rule application test performance in Figure 3). Complexity of learning was operationalized in this study as structural complexity, that is, the number of transformations applied to the present progressive form. The limited effect of structural complexity on the lag effect found in this study is not consistent with Donovan and Radosevich's (1999) findings. Their meta-analysis included many nonlinguistic tasks, and their classification of simple and complex tasks may not be applicable to L2 grammar learning. Although Donovan and Radosevich's findings has been frequently cited by previous L2 research to explain complexity as a potential moderating factor of spacing effects (e.g., Serrano, 2012), the applicability of these findings to L2 research may need to be reconsidered.

The current findings may be more closely compatible with the results of a meta-analysis by Janiszewski et al. (2003). These researchers found that the advantage of distributed learning over massed learning in verbal learning was pronounced, when compared with simple stimuli (e.g., the word *cat*), for semantically complex stimuli (e.g., homographs) but not for structurally complex target items (e.g., the sentence "The cat is on the red brick wall"). Although their findings were based on spacing effects, not lag effects, these findings suggest that other operationalizations for linguistic complexity may yield different results for complexity such as ones related to the abstractness of form-meaning connections (e.g., DeKeyser, 2005).

Further Research: Potential Moderating Factors

The present study suggested several factors that future research should explore for the effects of different levels of learning distribution. Among them, the type of linguistic knowledge (e.g., receptive vs. productive knowledge) may be an important moderating factor of optimal ISIs because this is one of the major differences that underlies the discrepancies found with previous research. Other factors are not limited to, but include, modality (written vs. spoken), amount of prior knowledge, the number of training sessions and ISIs, and linguistic knowledge domains (e.g., morphology, syntax, pragmatics). Furthermore, complexity still remains a potential factor that is worth further investigation. Different dimensions of learning complexity (e.g., task complexity) and operationalizations of linguistic complexity may interact with different levels of learning distribution. Future research should thus not only compare the effects of different learning distribution on skill/knowledge retention but also determine how multiple other factors might moderate lag effects.

Conclusion

The objectives of this study were to replicate and extend Suzuki and DeKeyser's (2015) original research in order to better understand the effects of learning distribution on L2 grammar learning. Although previous L2 research had revealed optimal ISI-RI ratios that were consistent with those found in cognitive psychology research (Bird, 2010; Rogers, 2015), Suzuki and DeKeyser demonstrated otherwise. By improving several aspects of the original research design, this study suggested that the optimal ISI-RI ratio may be influenced by factors pertaining to L2 skills, lending some support for Suzuki and DeKeyser's findings. Furthermore, this study found limited evidence for the interaction of rule complexity with different distributions of learning practice. These findings suggest that L2 grammar learning involves a number of factors, such as learning difficulty and targeted L2 skills, that may moderate the benefits offered by different ways in which learning could be distributed. The present study thus suggests promising avenues for further investigations into the role of different patterns of learning distribution targeting different aspects of L2 learning.

Notes

1 This review mainly focuses on L2 grammar learning related to classroom-level pedagogical decisions (e.g., a grammatical feature is reviewed in an interval of days or weeks). Another related area of research has examined the effects of time distribution at the program level (e.g., how to allocate 100-hour instruction over semesters), which is beyond the scope of this paper (for an extensive review of program-level studies, see Serrano, 2012). A study by Miles (2014) examined the spacing effect, rather than the lag effect, for L2 grammar learning by comparing the distributed learning condition (three study sessions spread over 5 weeks) and the massed learning condition (three sessions were massed in a single learning session).

2 Both accuracy and speed measures can index procedural knowledge because proceduralization refers to uttering accurate target sentences using the relevant declarative knowledge (DeKeyser, 2015). Speed measures can reflect the degree of proceduralization (which can be seen as a very initial stage of automatization) and automatization.

3 Some of the verbs are inflected in Spanish but treated as uninflected in this experiment.

4 As pointed out by anonymous reviewers, increasing the number of verbs in the training phase might have impacted the results (cf., Brooks, Kempe, & Sionov, 2006).

5 Previous L2 research (Bird, 2010; Rogers, 2015; Suzuki & DeKeyser, 2015) referred to Rohrer and Pashler's commentary paper (2007), which reported the preliminary results from the experiment by Cepeda et al. (2008) (D. Rohrer, personal communication, April 04, 2016).

6 Suzuki and DeKeyser's (2015) study did not present vocabulary during grammar practice, but this study presented an uninflected verb for every practice trial so that the learners in both groups would have an equal opportunity to practice present progressive forms.

7 A different set of 12 verbs (two verbs per category) were used for the monitoring tests during the training phase for every rule application test. The number of the items was decreased during the training sessions (i.e., for the monitoring tests) to reduce interference in learning of the actual verbs.

