Mixing Grammar Exercises Facilitates Long-Term Retention: Effects of Blocking, Interleaving, and Increasing Practice

TATSUYA NAKATA[1] and YUICHI SUZUKI[2]
[1]*Hosei University, Faculty of Letters, Hosei University, 2-17-1, Fujimi, Chiyoda-ku, Tokyo 102-8160 Japan*
Email: t-nakata@hosei.ac.jp
[2]*Kanagawa University, Faculty of Foreign Languages, Kanagawa University, 3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa 221-8686 Japan*
Email: szky819@kanagawa-u.ac.jp

<ABSTRACT>

Cognitive psychology research has shown that interleaving, wherein learners practice multiple skills or concepts at once, facilitates learning more than does blocking, wherein learners practice only one skill or concept at a time. Despite the advantage of interleaving over blocking observed across a number of domains, limited attention has been devoted to the effects of interleaving on second language (L2) learning. This study examined the effects of blocking and interleaving on L2 grammar learning. In this study, 115 Japanese learners studied 5 English grammatical structures under 1 of 3 conditions: blocking, interleaving, and increasing (i.e., blocking followed by interleaving). Learning was measured using a grammaticality judgment test administered immediately and 1 week after the treatment. Although interleaving led to the highest number of incorrect responses during training, it was more effective than blocking in the 1-week delayed posttest. These results indicate that the advantage of interleaving extends to L2 grammar learning. Furthermore, learners' levels of prior knowledge were found to moderate the effects of interleaving. Specifically, participants with lower pretest scores benefited more from interleaving compared to those with higher pretest scores. Pedagogically, the findings suggest that grammar learning may be enhanced by incorporating interleaved practice.

Keywords: interleaving; blocking; increasing practice; grammar acquisition; contextual interference; practice distribution

Practice is a critical component in second language (L2) learning (DeKeyser, 2007; Ellis & Shintani, 2014; Suzuki, 2017). According to skill acquisition theory (DeKeyser, 2015), instructed L2 learners first acquire declarative knowledge (e.g., knowledge about grammatical rules), which allows them to develop procedural and automatic knowledge, and thus use the L2 more fluently and efficiently. In developing procedural and automatic knowledge, extensive and deliberate practice plays an important role (DeKeyser, 2007; Lyster & Sato, 2013). The importance of practice for L2 learning prompts the question of how we can optimize the effectiveness of L2 practice.

An extensive body of literature suggests that manipulating the practice schedule exerts a marked influence on learning (e.g., Bird, 2010; Muñoz, 2012; Nakata & Webb, 2016; Rogers, 2017; Suzuki, 2017; Suzuki & DeKeyser, 2017a). One of the most robust findings related to the practice schedule pertains to the spacing effect. According to the spacing effect, spaced learning, where practice opportunities are distributed over multiple occasions, yields superior retention when compared to massed learning, where practice opportunities take place successively without any intervals (e.g., Rogers, 2017). Research has shown that spaced learning often leads to superior retention relative to massed learning for a wide range of target features, including L2 grammar (Miles, 2014) and vocabulary (e.g., Nakata, 2015).[1]

Learning could also be facilitated by manipulating the practice schedule via interleaved practice instead of blocked practice (for a review, see Kang, 2016). In blocked practice, learners practice only one skill or concept at a time, whereas in interleaved practice, multiple skills or concepts are practiced at once. Research shows that learners typically produce more errors during learning in interleaved practice than in blocked practice because interleaving causes interference (e.g., Kang & Pashler, 2012; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010). In other words, when multiple skills or concepts are practiced simultaneously, learners tend to confuse one with another. This interfering effect that results from practicing multiple skills or concepts at once is referred to as contextual interference (e.g., Schneider, Healy, & Bourne, 1998, 2002). Yet, empirical evidence shows that, although interleaving results in lower performance than blocking during learning, it often increases long-term retention (e.g., Kang, 2016), a phenomenon known as the interleaving effect.

It should be noted here that, although spacing and interleaving are separate constructs, they are often confounded. Specifically, interleaved practice typically corresponds to spaced learning, whereas blocked practice corresponds to massed learning. For instance, if learners are asked to study three related concepts (A, B, and C) based on a blocked or interleaved schedule, these could be arranged as in Figure 1.

FIGURE 1
Example of Typical Blocked and Interleaved Schedules

Blocked schedule
A A A B B B C C C

Interleaved schedule
A B C A B C A B C

In the blocked schedule in Figure 1, each of the three concepts is studied sequentially. The treatment, therefore, does not involve any spacing between practice opportunities for a given

concept. In interleaved practice, in contrast, practice opportunities for a given concept (e.g., A) are separated by those for the other two concepts (e.g., B and C), which corresponds to spaced learning.

It is, however, possible to isolate the interleaving and spacing effects. One way to do so would be to use filler tasks (which are unrelated to the target concepts or skills) in the blocked schedule to introduce temporal spacing. An example of a blocked schedule with temporal spacing is given in Figure 2.

FIGURE 2
Example of a Blocked Schedule with Temporal Spacing

Blocked schedule (spaced)

A f f A f f A B f f B f f B C f f C f f C

Note. *f* denotes a filler task.

In the schedule given in Figure 2, practice opportunities for a given concept (e.g., A) are not separated by those for the other two (e.g., B or C). The treatment, therefore, is a type of blocked schedule. At the same time, because the practice opportunities for a given concept are separated by filler tasks, the treatment involves spacing. Although interleaving and spacing were confounded in most existing studies, Taylor and Rohrer (2010) and Kang and Pashler (2012) compared the effects of blocking and interleaving while controlling for the amount of spacing using filler tasks, as shown in Figure 2, for the learning of mathematics and painting styles, respectively. They found that interleaving resulted in better long-term retention than blocking, even though both had equivalent spacing. These findings suggest that the benefits of interleaving cannot be solely explained by the spacing effect.

Considering that benefits of interleaving have been observed across a number of domains, such as motor skill acquisition (learning to play sports or musical instruments; e.g., Porter et al., 2007), category learning (learning species of birds, Birnbaum et al., 2013; learning different categories of chemical compounds, Eglington & Kang, 2017; learning painting styles of different artists, Kang & Pashler, 2012), and mathematical problem solving (Rohrer & Taylor, 2007; Taylor & Rohrer, 2010), it would be valuable to investigate whether interleaving also facilitates L2 learning. In the present study, the effects of blocking and interleaving in the acquisition of L2 grammar were compared. Examining the effects of blocking and interleaving for grammar learning is useful because, in most traditional L2 textbooks, especially ones based on a structural syllabus, grammar is practiced in a blocked manner. For instance, Pan et al. (2018) analyzed 25 popular Spanish coursebooks to examine how preterite and imperfect past tenses, which are highly similar and confusable, are introduced. They found that, in 24 of these textbooks, the two grammatical structures are introduced in a blocked manner. That is, the two tenses are introduced separately, typically in different chapters. The interleaving effect, however, predicts that instead of presenting one of the two tenses at a time, it may be more beneficial to simultaneously introduce the students to both tenses. Given the potential benefits of interleaving practice for L2 grammar learning, the present study aims to examine whether interleaving facilitates the acquisition of L2 grammar. The findings of this study might allow us to identify the more effective practice schedule for grammar learning.

<A>LITERATURE REVIEW

<B>*Theoretical Background*

Several theories have been proposed to account for the benefits of interleaving over blocking practice. First, researchers argue that interleaving facilitates learning because it introduces spacing. As previously discussed, interleaving typically corresponds to spaced learning and blocking is equivalent to massed learning. The spacing effect (e.g., Rogers, 2017), therefore, predicts that interleaving facilitates learning more than does blocking. However, findings yielded by several studies have shown that interleaving resulted in better long-term retention than blocking, even though both had equivalent spacing (Kang & Pashler, 2012; Taylor & Rohrer, 2010), suggesting that the benefits of interleaving cannot be solely explained by the spacing effect.

Second, the Discriminative Contrast Hypothesis (Carpenter & Mueller, 2013; Kang & Pashler, 2012) might account for the benefits of interleaving, independently from the spacing effect. According to this hypothesis, interleaving helps learning because, by mixing exemplars from different categories, the differences between categories are made more apparent. For instance, suppose that learners are asked to answer grammar questions targeting the second and third conditionals in English. When questions for the second conditional are interleaved with those for the third conditional, this strategy may help learners identify features that distinguish between the two (e.g., the second conditional is used to refer to a hypothetical present situation, while the third conditional is used to refer to a hypothetical past situation; Ferguson, 2001). In contrast, when questions are presented in a blocked fashion, learners may notice commonalities underlying the second or third conditional but may not necessarily understand the distinction between the two conditionals.