8 The mixed logit models were constructed separately for three tests because complexity was modeled for the rule and present progressive tests only.

9 Küpper-Tetzel et al.'s (2014) study is the only one that has directly compared memory effects of two different nonzero ISIs that were more than a day apart. The data were not available in Cepeda et al.'s (2008) experiment for computing the relevant effect sizes.

10 I thank the reviewers for pointing out this issue.

11 Cepeda et al. (2008) examined the other two RIs (7 days and 350 days), but the 35-day and 70-day RIs were used as reference points because the RIs were similar to the previous L2 studies: 28 days (Suzuki & DeKeyser, 2015, and the present study), 42 days (Rogers, 2015), and 60 days (Bird, 2010).

References

<**REF**>Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. doi:10.1016/j.jml.2012.11.001

<**REF**>Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67,* 1–48. doi:10.18637/jss.v067.i01

<**REF**>Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*, 228–247. doi:10.1016/j.cogpsych.2010.05.004

<**REF**>Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *31*, 635–650. doi:10.1017/S0142716410000172

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

<**REF**>Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research*, 245–248. doi:10.1080/00220671.1981.10885317

<**REF**>Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer (Version 6.0.14). Retrieved from http://www.praat.org

<**REF**>Brooks, P. J., Kempe, V., & Sionov, A. (2006). The role of learner and input variables in learning inflectional morphology. *Applied Psycholinguistics*, *27*, 185–209. doi:10.1017/S0142716406060243

<**REF**>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. doi:10.1027/1618–3169.56.4.236

<**REF**>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. doi:10.1037/0033–2909.132.3.354

<**REF**>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. doi:10.1111/j.1467–9280.2008.02209.x

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

<**REF**>De Jong, N., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*, 893–916. doi:10.1017/S0142716412000069

<**REF**>DeKeyser, R. M. (1997). Beyond explicit rule learning. *Studies in Second Language Acquisition*, *19*, 195–221.

<**REF**>DeKeyser, R. M. (2005). What makes learning second language grammar difficult? A review of issues. *Language Learning*, *55*, 1–25. doi:10.1111/j.0023–8333.2005.00294.x

<**REF**>DeKeyser, R. M. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 94–112). New York: Routledge.

<**REF**>Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, *53*, 63–147. doi:10.1016/S0079–7421(10)53003–2

<**REF**>Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*, 162–170. doi:10.1037/0022–0663.79.2.162

<**REF**>Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*, 795–805. doi:10.1037/0021–9010.84.5.795

<**REF**>Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 75–93). New York: Routledge.

<**REF**>Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. doi:10.3758/BRM.41.4.1149

<**REF**>Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124. doi:10.3758/BF03195503

<**REF**>Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*, 1–16. doi:10.1016/S0022–5371(76)90002–5

<**REF**>Grote, M. G. (1995). Distributed versus massed practice in high school physics. *School Science and Mathematics*, *95*, 97–101. doi:10.1111/j.1949–8594.1995.tb15736.x

<**REF**>Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

<**REF**>Hulstijn, J., & de Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review*, *11*, 97–112.

<**REF**>Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, *30*, 138–149. doi:10.1086/374692

<**REF**>Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, *20*, 37–47. doi:10.1080/09658211.2011.631550

<**REF**>Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, *42*, 373–388. doi:10.1007/s11251–013–9285–2

<**REF**>Kim, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science*, *14*, 22–37. doi:10.1080/1464536X.2011.573008

<**REF**>Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

<**REF**>Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207. doi:10.1111/lang.12117

<**REF**>Mackey, A. (2012). *Why (or why not), when and how to replicate research*. Cambridge: Cambridge University Press.

<**REF**>Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology, 28*, 684–706. doi:10.1080/20445911.2016.1181637

<**REF**>Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System*, *42*, 412–428. doi:10.1016/j.system.2014.01.014

<**REF**>Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*, 677–711. doi:10.1017/S0272263114000825

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

<**REF**>Porte, G. (2012). *Replication research in applied linguistics*. Cambridge, UK: Cambrige University Press.

<**REF**>Prajapati, B., Dunne, M., & Armstrong, R. (2010). Sample size estimation and statistical power analyses. *Optometry Today*, *16*, 10–18.

<**REF**>Robinson, P. (1996). Learning simple and complex second language rules under implicit, incidental, rule-search, and instructed conditions. *Studies in Second Language Acquisition*, *18*, 27–67. doi:10.1017/S0272263100014674

<**REF**>Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*, 857–866. doi:10.1002/tesq.252

<**REF**>Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, *16*, 183–186. doi:10.1111/j.1467–8721.2007.00500.x

<**REF**>Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*, 1209–1224. doi:10.1002/acp.1266

<**REF**>Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three taradigms suggest new concepts for training. *Psychological Science, 3*, 207–217. doi:10.1111/j.1467–9280.1992.tb00029.x

<**REF**>Segalowitz, N. S. (2010). *Cognitive bases of second language fluency*. New York: Taylor & Francis.