Third, benefits of interleaving may also be explained by the transfer-appropriate processing (TAP) theory (Morris, Bransford, & Franks, 1977). According to this theory, performance is enhanced if the testing condition matches that employed while learning. In the posttests used in earlier studies, items targeting different categories were usually mixed together, just as in interleaved practice. The TAP theory, as a result, predicts that interleaving will lead to better posttest performance than blocking. Note that interleaved practice is also more representative of real-life situations. For instance, in actual language use, different grammatical structures are typically employed simultaneously and do not appear in any particular order. According to the TAP theory, hence, interleaving better prepares learners to function in real-life situations, where grammatical structures are not blocked by category.

<B>*Effects of Blocking and Interleaving in Cognitive Psychology and Motor Learning Research*

While only a few studies have been conducted to examine the effects of blocking and interleaving on L2 learning (e.g., grammar in Pan et al., 2018; vocabulary in Schneider et al., 1998, 2002; pronunciation in Carpenter & Mueller, 2013), an extensive body of research exists in the field of cognitive psychology and motor learning (Kang, 2016). Findings yielded by most non-L2 studies indicate that, although blocking leads to better performance during training, interleaving often results in better long-term retention (e.g., Kang & Pashler, 2012; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010). Yet, some researchers are of the view that the effects of interleaving may be moderated by several factors, such as the degree of between-category discriminability of the target features and learners' levels of prior knowledge, suggesting that the advantage of interleaving may disappear or even be reversed in certain conditions.

One factor that is found to influence the effects of interleaving is the degree of between-

category discriminability (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2013). Research suggests that, while blocking helps learners notice the commonalities within each category, interleaving is especially helpful for distinguishing among different categories (e.g., Kang, 2016). This distinction suggests that the effects of interleaving may be moderated by the nature of target features. For instance, suppose that categories to be learned are similar to each other, making it difficult to identify features that separate one category from others (i.e., between-category discriminability is low). In this case, being able to distinguish among similar categories is critical. As a result, interleaving may be particularly beneficial for target features with low between-category discriminability. In contrast, if categories to be learned are very different from each other (i.e., between-category discriminability is high), learners may have little difficulty distinguishing among different categories even in a blocked schedule. The advantage of interleaving over blocking may therefore be limited for target features characterized by high between-category discriminability.

Another factor that is found to influence the effects of interleaving is learners' prior knowledge. Several studies have shown that interleaving tends to be effective for experienced learners, while blocking is sometimes beneficial for novices (e.g., Guadagnoli, Holcomb, & Weber, 1999; Rey, Wughalter, & Whitehurst, 1982). Porter and Magill (2010) argued that these results could be in part explained by the desirable difficulty framework (Bjork, 1999), according to which challenging learners at the appropriate level of difficulty facilitates long-term retention. Because interleaving causes more contextual interference than blocking, interleaved practice may be too difficult and overwhelming for novice learners, which results in inefficient learning. Interleaving, in contrast, helps introduce the appropriate level of difficulty to experienced learners, which may facilitate learning by making learning desirably difficult.

<B>*Effects of a Combination of Blocking and Interleaving*

Although in most extant studies the researchers have examined the effects of either blocking or interleaving, a small number of non-L2 studies focused on the effects of a combination of the two (Porter et al., 2007; Porter & Magill, 2010; Wong et al., 2013). These studies have indicated that using both blocking and interleaving is more effective than relying on either method alone. Combining blocking and interleaving is considered beneficial because these strategies facilitate different processes. While blocking may be especially beneficial for learning the commonalities within each category, interleaving may help learners to effectively distinguish among different categories (e.g., Kang, 2016). Therefore, combining blocking and interleaving may facilitate more efficacious learning because it offers benefits of both strategies.

This could be achieved by presenting blocking first followed by interleaving mode, or vice versa. Porter and colleagues argued that blocking followed by interleaving facilitates learning more than when their order is reversed (Porter et al., 2007; Porter & Magill, 2010). Other researchers explained this effect by suggesting that, although interleaving tends to be effective for experienced learners, blocking is more beneficial for novices (e.g., Guadagnoli et al., 1999; Rey et al., 1982). As training progresses, learners' proficiency is expected to improve, indicating that blocking should be used in the early stages of learning, when the proficiency level is low. In the later stages, in contrast, interleaving is more advantageous because the proficiency level of the learner increases as a function of prior practice. This kind of practice—blocking followed by interleaving—is referred to as *increasing practice* because contextual interference is gradually increased. By ensuring a continuous match between learner's proficiency level and task difficulty, increasing practice introduces the appropriate level of difficulty throughout the

treatment (Porter et al., 2007; Porter & Magill, 2010), which enhances retention, according to the desirable difficulty framework (Bjork, 1999).

<*B*>Effects of Blocking and Interleaving on L2 Learning

Effects of blocking and interleaving have also been investigated in the context of learning L2 pronunciation (Carpenter & Mueller, 2013), vocabulary (Finkbeiner & Nicol, 2003; Schneider et al., 1998, 2002), and grammar (Pan et al., 2018). Carpenter and Mueller (2013), for instance, conducted a study as a part of which English-speaking college students learned pronunciation rules for French (e.g., *eau* is pronounced as a long *o* sound as in *cadeau* 'gift' or *tableau* 'blackboard'). In the blocked condition, participants were exposed to, both visually and orally, words conforming to the same pronunciation rule in a sequence (e.g., *bat<u>eau</u>, fard<u>eau</u>, ram<u>eau</u>, . . . tandi<u>s</u>, brebi<u>s</u>, verni<u>s</u>, . . . dard<u>er</u>, compl<u>er</u>, vals<u>er</u>*, etc.). In the interleaved condition, the words to which different rules applied were presented in a random fashion (e.g., *bat<u>eau</u>, tandi<u>s</u>, dard<u>er</u>, fard<u>eau</u>, brebi<u>s</u>, compl<u>er</u>, ram<u>eau</u>, verni<u>s</u>, vals<u>er</u>*). Contrary to earlier non-L2 studies finding benefits of interleaving, Carpenter and Mueller's results indicate that blocking leads to better retention than does interleaving. These findings may be in part explained by two moderating factors discussed in the previous section. First, because the target pronunciation rules were markedly different from each other (e.g., *eau, s, er*), the between-category discriminability was perhaps high. This might have attenuated the benefits of interleaving. Second, participants in their study did not have any prior knowledge of the target pronunciation rules in French. Because blocking tends to be effective for novices (e.g., Guadagnoli et al., 1999; Rey et al., 1982), participants might have benefitted more from blocking than from interleaving.

Three studies have compared the effects of blocking and interleaving on L2 vocabulary learning (Finkbeiner & Nicol, 2003; Schneider et al., 1998, 2002). In the study conducted by Finkbeiner and Nicol, 24 American university students learned 32 pseudowords from four semantic categories (animals, kitchen utensils, furniture, and body parts) under either a blocked or mixed condition. In the blocked condition, eight items from the same semantic category were studied in a row, whereas in the mixed condition, target items were interleaved with items from other categories. Finkbeiner and Nicol found that, on a translation task conducted immediately after learning, participants' response speed was faster for items studied in the mixed condition than for items studied in the blocked condition, suggesting the advantage of interleaving for L2 vocabulary learning. In contrast, Schneider et al. (1998, 2002) failed to show any advantage of interleaving over blocking. In their 1998 study, English-speaking college students studied 25 French words under blocked and mixed conditions. The target items belonged to five semantic categories (body parts, vehicles, kitchen utensils, food, and clothes). Although the blocked condition resulted in a greater percentage of correct translations than the interleaved condition during the initial learning phase, no significant difference was found on the immediate or delayed posttest. In their 2002 study, Schneider and colleagues showed that the blocked condition resulted in higher scores than the mixed condition on an immediate posttest. No significant difference, however, was found on a 1-week delayed posttest.