<**REF**>Serrano, R. (2012). Is time concentration good or bad for learning? Reflecting on the results from the cognitive psychology and from the SLA literature. In C. Muñoz (Ed.), *Intensive exposure experiences in second language learning* (pp. 3–22). Clevedon, UK: Multilingual Matters.

<**REF**>Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*, 263–308. doi:10.1111/j.1467–8721.2007.00500.x

<**REF**>Suzuki, Y., & DeKeyser, R. M. (2015). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*. Published online 20 November 2015. doi:10.1177/1362168815617334

<**REF**>Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 437–444. doi:10.1037/0278–7393.28.3.437

<**REF**>Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. *The Psychology of Learning and Motivation*, *60*, 113–189. doi:10.1016/B978–0–12–800090–8.00004–4

<**REF**>Verkoeijen, P. P., Rikers, R. M., & Özsoy, B. (2008). Distributed rereading can hurt the spacing effect in text memory. *Applied Cognitive Psychology*, *22*, 685–695. doi:10.1002/acp.1388

<**REF**>Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, *9*, 185–211. doi:10.3758/bf03196276

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. List of Verbs.

Appendix S2. Test Items in the Generalization Test.

Appendix S3. Explicit Grammatical Information Sheet.

Appendix S4. Results of *t* Tests Comparing Performance During Training Sessions.

Appendix S5. Group Comparisons for Posttests 1 and 2.

Appendix S6. Results of Mixed Effects Models.

Appendix S7. Data Reanalyses Using a Learning Criterion.

Appendix S8. Graphical Representation of Data for Reanalyses Using a Learning Criterion.

Table 1 Verb category and conjugation

Category	Complexity	Uninflected form	Present progressive
–ar	simple	lavar (laugh)	lav**iando**
–er	simple	poner (sleep)	pon**iendo**
–ir	simple	partir (dance)	part**iondo**
–as	complex	montas (clean)	m**a**nt**iando**
–es	complex	detenes (read)	d**i**ten**iendo**
–is	complex	recibis (smoke)	r**o**cib**iondo**

Table 2 Ratio of inter-session intervals (ISI) to retention interval (RI)

	7-day RI (Posttest 1)	28-day RI (Posttest 2)
3.3-day ISI	47.1%	11.8%
7-day ISI	100.0%	25.0%

Table 3 Training procedures used in the study

Session 1		Sessions 2–4	
Task	Length (minutes)	Task	Length (minutes)
1. Questionnaire and consent form	5	1. Monitoring Tests A (2A, 3A, 4A)	7
2. Vocabulary practice	14	2. Vocabulary practice	16
3. Explicit information sheet and explanation	5	3. Explicit information sheet	1
4. Grammar practice	20	4. Grammar practice	20
5. Monitoring Test 1	7	5. Monitoring Tests B (2B, 3B, 4B)	7

Table 4 Original and recalculated ISI-RI ratios from current and previous studies

		Original ISI-RI ratio		Recalculated ISI-RI ratio	
Study	Results	Shorter ISI	Longer ISI	Shorter ISI	Longer ISI
Bird (2010)	supported	5%	23%	7%	24%
Rogers (2015)[a]	supported	5%	17%	—	
Suzuki & DeKeyser (2015)	not supported	3%	25%	19%	33%
The present study	not supported	13%	25%	20%	33%

Note. [a]Only one delayed posttest was used in Rogers' (2015) study.