While L2 vocabulary studies have yielded inconsistent results regarding interleaving effects, the effects of interleaved and blocked practice may be more relevant for grammar learning than for vocabulary learning, as they involve fundamentally different processes. Specifically, grammar learning necessitates acquisition of an underlying linguistic pattern within each category (e.g., sentences using the first conditional are formed using the following structure: *if* + subject + simple present, subject + auxiliary verb + infinitive; Ferguson, 2001). In

contrast, a form of vocabulary learning investigated in the previous studies involved associating L2 words with their first language (L1) translations, which did not necessitate knowledge of any underlying pattern within each category (there was no rule, for example, that words referring to animals must have a particular ending). As a result, the context of grammar learning may provide a richer ground for examining the theoretical underpinnings of the advantage of blocking (i.e., blocking helps to find the commonalities within each category) and interleaving (i.e., interleaving aids in distinguishing among similar categories) than vocabulary learning.

Pan et al. (2018) were the first to examine the effects of interleaving on L2 grammar learning. As a part of their study, the researchers conducted four experiments, in which English-speaking college students were introduced to two grammar rules in Spanish (preterite and imperfect past tenses) under blocked or interleaved conditions. In Experiments 1 and 2, the treatment was conducted in a single session, whereas in Experiments 3 and 4, it was conducted over two sessions that were spaced 1 week apart. Although Pan and colleagues did not find any significant difference between blocked and interleaved practice in their first two experiments, interleaved practice led to higher posttest scores than blocked practice in Experiments 3 and 4. The researchers attributed these findings to the larger spacing in the interleaved condition relative to the blocked condition. Specifically, in Experiments 3 and 4, whereas study opportunities for a given target structure were massed into one treatment session in the blocked condition, they were distributed across two weekly sessions in the interleaved condition.

The findings reported by Pan et al. (2018) are valuable because their study is the first to demonstrate the benefits of interleaving for L2 grammar learning. One limitation of their research design, however, is that the participants did not have any prior knowledge of the target language (Spanish). Thus, the obtained results may not necessarily be applicable to L2 learners that are not complete novices. In the present study, the effects of blocking and interleaving on the acquisition of L2 grammar knowledge were also compared. However, to address the aforementioned limitation, participants had prior exposure to the target structures. This allowed us to examine whether the effects of blocking and interleaving are moderated by the learners' prior knowledge.

<B>*Research Questions and Hypotheses*

The current study addresses the following three research questions:

RQ1. Does interleaving facilitate L2 grammar learning more than blocking?

RQ2. Does increasing practice (blocking followed by interleaving) facilitate L2 grammar learning more than blocking or interleaving alone?

RQ3. Does the prior knowledge learners possess moderate the effects of blocking, interleaving, and increasing practice on L2 grammar learning?

The following three hypotheses related to these research questions were formed:

H1. Interleaving facilitates L2 grammar learning more than blocking.

H2. Increasing practice facilitates L2 grammar learning more than blocking or interleaving alone.

H3. Blocking is effective for learners with a low level of prior knowledge, whereas interleaving is effective for learners with a high level of prior knowledge.

The first hypothesis predicts that interleaving facilitates L2 grammar learning more than does blocking. This hypothesis is based on the findings yielded by Experiments 3 and 4 performed by Pan et al. (2018), as well as results obtained in other extant studies conducted in

the fields of motor skill acquisition, category learning, and mathematical operations.

The second hypothesis predicts the advantage of increasing practice over blocking or interleaving for two reasons. First, combining blocking and interleaving may be more effective than either strategy alone because it offers benefits of both blocking (i.e., it helps learners detect the commonalities within each category) and interleaving (i.e., it helps distinguish among different categories). Second, with increasing practice, difficulty is gradually increased by using blocking first, followed by interleaving. This practice order helps introduce the appropriate level of difficulty throughout the treatment, which facilitates learning (Bjork, 1999; Porter et al., 2007; Porter & Magill, 2010).

The third hypothesis predicts that the effects of blocking and interleaving are moderated by the learners' prior knowledge. Several non-L2 studies indicate that, although interleaving tends to be effective for experienced learners, blocking is sometimes beneficial for novices (e.g., Guadagnoli et al., 1999; Rey et al., 1982). This hypothesis predicts that a similar pattern of results may be observed for L2 grammar learning. No prediction was made for increasing practice because the role of prior knowledge in increasing practice has not been previously examined.

<A>METHOD

<B>*Participants*

The original pool of participants consisted of 151 Japanese students (aged 18−22) from two universities in Japan, each having at least 6 years of experience of studying English. Based on their scores on a standardized test (Test of English for International Communication [TOEIC] Bridge), their English proficiency was estimated to fall between the A2 (elementary) and B1 (intermediate) levels in the Common European Framework of Reference for Languages (CEFR) benchmark. The participants' L2 English proficiency was assessed by the junior Minimal English Test (jMET; Goto, Maki, & Kasai, 2010). The jMET is a dictation test, and the test scores are found to correlate highly with general L2 English proficiency as measured by scores on the reading and listening sections of Japanese university entrance exams. The average score on the jMET was 39.76 (SD = 8.63). The participants were randomly divided into three groups, denoted as blocked, increasing, and interleaved, to ensure that there would be no statistically significant differences in the jMET scores. Data from 32 participants who indicated in the post-study questionnaire that they had studied the target grammatical structures outside the experiment during the period between the immediate and delayed posttests were excluded from analysis (10 from the blocked, 11 from the increasing, and 11 from the interleaved group). Four additional participants who did not complete the questionnaire were also removed. The remaining 115 participants consisted of 39, 40, and 36 students from the blocked, increasing, and interleaved groups, respectively. According to the findings yielded by a one-way ANOVA, there were no statistically significant differences in the jMET scores among the remaining participants forming the three groups, F (2,112) = .538, p = .585, η^2 = .010. Nonetheless, because absence of statistical significance does not guarantee that the proficiency scores are equivalent among the three groups, the jMET scores were used as a covariate in the data analysis (see *Results*).

<B>*Target Structures*

Two sets of grammatical structures from the English tense–aspect–mood system were used as target structures in this study: (a) simple past and present perfect, and (b) the first, second, and third conditional. These grammatical structures were chosen because they are similar

to each other within each set. The simple past tense is used to refer to an event that happened at a particular point in the past (e.g., *My father came to my school last week*). The present perfect is used when there is a link between the past and present. For instance, it is used to refer to an event that started in the past and continues in the present (e.g., *My parents have lived in this house since 1976*). The simple past and present perfect are similar, in that both structures are used to refer to an event that took place in the past, and learners often confuse the two structures (Bird, 2010; e.g., **My father has come to my school last week* or **My parents lived in this house since 1976*), which might make them conducive to interleaving effects.

The first conditional is used to express an action that might happen in the future (e.g., *If it rains tomorrow, I'll stay at home*), and is formed using the following structure: conjunction (e.g., *if*) + subject + simple present, subject + auxiliary verb (e.g., *will*) + infinitive (Ferguson, 2001). The second conditional is used to refer to a hypothetical present situation (e.g., *If I lived in New York, I would go to musicals every day*), and is formed using the following structure: *if* + subject + simple past, subject + past form of auxiliary verb (e.g., *would*) + infinitive. Finally, the third conditional is used to refer to a hypothetical past situation (e.g., *If they had missed the train, they would have called*), and is formed using the following structure: *if* + subject + past perfect, subject + past form of auxiliary verb (e.g., *would*) + *have* + past participle. These three types of conditionals are similar in that each involves the conditional aspect. Consequently, they might interfere with each other, making them amenable to interleaving effects (e.g., **If I live in New York, I would go to musicals every day* or **If they missed the train, they would have called*).

The five target structures examined in the present study are not only similar to each other within each set (a and b) but also across sets. Specifically, the simple past and second conditional are similar in that both involve the simple past tense. The present perfect and third conditional involve the perfect aspect, and the simple past and third conditional involve pastness. Due to these similarities among the five structures, it was expected that students would have difficulty distinguishing them, making them conducive to interleaving effects. Note that because all five target structures are typically taught at Japanese secondary schools, all study participants had some exposure to the target structures prior to the experiment. These five structures, however, were chosen because most students have difficulty using them correctly.

<B>*Instruments*

The study was conducted in a computer laboratory. Each student had access to a computer, and the entire experiment was conducted using computer software developed by one of the authors.