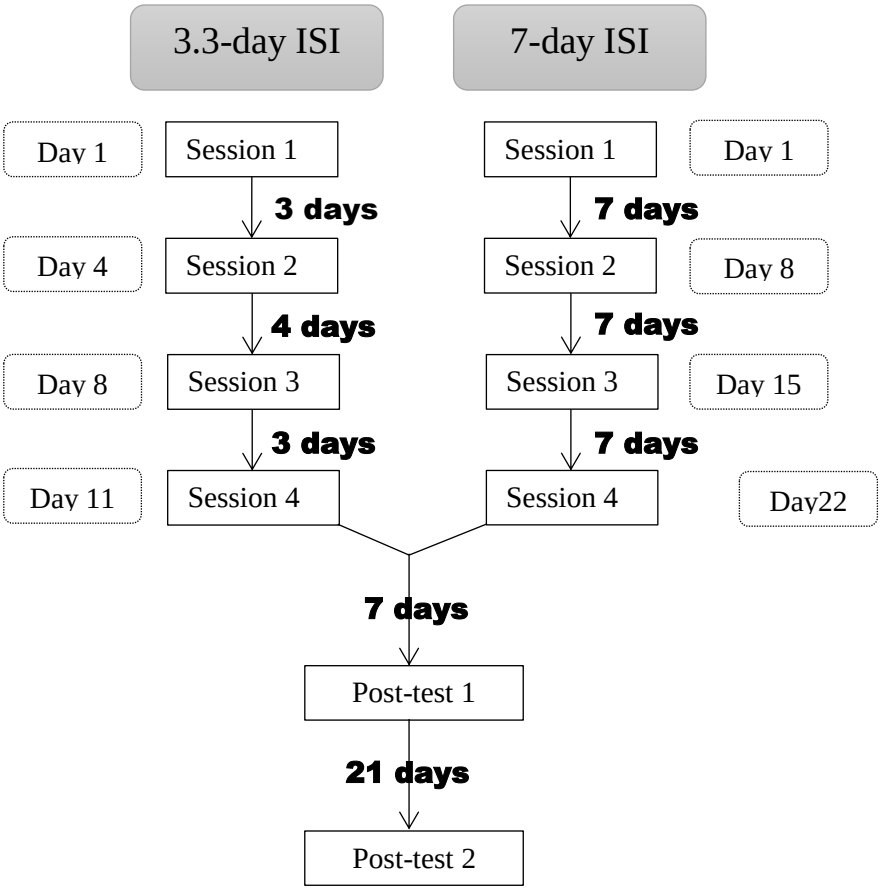
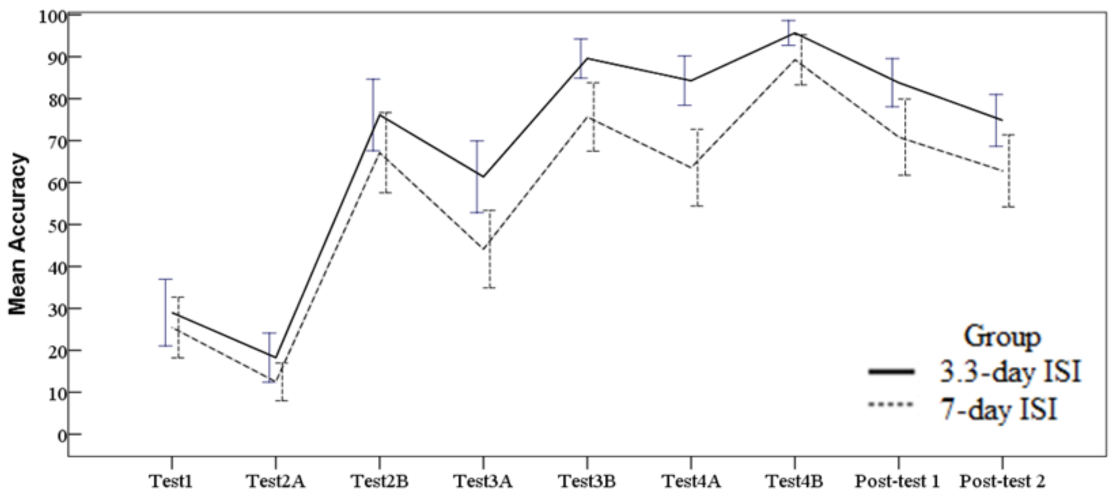
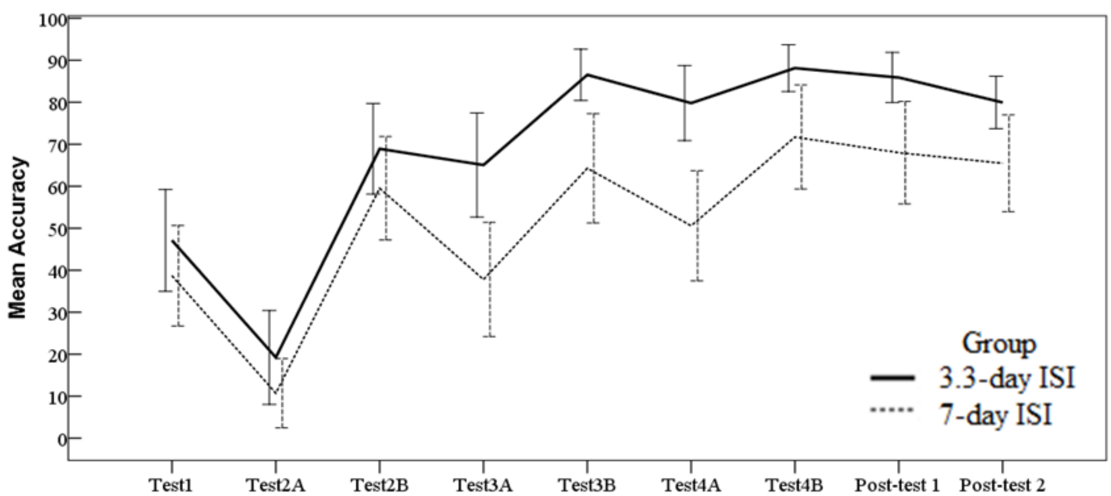


Figure 1 Research design.