<C>*Training Materials.*

During the treatment, the participants practiced five target grammatical structures in a multiple-choice format (see Figure 3). They were presented with a sentence where a verb or verb phrase was replaced with a blank (e.g., "I _____ a car for my daughter last Christmas") together with four options (e.g., "will buy," "have bought," "buy," "bought"). Participants were instructed to choose the most appropriate verb or verb phrase to complete the sentence. To make the intended meaning clear, the L1 (Japanese) translation of each sentence was also provided. The training materials consisted of 50 multiple-choice questions. There were 10 questions from each of the following five categories: simple past, present perfect, first conditional, second conditional, and third conditional. The three incorrect options (distractors) for each multiple-choice question were created so that, for questions in a given category, the same three verb

tenses were always used as incorrect options. For instance, for the questions targeting the simple past tense (e.g., *bought*), three incorrect options were always the simple present (e.g., "buy"), present perfect (e.g., "have bought"), and future tense (e.g., "will buy") of the target verb. To reflect real-life study situations, the treatment was self-paced and participants were able to spend as much time as they needed on each question. Since the study time was not controlled, it was treated as a covariate in the data analysis (see *Results*). After each response, the correct answer and metalinguistic explanation of the target structure were provided as feedback for 12 seconds (Figure 4; also see Online Supplement A for details).

Note that the treatment used in this study (multiple-choice fill-in-the-blank questions) is categorized as a type of "instruction that expects learners to focus on forms in isolation" (Norris & Ortega, 2000, p. 420). This type of controlled practice alone is insufficient for L2 acquisition (Ellis & Shintani, 2014). Nonetheless, it was chosen for the present study for three reasons. First, although controlled practice is not sufficient in isolation, it can be an efficient technique for acquiring explicit, declarative knowledge about target structures. This kind of knowledge can serve as a basis for developing procedural and automatic knowledge (DeKeyser, 2015), which allows learners to use the L2 more fluently and efficiently. Second, controlled practice including fill-in-the-blank questions is still used widely in L2 English textbooks (Nitta & Gardner, 2005) and in many English-as-a-foreign-language contexts. Third, using controlled practice enables experimenters to have strict control over the treatment and manipulate the practice schedule relatively easily.

FIGURE 3
Example of Multiple-Choice Questions During the Treatment

Note. The direction reads, "Choose the most appropriate word or phrase from the four options to complete the sentence." The Japanese translation for the sentence *I bought a car for my daughter last Christmas* is given above the English sentence. The correct answer is "bought."

FIGURE 4
Example of Feedback Given After a Learner's Response

Note. The metalinguistic explanation reads, "The past tense is used to refer to an event that happened at a particular point in the past."

Although all three groups answered the same 50 multiple-choice questions during the treatment, the item order was different in each case. In the blocked group, the 50 questions were blocked by category. Specifically, for approximately half of the participants in this group (18 out of 39), the 50 questions were arranged as follows: 10 simple past → 10 present perfect → 10 first conditional → 10 second conditional → 10 third conditional (see Figure 5, top). For the remaining participants (21 out of 39), the 50 questions were arranged as follows: 10 first conditional → 10 second conditional → 10 third conditional → 10 simple past → 10 present perfect. These two item orders were used to minimize the potential order effects. In both cases, questions on the simple past were immediately followed by those on the present perfect (10 simple past → 10 present perfect). This was done to ensure that questions related to the simple past and present perfect, which are similar and somewhat confusable, would not be interleaved by questions pertaining to the conditionals. For the same reason, questions on the three types of conditionals were presented in sequence (10 first conditional → 10 second conditional → 10 third conditional) in both cases.

In the interleaved group, questions from five categories were mixed, such as simple past, present perfect, first conditional, second conditional, third conditional, second conditional, simple past, first conditional, present perfect, third conditional, etc. Questions from the same grammatical category never appeared twice in a row (see Figure 5, middle). In the increasing group, the first 25 questions were blocked by category as in the blocked group. For the remaining 25 questions, questions related to the aforementioned five grammatical categories were mixed as in the interleaved group (see Figure 5, bottom).

FIGURE 5
Sample Item Orders in the Blocked, Interleaved, and Increasing Conditions

Blocked condition

1. Simple past	2. Simple past	3. Simple past	4. Simple past	5. Simple past
6. Simple past	7. Simple past	8. Simple past	9. Simple past	10. Simple past
11. Present perfect	12. Present perfect	13. Present perfect	14. Present perfect	15. Present perfect
16. Present perfect	17. Present perfect	18. Present perfect	19. Present perfect	20. Present perfect
21. 1st conditional	22. 1st conditional	23. 1st conditional	24. 1st conditional	25. 1st conditional
26. 1st conditional	27. 1st conditional	28. 1st conditional	29. 1st conditional	30. 1st conditional
31. 2nd conditional	32. 2nd conditional	33. 2nd conditional	34. 2nd conditional	35. 2nd conditional
36. 2nd conditional	37. 2nd conditional	38. 2nd conditional	39. 2nd conditional	40. 2nd conditional
41. 3rd conditional	42. 3rd conditional	43. 3rd conditional	44. 3rd conditional	45. 3rd conditional
46. 3rd conditional	47. 3rd conditional	48. 3rd conditional	49. 3rd conditional	50. 3rd conditional

Interleaved condition

1. Simple past	2. Present perfect	3. 1st conditional	4. 2nd conditional	5. 3rd conditional
6. 2nd conditional	7. Simple past	8. 1st conditional	9. Present perfect	10. 3rd conditional
11. Simple past	12. 2nd conditional	13. 1st conditional	14. Present perfect	15. 3rd conditional
16. Simple past	17. Present perfect	18. 3rd conditional	19. 2nd conditional	20. 1st conditional
21. Present perfect	22. Simple past	23. 3rd conditional	24. 2nd conditional	25. 1st conditional
26. 3rd conditional	27. 2nd conditional	28. Simple past	29. 1st conditional	30. Present perfect
31. Simple past	32. 1st conditional	33. 2nd conditional	34. Present perfect	35. 3rd conditional
36. Present perfect	37. 3rd conditional	38. Simple past	39. 2nd conditional	40. 1st conditional
41. Simple past	42. 1st conditional	43. Present perfect	44. 3rd conditional	45. 2nd conditional
46. Simple past	47. Present perfect	48. 1st conditional	49. 2nd conditional	50. 3rd conditional

Increasing condition

1. Simple past	2. Simple past	3. Simple past	4. Simple past	5. Simple past
6. Present perfect	7. Present perfect	8. Present perfect	9. Present perfect	10. Present perfect
11. 1st conditional	12. 1st conditional	13. 1st conditional	14. 1st conditional	15. 1st conditional
16. 2nd conditional	17. 2nd conditional	18. 2nd conditional	19. 2nd conditional	20. 2nd conditional
21. 3rd conditional	22. 3rd conditional	23. 3rd conditional	24. 3rd conditional	25. 3rd conditional
26. 3rd conditional	27. 2nd conditional	28. Simple past	29. 1st conditional	30. Present perfect
31. Simple past	32. 1st conditional	33. 2nd conditional	34. Present perfect	35. 3rd conditional
36. Present perfect	37. 3rd conditional	38. Simple past	39. 2nd conditional	40. 1st conditional
41. Simple past	42. 1st conditional	43. Present perfect	44. 3rd conditional	45. 2nd conditional
46. Simple past	47. Present perfect	48. 1st conditional	49. 2nd conditional	50. 3rd conditional

<C>*Pretest and Posttest.*

The grammaticality judgment test was administered as the pretest and posttest (immediate and 1-week delayed). The grammaticality judgment test was chosen as a dependent measure because it is a commonly used tool for assessing L2 linguistic knowledge (Gass, 2018). Participants were presented with 40 sentences sequentially and were instructed to press the left arrow on the keyboard for a grammatical sentence and the right arrow for an ungrammatical sentence. The test targeted declarative–explicit knowledge of grammatical features developed via controlled practice in rule identification and form comparison. Although the grammaticality judgment test was not timed, the participants were instructed to respond as quickly as possible. The 40 items consisted of 8 items from each of the five target structures. Half of the items were grammatical (e.g., "My father came to my school last week"), and the other half contained errors related to the verb tense and were ungrammatical (e.g., *"Mary has bought a nice bag three days ago"). To ensure that participants would not judge the sentence as grammatically incorrect for reasons other than the verb tense, they were informed that no sentences would contain errors unrelated to the verb tense. At the beginning of the test, four practice items were given to familiarize the participants with the test format. The four practice items did not involve any of the target structures (e.g., "He plays basketball well" or *"This is mine pen") to eliminate the potential influence on student performance on the critical items.

To reduce a potential practice effect, three forms of the test that differed in terms of the noun phrase and adverb phrase were used, as illustrated by the following examples:
EXAMPLE
Form A: *Daniel built a house last September.*
Form B: *Paul built a house last October.*
Form C: *John built a house last summer*.

The test items appeared in a different randomized order in each form. To control for the test form effects, the administration of the three forms was counterbalanced across participants. Specifically, a subgroup of participants (20 out of 115) had Form A for the pretest, Form B for the immediate posttest, and Form C for the delayed posttest, while another subgroup (18 out of 115) had Form C for the pretest, Form A for the immediate posttest, and Form B for the delayed posttest, etc.

The reliability of the grammaticality judgment test indexed by Cronbach alpha was .688 (pretest), .851 (immediate), and .817 (delayed). As the pretest reliability is slightly lower than .70, it requires some caution in interpreting the results, whereas the two posttests showed moderate reliability (Brown, 2014).

<B>*Procedure*

The study was conducted during two regular classes. The participants received explanations about the study before taking part in the pretest. In the pretest, participants were presented with 40 English sentences one by one and were asked to judge whether each sentence was grammatically correct or not. The pretest was followed by the treatment, where the participants answered 50 multiple-choice questions and studied five target grammatical structures. After the treatment, the participants were required to answer 10 two-digit additions (e.g., 17 + 75 = ?) as a filler task. This task was included to minimize the potential order effects. Following the filler task, the participants took the grammaticality judgment test as the immediate posttest. After the immediate posttest, participants were asked to evaluate the effectiveness of the treatment for learning the target structures on a 7-point Likert scale, where 1 meant *not effective at all* and 7 corresponded to *very effective*. Although the participants were not specifically asked to focus on item distribution, as only learning schedules differed across the three groups, they would be the source of any potential differences among these groups. One week after the immediate posttest, an unannounced delayed posttest was given.

<B>*Analysis*

Responses to the grammaticality judgment tests were analyzed in terms of accuracy rate (proportion of correct responses) and *d*-prime score. *D*-prime scores provide an unbiased index of grammatical sensitivity, namely the ability to distinguish grammatical from ungrammatical items. They are considered a more sensitive index than accuracy rates, because they take response bias into account (Macmillan & Creelman, 2004). Specifically, *d*-prime scores are calculated by subtracting the *z* score of false alarm rate (e.g., responding “yes” to ungrammatical items) from the *z* score related to hit rate (responding “yes” to grammatical items). For instance, the grammaticality judgment test used in this study consisted of 20 grammatical items and 20 ungrammatical items. Suppose Person A answered 16 grammatical items correctly and 16 ungrammatical items correctly, while Person B answered 14 grammatical items correctly and 18 ungrammatical items correctly. Both individuals successfully answered 32 items out of 40, and their accuracy rate is 80%. However, Person A rated more ungrammatical items as correct (4) than did Person B (2). The *d*-prime score takes this type of response bias into account by penalizing participants who have a high false alarm rate. As a result, the *d*-prime score will be higher for Person B (1.81) than Person A (1.68).

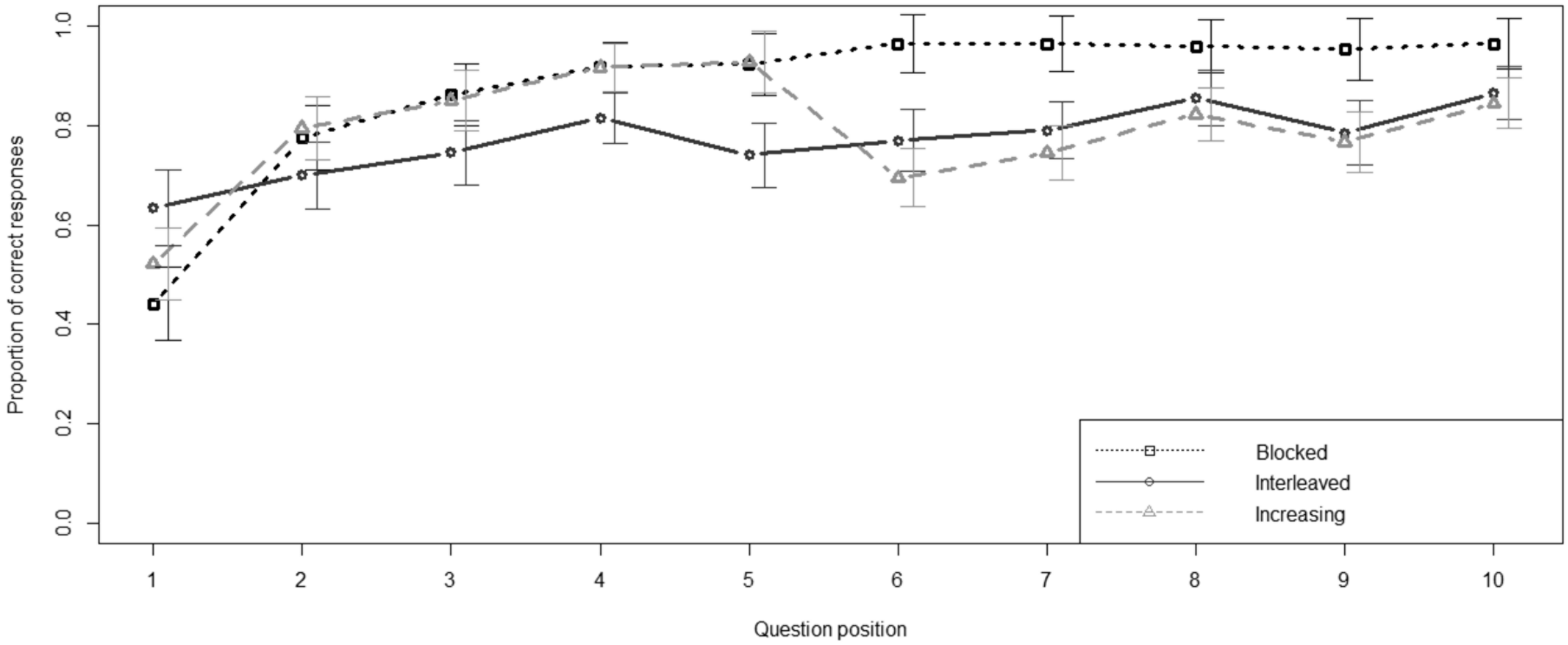
<A>RESULTS

<B>*Performance During Treatment*

During the treatment, the participants answered 50 multiple-choice questions, comprising 10 questions for each of the five target grammatical structures. Figure 6 illustrates the proportion of correct responses as a function of question position during the treatment when collapsed across the five grammatical structures. For instance, Figure 6 shows that, in the blocked group, the average accuracy rate was 44.10% for the first question in each grammatical structure. It also shows different learning curves for the three groups. First, in the blocked group, although the accuracy rate initially increased as a function of question position, there were no measurable gains in the latter half of the treatment. Second, in the interleaved group, the accuracy rate continued to increase toward the end of the treatment, in contrast to the blocked group. Third, for the first half of the questions, the learning curve for the increasing group was similar to that of the blocked group. This finding can be attributed to the identical practice schedule for the first five questions in each grammatical structure for the blocked and increasing groups (see Figure 5). The accuracy rate of the increasing group, however, decreased when responding to the latter half of the questions, where the questions were interleaved rather than blocked (see Online Supplement 2 for detailed statistical analyses).

FIGURE 6

Proportion of Correct Responses During the Treatment as a Function of Question Position



Note. Error bars indicate 95% confidence intervals.

When collapsed across all 50 questions, the average accuracy rate was 87.23% (95% CI [84.33%, 90.13%], *SD* = 8.94%) for the blocked group, 78.83% (95% CI [74.76%, 82.90%], *SD* = 12.03%) for the increasing group, and 77.00% (95% CI [72.21%, 81.79%], *SD* = 14.98%) for the interleaved group. Results yielded by a one-way ANOVA revealed significant differences among the three groups, F (2,114) = 7.72, p = .001, η^2 = .121. The Bonferroni method of multiple comparisons showed that the blocked practice led to significantly better performance than increasing (p = .010, d = 0.80) and interleaved practice (p = .001, d = 0.83). The effect sizes (d) are considered medium according to the guidelines (small: d = 0.4; medium: d = 0.7; large: d = 1.0) proposed by Plonsky and Oswald (2014). However, no significant difference was found between the increasing and interleaved groups (p = .792), and only a very small effect size was noted (d = 0.13).