Mean Accuracy
Group
3.3-day ISI
7-day ISI
Test1
Test2A
Test2B
Test3A
Test3B
Test4A
Test4B
Post-test 1
Post-test 2


Mean Accuracy
Group
3.3-day ISI
7-day ISI
Test1
Test2A
Test2B
Test3A
Test3B
Test4A
Test4B
Post-test 1
Post-test 2

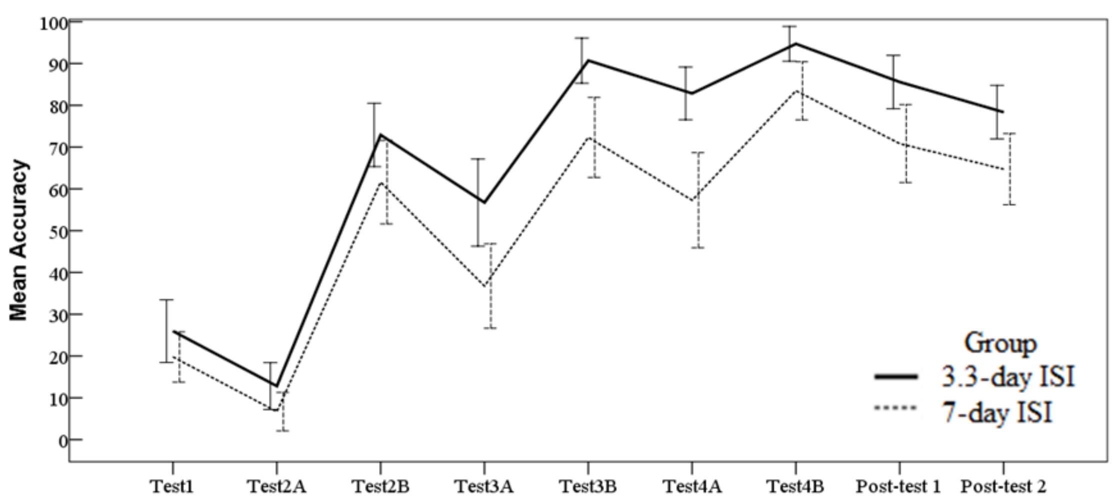

Mean Accuracy
Group
3.3-day ISI
7-day ISI
Test1
Test2A
Test2B
Test3A
Test3B
Test4A
Test4B
Post-test 1
Post-test 2

Figure 2 Accuracy scores and 95% confidence intervals for the two groups across time in the vocabulary test (top panel), rule application test (middle panel), and present progressive test (bottom panel).

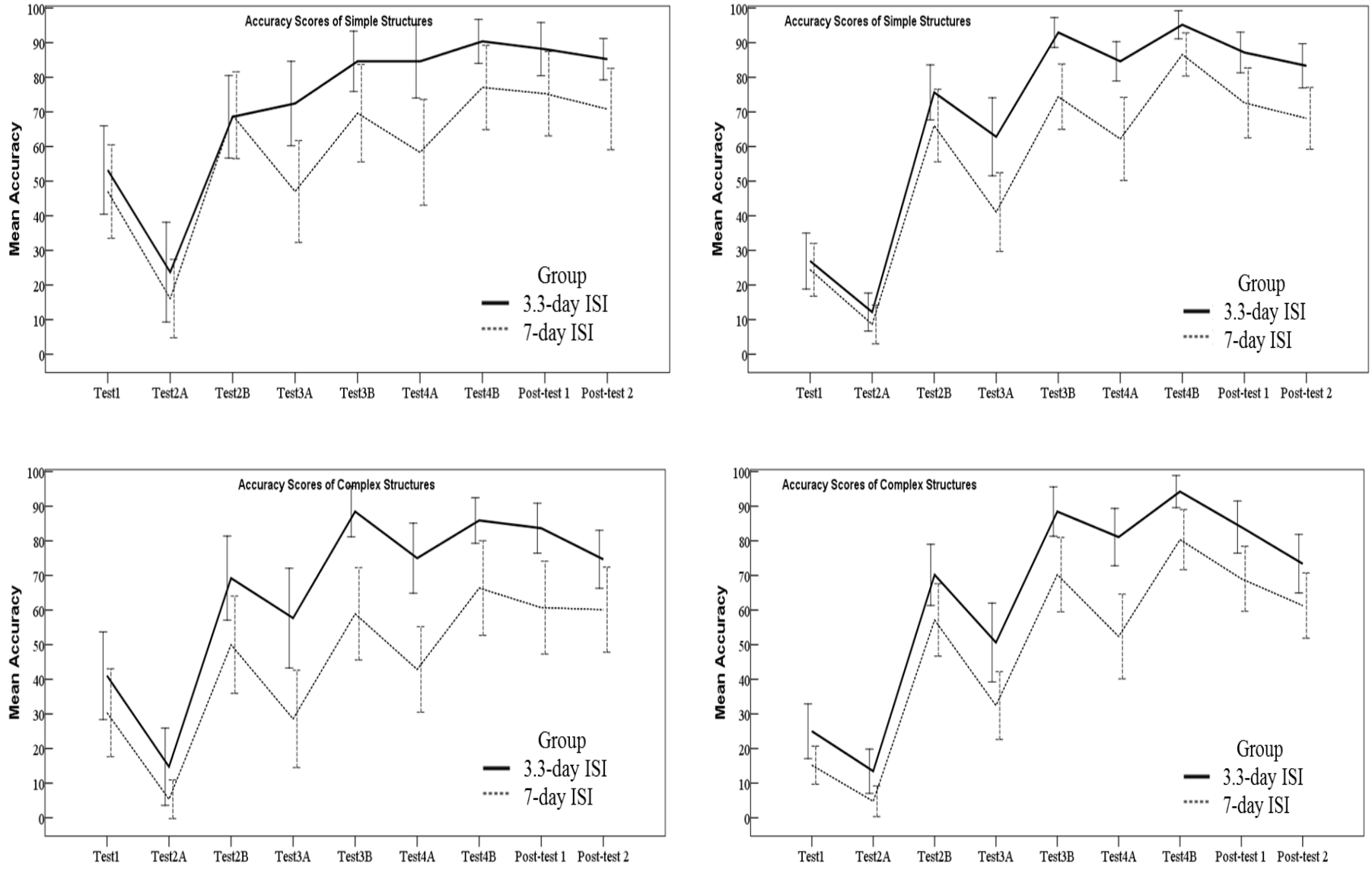


Figure 3 Accuracy scores and 95% confidence intervals for the two groups across time for simple structures (upper panel) and complex structures (lower panel) in the rule application test (left) and present progressive test (right).