Since participants were allowed to spend as much time as they needed on answering the multiple-choice questions, the treatment duration was different for each participant. The average treatment duration (with standard deviations given in parentheses) was 17.56 (3.84), 18.16 (1.98), and 20.26 (4.48) minutes for the blocked, increasing, and interleaved groups, respectively. Due to the statistically significant difference in the treatment duration among the three groups, F(2,112) = 6.01, p = .003, treatment duration was modelled as a covariate in the subsequent analyses.

<B>*Performance on the Pretest and Posttests*

Table 1 presents accuracy rates and *d*-prime scores of the grammaticality judgment tests. According to the Shapiro-Wilk test, all *d*-prime scores were normally distributed (ps > .05). A mixed ANCOVA was conducted on the *d*-prime scores with condition (blocked, increasing, and interleaved) as a between-participant variable and time (immediate and delayed posttest) as a within-participant variable. The covariates were *d*-prime pretest scores, treatment duration (log transformed, in order to reduce skewness), and jMET scores. The results of the mixed ANCOVA revealed no significant main effect of condition (see Table 2). However, a marginally significant interaction between time and condition was found, indicating that the effects of condition varied depending on the timing of posttests (immediate or delayed). Furthermore, the three-way interaction among time, condition, and pretest was also marginally significant, suggesting that the pretest scores may further moderate the interaction between time and condition.[2]

TABLE 1
Accuracy and D-Prime Scores on the Grammaticality Judgment Tests

	Pretest		Immediate Posttest		Delayed Posttest	
	M	*SD*	*M*	*SD*	*M*	*SD*
Accuracy (%)						
Blocked	70.51	11.21	80.51	13.51	78.53	13.00
Increasing	70.90	12.72	83.54	12.81	80.14	14.64
Interleaved	67.88	13.06	80.00	17.02	79.63	14.75
D-prime						
Blocked	1.17	0.72	1.87	0.97	1.76	0.97
Increasing	1.16	0.77	2.18	1.04	1.92	1.09
Interleaved	1.04	0.84	1.95	1.21	1.89	1.09
Adjusted *d*-prime						
Blocked			1.73	0.80	1.60	0.72
Increasing			2.12	0.77	1.86	0.70
Interleaved			2.14	0.80	2.07	0.73

Note. *M* = mean; *SD* = standard deviation. *n* = 39, 36, and 40 for the blocked, increasing, and interleaved groups, respectively. Adjusted *d*-prime scores refer to *d*-prime scores adjusted for the following three covariates: *d*-prime pretest score, treatment duration, and junior Minimal English Test (jMET) score.

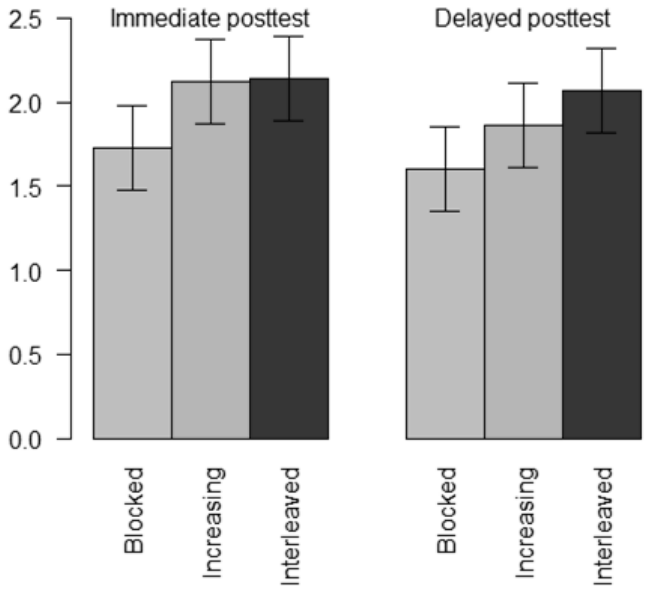
TABLE 2
Results of Mixed ANCOVA on the Grammaticality Judgment Tests

Effect	*df*	MS	*F*	*p*	η_p^2
Intercept	1	14.647	16.853	.000	.136
Condition	2	1.844	2.122	.125	.038
Time	1	0.003	0.017	.898	.000
Pretest	1	92.487	106.415	.000	.499
jMET	1	0.040	0.046	.831	.000
Treatment Duration	1	12.188	14.023	.000	.116
Condition × Pretest	2	0.931	1.072	.346	.020
Time × Condition	2	0.586	2.816	.064	.050
Time × Pretest	1	0.003	0.015	.904	.000
Time × jMET	1	0.234	1.126	.291	.010
Time × Treatment Duration	1	0.017	0.081	.777	.001
Time × Condition × Pretest	2	0.494	2.377	.098	.043

Note. jMET = junior Minimal English Test.

To probe the interaction between time and condition further, two univariate ANCOVAs were conducted separately for the immediate and delayed posttests. For the immediate posttest, neither the main effect of condition nor the interaction between condition and pretest was significant. Table 1 and Figure 7 show the adjusted means of *d*-prime scores for the three conditions. Adjusted *d*-prime scores refer to *d*-prime scores adjusted for the following three covariates: *d*-prime pretest score, treatment duration, and jMET score. The interleaved (M = 2.14) and increasing conditions (M = 2.12) yielded higher adjusted *d*-prime scores than the blocked condition (M = 1.73) at the descriptive level. However, the 95% confidence intervals overlapped among the three conditions, producing small effect sizes ($0.02 \leq d \leq 0.51$).

FIGURE 7
Adjusted Mean of *D*-Prime Scores on the Immediate and Delayed Posttests



Note. Error bars indicate 95% confidence intervals. Mean *d*-prime scores were adjusted for the following three covariates: *d*-prime pretest score, treatment duration, and junior Minimal English Test (jMET) score.

In contrast, on the delayed posttest, the main effect of condition was significant, F (2,107) = 4.512, p = .013, η_p^2 = .078. As shown in Table 1 and Figure 7 , the adjusted mean of *d*-prime score was the highest for the interleaved group (M = 2.07), followed by the increasing (M = 1.86) and blocked groups (M = 1.60). The Bonferroni method of multiple comparisons showed that the interleaved group significantly outperformed the blocked group (p = .021), and a close-to-medium effect size was observed (d = 0.64). However, there were no statistically significant differences between the blocked and the increasing group (p = .339, d = 0.37) or between the increasing and the interleaved group (p = .667, d = 0.28), yielding small effect sizes. Furthermore, the interaction between condition and pretest was marginally significant on the delayed posttest, F (2,107) = 2.542, p = .083, η_p^2 = .045. This interaction was examined further to establish whether the benefit of interleaving differed depending on the pretest scores. This part of the analysis is more exploratory in nature, and the results should be interpreted with caution. After segregating the effects of treatment duration and jMET score, correlations (partial correlations) between the pretest score and the absolute score gain from the pretest to the delayed posttest were computed. No significant correlation was found in the blocked (r = −.103, p = .543) or increasing group (r = .192, p = .276). However, the correlation was negative and moderate in the interleaved group (r = −.420, p = .009). This result suggests that the participants with lower pretest scores benefited more from interleaved practice compared to those with higher pretest scores.

<B>*Judgments of Learning*

After the immediate posttest, participants were asked to evaluate the treatment effectiveness for learning the target structures on a 7-point scale, anchored at 1 = *not effective at all* and 7 = *very effective*. The average rating was 5.31 (*SD* = 1.28, 95% CI [4.95, 5.67]), 5.61 (*SD* = 1.20, 95% CI [5.23, 5.99]), and 5.63 (*SD* = 1.21, 95% CI [5.27, 5.98]) for the blocked, increasing, and interleaved groups, respectively. No significant differences were found among the ratings given by the three groups, F (112,2) = 0.820, p = .443, η^2 = .014. These findings suggest that the participants considered the three practice schedules equally effective.

<A>DISCUSSION

The aim of the current study was to establish whether interleaving would facilitate L2 grammar learning more than blocking (RQ1). Although no statistically significant difference was found between the interleaved and blocked groups on the immediate posttest, the interleaved group significantly outperformed the blocked group on the 1-week delayed posttest. These findings are consistent with those reported in non-L2 research (see *Literature Review* section) but are somewhat inconsistent with those obtained in the L2 grammar study conducted by Pan et al. (2018). Although these researchers found benefits of interleaving over blocking when the treatment was conducted across multiple sessions (Experiments 3 and 4), no significant difference emerged between blocking and interleaving when the treatment was delivered during a single session (Experiments 1 and 2). The results obtained in the present study, however, indicate presence of the interleaving effect, even though the treatment was conducted in a single-session format. The advantage of interleaved practice observed in this study may be partly attributed to the relatively high level of learners' prior knowledge. Although participants in Pan et al.'s (2018) study had no prior knowledge of the target grammatical structures (preterite and imperfect past tenses in Spanish), the average pretest score in this study was 69.76% (*d*-prime = 1.12), which is above the chance level of 50%. Earlier research suggests that interleaving is especially effective for experienced learners (e.g., Guadagnoli et al., 1999; Rey et al., 1982). As a result, the benefits of interleaving were perhaps more pronounced in this study.

In an L2 pronunciation study, Carpenter and Mueller (2013) found that blocked practice led to better retention than interleaved practice. The present study, in contrast, showed that interleaving was more beneficial than blocking. The incongruence in the reported findings may be partially due to three factors. First, the target grammatical structures used in this study were somewhat similar to each other (e.g., three types of conditionals), while the French ortho-phonological rules used by Carpenter and Mueller were not (e.g., *eau, s, er*). In other words, the target features in this study had lower between-category discriminability than those employed by Carpenter and Mueller (2013). Previous research suggests that interleaving tends to be beneficial for target features with low between-category discriminability because, according to the Discriminative Contrast Hypothesis, by mixing exemplars from different categories, interleaving helps learners to distinguish between similar categories (Kang & Pashler, 2012). This may be one of the reasons behind the advantage of interleaving found in this study, but not in the study conducted by Carpenter and Mueller. Second, while participants in Carpenter and Mueller's study had no prior knowledge of the target pronunciation rules, the participants in this study had a relatively high level of prior knowledge. As a result, the benefits of interleaving were perhaps observed in this study, whereas blocking was more effective than interleaving in Carpenter and Mueller's study. Third, whereas Carpenter and Mueller examined the learning of form–form connections (i.e., spelling and sound), the present investigation focused on the learning of form–

meaning connections for morphosyntactic features. The difference in the nature of the target features might also be responsible for the inconsistent results.

In addition to the Discriminative Contrast Hypothesis, the benefits of interleaving in this study might also be explained by the spacing effect. While the questions targeting a particular grammatical category were studied sequentially in the blocked condition (which corresponds to massed learning), questions from a particular grammatical structure were separated by those pertaining to other structures in the interleaved condition (which corresponds to spaced learning). Because spaced learning leads to better long-term retention than would be achieved by massed learning (spacing effect), interleaving was perhaps more effective than blocking in this study. However, some researchers have failed to find the advantage of interleaving over blocking, despite the predictions of the spacing effect (e.g., Carpenter & Mueller, 2013). The current study findings suggest that the effects of the moderating factors, such as the ones mentioned previously (between-category discriminability and prior knowledge), may sometimes outweigh the benefits of spacing.

The second research question addressed the extent to which increasing practice would facilitate L2 grammar learning as opposed to blocked or interleaved practice alone. We hypothesized that increasing practice would facilitate L2 grammar learning more than blocking or interleaving would, for two reasons. First, a combination of blocking and interleaving is expected to facilitate learning because it may allow learners to detect the commonalities within each category while at the same time helping them distinguish among different categories. Second, increasing practice may be beneficial for learning because difficulty is increased gradually by using blocking first and then interleaving. This strategy provides a continuous match between the proficiency level of the learner and task difficulty and helps introduce the appropriate level of difficulty throughout the treatment (Porter et al., 2007; Porter & Magill, 2010), which enhances retention, according to the desirable difficulty framework (Bjork, 1999). This hypothesis, however, was not supported by the study findings, as on both immediate and delayed posttests, the increasing group failed to outperform the blocked or interleaved group. At the same time, although the posttest scores achieved by the increasing and interleaved groups were not statistically significantly different, the former group required a shorter treatment duration (M = 18.16 minutes) than the latter (M = 20.26 minutes). These findings suggest that, while both treatments appear equally effective, the increasing format could potentially be more efficient.

One of the reasons why increasing practice was not very effective in this study might be a relatively high level of prior knowledge on the part of learners. The theoretical underpinning for the advantage of increasing practice is based on two assumptions. First, while interleaving tends to be effective for experienced learners, blocking is posited to benefit novices (e.g., Guadagnoli et al., 1999; Rey et al., 1982). Second, as training progresses, learners' proficiency is expected to improve. These two assumptions suggest that, in the early stages of learning (when the proficiency level is low), blocking should be used, whereas interleaving should be introduced in the later stages of learning (when the proficiency level of the learner is relatively high). In the present study, however, since the participants had a relatively high level of prior knowledge, blocking in the early stage might not have been particularly effective, and it is possible that interleaving should have been used at the outset. This may be one of the factors that reduced the effectiveness of the increasing practice. This interpretation can be supported by the relative effectiveness of the three practice schedules used in this study. The interleaved condition, which included more interleaved questions (50) than the increasing condition (25), was the most

effective on the delayed posttest, whereas the blocked condition, which had no interleaved questions, turned out to be the least effective. These findings suggest that the amount of interleaved practice might have been the key factor for facilitating grammar learning in this study because of the relatively high level of prior knowledge the participants possessed.

The third research question concerned the role of prior knowledge in the effects of blocked, interleaved, and increasing practice. Based on the extant research on motor skill acquisition, we hypothesized that blocking would be effective for learners with a low level of prior knowledge, whereas interleaving would be more beneficial for learners with a high level of prior knowledge. The current study findings revealed a marginally significant interaction between condition and pretest score on the delayed posttest, suggesting that prior knowledge may moderate the practice schedule effects. While no statistically significant correlation was found between the pretest score and the score gain on the delayed posttest in the blocked (r = −.103) or increasing group (r = .192), a statistically significant, moderate and negative correlation was detected in the interleaved group (r = −.420).

These findings suggest that prior knowledge might play a role especially in interleaved practice. Contrary to our prediction, the participants with lower pretest scores benefited more from interleaving when compared to those with higher pretest scores. This finding may seem somewhat inconsistent with the results reported by other researchers indicating that interleaving is especially effective for higher-level learners, whereas blocking is sometimes beneficial for novices (e.g., Guadagnoli et al., 1999; Rey et al., 1982). However, recall that the participants with lower pretest scores in this study were not complete novices and can be regarded as relatively advanced learners. Consequently, the findings of this study may not necessarily be at odds with those yielded by prior research. The participants with lower pretest scores benefited from interleaving possibly because the level of difficulty was appropriate for them (Bjork, 1999). The participants with higher pretest scores, however, might have found interleaved practice relatively easy, which resulted in smaller benefits of interleaved practice. This is probably the reason behind the negative correlation between the pretest scores and score gains for the interleaved group. Note that, because the level of prior knowledge can only be assessed as high or low in relative terms, this interpretation is only speculative, and more rigorous research is needed to scrutinize the role of prior knowledge in the interleaving effects.

The correlation between the pretest scores and absolute score gains was not significant in the blocked group possibly because blocking was not challenging enough for most participants (the accuracy rate during learning reached a plateau in the blocked condition, as shown in Figure 6 and Online Supplement 2). Blocked practice is deemed less demanding than interleaved practice because it allows multiple practice items from the same grammatical category to be presented in sequence. This might have allowed learners that took part in the present study to induce grammatical rules regardless of their L2 proficiency, neutralizing the effects of prior knowledge. This interpretation is consistent with L2 acquisition and educational psychology research demonstrating that individual difference factors, such as prior knowledge and aptitude, exert little influence under a treatment with low information processing demand (DeKeyser, 2013; Sanz et al., 2016). Similarly, no significant correlation was found in the increasing group possibly because increasing practice, where the first half of the questions were blocked by category and learning difficulty was gradually increased, helped lessen the learning burden and neutralized the role of prior knowledge.