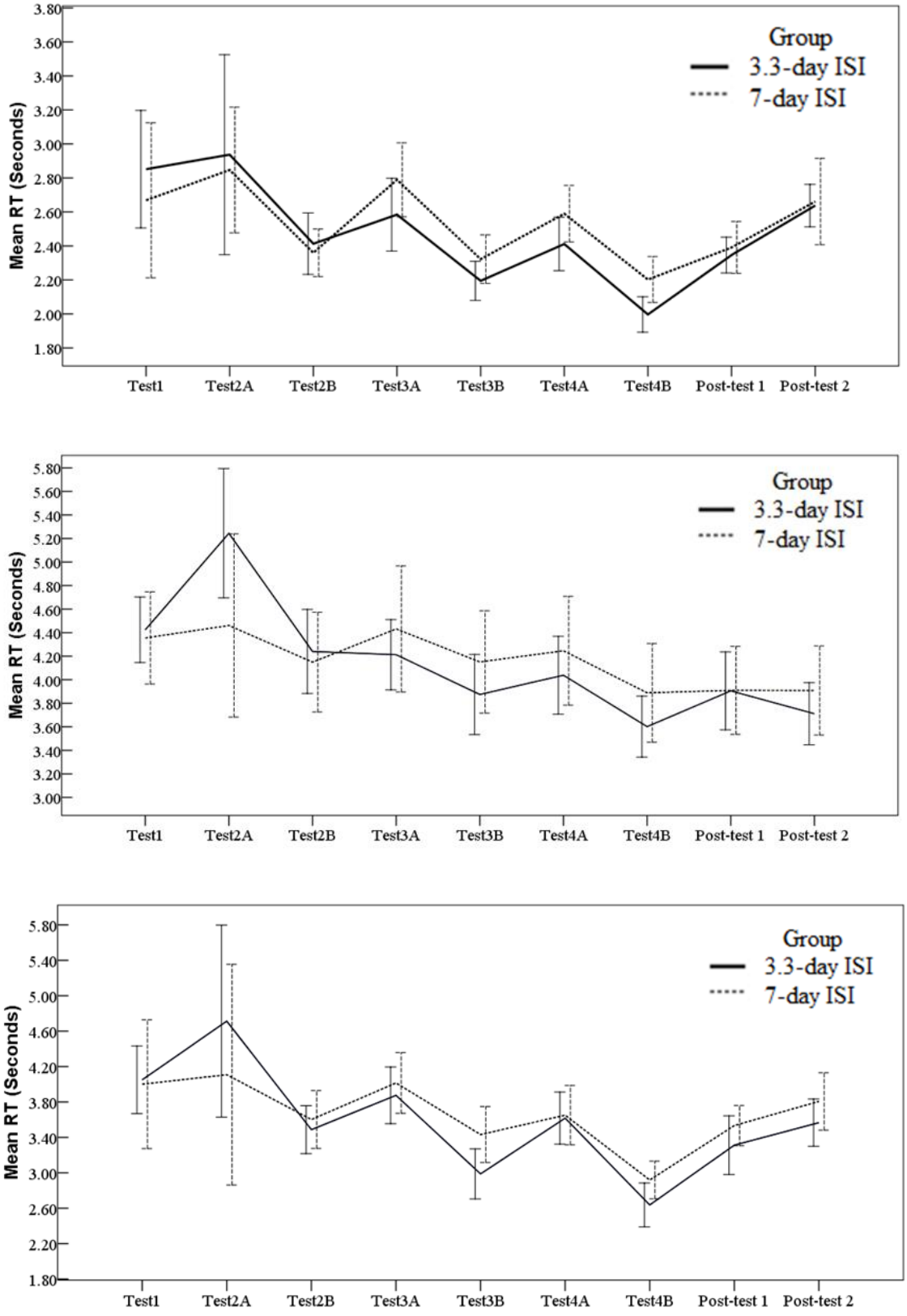
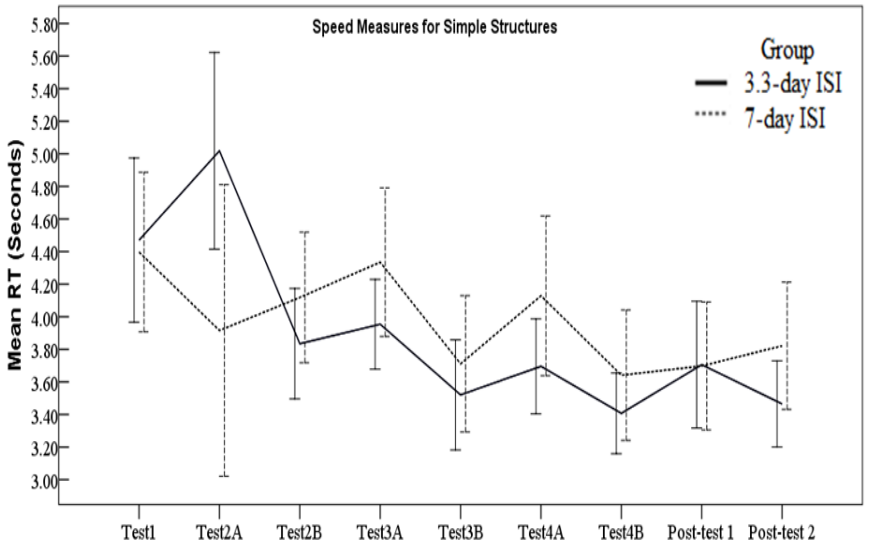
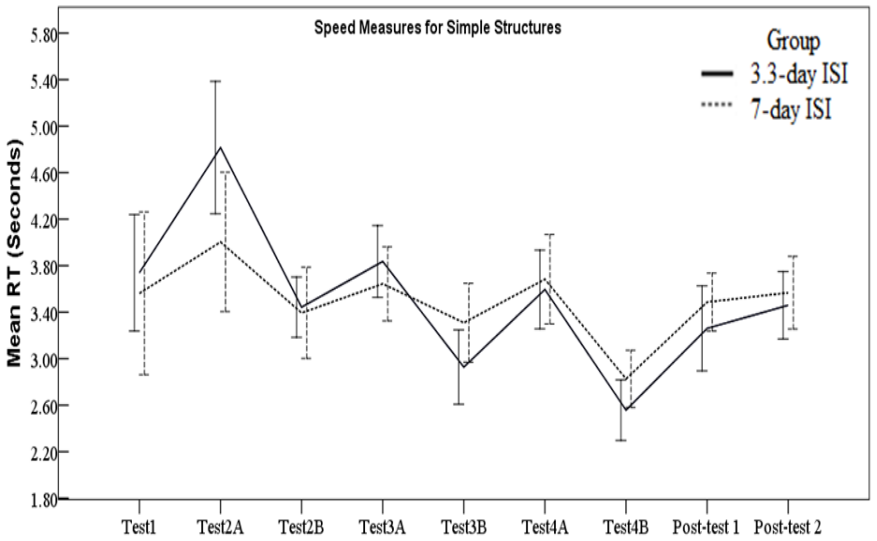
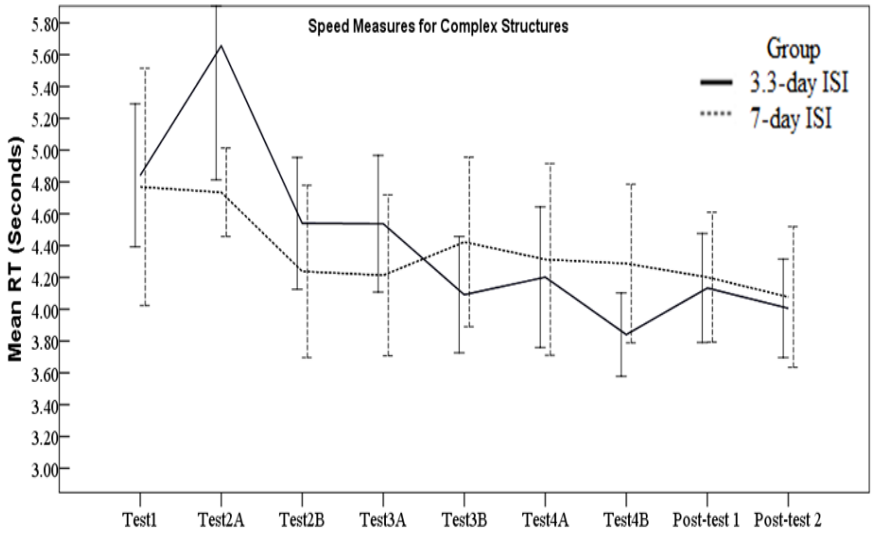