<A>CONCLUDING REMARKS

The findings yielded by the present study suggest that the benefits of interleaving for L2 grammar learning observed in earlier research (Pan et al., 2018) may extend to the learning of partially familiar grammatical structures. Pedagogically, the findings suggest that grammar learning may be enhanced by incorporating interleaved practice. These benefits can be derived by, for example, including multiple features that the learners have studied in the past in a language-focused task (e.g., a dictogloss task that contains mixed use of the simple past and present perfect; see Wajnryb, 1990). Both the present investigation and Pan et al.'s recent study demonstrated the advantage of interleaved practice over blocked practice, suggesting that the interleaving effect on grammar learning is reliable. Furthermore, the magnitude of effect sizes found in the present study ($d = 0.64$) as well reported by Pan et al. (2018) for Experiments 3 and 4 ($0.53 \leq d \leq 0.79$) indicates that incorporating interleaving into the curriculum is desirable (Hattie, 2008) for grammar learning.

A questionnaire administered after the immediate posttest showed that the learners considered blocking as effective as interleaving, although the latter schedule led to better long-term retention. These findings highlight the importance of raising awareness about the effects of interleaved practice. The participants in the present study were unaware of the benefits of interleaving potentially because it led to a lower proportion of correct responses during learning ($M = 77.00\%$) compared with blocking ($M = 87.23\%$). Since learners tend to assess long-term retention based on performance during the learning phase (e.g., Bjork, 1999), the participants in the interleaved group perhaps felt that learning was not progressing smoothly, potentially resulting in judgments of learning that were similar to those of the blocked group. The results reported here also demonstrate the value of highlighting that conditions that initially confuse learners and induce a low level of performance during training can be beneficial over time, whereas conditions that increase learning phase performance can be harmful in the long term (desirable difficulty framework; Bjork, 1999).

Although the findings presented in this work are valuable, this study is not without limitations. For instance, the present findings suggested that learners' prior knowledge might interact with the interleaving effect. At the same time, it is possible that the interleaving effect is moderated by individual differences in learner-related variables, such as working memory capacity or language analytic ability (e.g., Suzuki & DeKeyser, 2017b). Further research examining the effects of cognitive aptitudes on the interleaving effect would thus be a useful follow-up to this study.

Another limitation stems from the rather narrow concept of practice. The treatment in this study comprised solely multiple-choice fill-in-the-blank questions, which is a form of controlled grammar practice. Although the use of controlled practice may offer some benefits (e.g., it facilitates the acquisition of explicit, declarative knowledge, while allowing experimenters to have strict control over the treatment), it is not sufficient for L2 acquisition (e.g., Ellis & Shintani, 2014). Future researchers should thus investigate the interleaving effects as a part of less controlled grammar practice, such as picture description tasks. Moreover, in this study, an untimed grammaticality judgment test was employed as the posttest that measured only declarative–explicit knowledge. In future research, it may be useful to employ posttest measures that investigate learners' ability to use target structures fluently, such as oral elicited imitation tasks (Suzuki & Sunada, 2018).

Furthermore, the practice schedule was confounded with spacing in the current study. Specifically, while interleaved practice corresponded to spaced learning, blocked practice

corresponded to massed learning. Although this design reflects authentic learning, separating the effects of interleaving and spacing would allow us to better understand the mechanisms of interleaving effects (e.g., Kang & Pashler, 2012; Taylor & Rohrer, 2010). Further research isolating the effects of interleaving and spacing for L2 grammar learning is thus warranted.

Cognitive psychology research has long demonstrated that learning can be increased significantly by manipulating the practice schedule through spacing and interleaving. Although there has been a growing interest in the effects of spacing on L2 learning in recent years (e.g., Bird, 2010; Muñoz, 2012; Nakata & Suzuki, 2019; Nakata & Webb, 2016; Rogers, 2017; Suzuki, 2017; Suzuki & DeKeyser, 2017a), the interleaving effects have received relatively little attention in extant research. The present study demonstrated that interleaving can potentially enhance L2 grammar acquisition. Because interleaving allows instructors and curriculum designers to significantly improve learning by simply rearranging practice questions, further investigations into the effects of interleaving on L2 development would be valuable for both researchers and practitioners.

ACKNOWLEDGMENTS

This research was supported in part by Grant-in-Aid for Young Scientists (A) (#16H05943) awarded to the first author from Japan Society for the Promotion of Science. The authors are very grateful to two anonymous reviewers and the editor, Marta Antón, for their invaluable advice. We extend our gratitude also to Tomohiro Tsuchiya for his cooperation with data collection.

NOTES

1. It is true that several researchers found massing superior to spacing, a phenomenon known as the Peterson paradox. However, this phenomenon has been observed under very limited conditions, that is, when spacing intervals are very short (4−8 seconds), and learning is measured after a-less-than-8-second delay following the treatment (Cepeda et al., 2006).
2. These marginally significant interactions are worth exploring because the current study is one of the early attempts to examine the effects of interleaving on L2 grammar learning. Furthermore, the magnitude of effect sizes ($0.04 < \eta_p^2 < 0.05$) was nonnegligible, approaching medium size (Cohen, 1988).

REFERENCES

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics, 31*, 635–650.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition, 41*, 392–402.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Brown, J. D. (2014). Classical theory reliability. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1165–1181). Oxford, UK: Wiley–Blackwell.

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition, 41*, 671–682.

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, 1–15.

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.

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

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge: Cambridge University Press.

DeKeyser, R. (2013). Aptitude. In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 27–31). New York: Routledge.

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

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition, 6*, 475–485.

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. New York: Routledge.

Ferguson, G. (2001). If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes, 20*, 61–82.

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

Gass, S. (2018). SLA elicitation tasks. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 313–337). London: Palgrave Macmillan UK.

Goto, K., Maki, H., & Kasai, C. (2010). The Minimal English Test: A new method to measure English as a second language proficiency. *Evaluation & Research in Education, 23*, 91–104.

Guadagnoli, M. A., Holcomb, W. R., & Weber, T. J. (1999). The relationship between contextual interference effects and performer expertise on the learning of a putting task. *Journal of Human Movement Studies, 37*, 19–36.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to*

achievement. New York: Routledge.
Kang, S. H. (2016). The benefits of interleaved practice for learning. In J. C. Horvath, J. M. Lodge, & J. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 79–93). New York: Routledge.
Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*, 97–103.
Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. García Mayo, J. Gutierrez–Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71–92). Philadelphia/Amsterdam: John Benjamins.
Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System, 42*, 412–428.
Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.
Muñoz, C. (2012). *Intensive exposure experiences in second language learning*. Bristol, UK: Multilingual Matters.
Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition, 37*, 677–711.
Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition. 41*, 287–311.
Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition, 38*, 523–552.
Nitta, R., & Gardner, S. (2005). Consciousness-raising and practice in ELT coursebooks. *ELT Journal, 59*, 3–13.
Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417–528.
Pan, S. C., Tajrana, J., Loveletta, J., Osuna, J., & Rickard, T. (2018). Does interleaved practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills. *Journal of Educational Psychology*. Advance online publication. doi: http://dx.doi.org/10.1037/edu0000336 .
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878–912.
Porter, J. M., Landin, D., Hebert, E. P., & Baum, B. (2007). The effects of three levels of contextual interference on performance outcomes and movement patterns in golf skills. *International journal of Sports Science & Coaching, 2*, 243–255.
Porter, J. M., & Magill, R. A. (2010). Systematically increasing contextual interference is beneficial for learning sport skills. *Journal of Sports Sciences, 28*, 1277–1285.
Rey, P. D., Wughalter, E. H., & Whitehurst, M. (1982). The effects of contextual interference on females with varied experience in open sport skills. *Research Quarterly for Exercise and Sport, 53*, 108–115.
Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics, 38*, 906–911.

Rohrer, D., & Taylor, K. M. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*, 481–498.

Sanz, C., Lin, H.–J., Lado, B., Stafford, C. A., & Bowden, H. W. (2016). One size fits all? Learning conditions and working memory capacity in Ab initio language development. *Applied Linguistics, 37*, 669–692.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 77–90). Mahwah, NJ: Erlbaum.

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

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning, 67*, 512–545.

Suzuki, Y., & DeKeyser, R. (2017a). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research, 21*, 166–188.

Suzuki, Y., & DeKeyser, R. (2017b). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics, 38*, 27–56.

Suzuki, Y., & Sunada, M. (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism: Language and Cognition, 21*, 32–46.

Taylor, K. M., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*, 837–848.

Wajnryb, R. (1990). *Grammar dictation*. Oxford: Oxford University Press.

Wong, A. W.–K., Whitehill, T. L., Ma, E. P.–M., & Masters, R. (2013). Effects of practice schedules on speech motor learning. *International Journal of Speech-Language Pathology, 15*, 511–523.

Zulkiply, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition, 41*, 16–27.