Figure 4 Speed measures and 95% confidence intervals for the two groups across time in the vocabulary test (top panel), rule application test (middle panel), and present progressive test (bottom panel).


Speed Measures for Simple Structures
Group
3.3-day ISI
7-day ISI
Mean RT (Seconds)
5.80
5.60
5.40
5.20
5.00
4.80
4.60
4.40
4.20
4.00
3.80
3.60
3.40
3.20
3.00
Test1
Test2A
Test2B
Test3A
Test3B
Test4A
Test4B
Post-test 1
Post-test 2


Speed Measures for Simple Structures
Group
3.3-day ISI
7-day ISI
Mean RT (Seconds)
5.80
5.40
5.00
4.60
4.20
3.80
3.40
3.00
2.60
2.20
1.80
Test1
Test2A
Test2B
Test3A
Test3B
Test4A
Test4B
Post-test 1
Post-test 2


Speed Measures for Complex Structures
Group
3.3-day ISI
7-day ISI
Mean RT (Seconds)
5.80
5.60
5.40
5.20
5.00
4.80
4.60
4.40
4.20
4.00
3.80
3.60
3.40
3.20
3.00
Test1
Test2A
Test2B
Test3A
Test3B
Test4A
Test4B
Post-test 1
Post-test 2

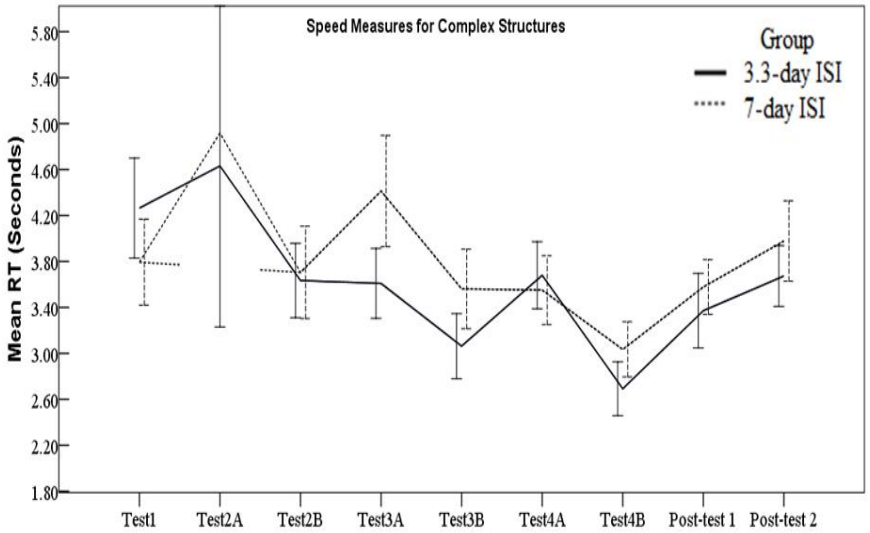

Speed Measures for Complex Structures
Group
3.3-day ISI
7-day ISI
Mean RT (Seconds)
5.80
5.40
5.00
4.60
4.20
3.80
3.40
3.00
2.60
2.20
1.80
Test1
Test2A
Test2B
Test3A
Test3B
Test4A
Test4B
Post-test 1
Post-test 2

Figure 5 Speed measures and 95% confidence intervals for the two groups across time for simple structures (upper panel) and complex structures (lower panel) in the rule application test (left) and present progressive test (right). The speed measure for Monitoring Test 2A in the distributed learning group was not computed due to insufficient number of valid (accurate) responses.