The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax

Yuichi Suzuki

Kanagawa University, Japan

Satoko Yokosawa

Tsurumi Sogo High School, Japan

David Aline

Kanagawa University, Japan

Abstract

Prior research showed that interleaved practice (studying multiple skills at once) is more effective than blocked practice (studying only one skill at a time). This study aims to replicate the benefits of interleaved practice on the proceduralization of second language (L2) syntax and further examines the role of working memory (WM) in different practice schedules. Sixty English learners studied five types of relative-clause constructions under either blocked- or interleaved-practice conditions. The blocked-practice group engaged in systematic form-focused speaking practice with exemplars blocked by syntactic category, while the interleaved-practice group received mixed exemplars from the different categories. The proceduralization of grammatical knowledge was measured by analyzing the accuracy and speed indices from a picture description test, which was administered immediately and one week after the training session. Learners' WM capacity was measured using a listening-span task. Results showed that interleaved practice led to more accurate performance on both immediate and delayed posttests than blocked practice. The advantage of interleaved practice was less pronounced for the speed dimension of performance. Furthermore, interleaved practice facilitated skill development regardless of learners' WM capacity, whereas in the blocked-practice condition, learners with higher WM capacity benefited more than those with lower WM capacity in speeding up of relative clause use, which presumably reflects the proceduralization-automatization stage.

Introduction

Second language (L2) practice encompasses fundamental aspects of L2 learning such as input and output practice (Shintani, Li, & Ellis, 2013; Swain, 2005) and interaction with corrective feedback (Lyster & Saito, 2010; Mackey & Goo, 2007). The concept of practice has recently enjoyed a renewed interest among applied linguists and L2 researchers

(Dekeyser, 2007; Jones, 2018). Following the line of prior work, practice is defined as "specific activities in the second language, engaged in systematically, deliberately, with the goal of developing knowledge of and skills in the second language" (DeKeyser, 2007, p. 1).

An important question concerning L2 practice is how practice can be designed systematically to facilitate L2 acquisition. One area of investigation concerns the sequence or schedule of practice. Blocked practice, on the one hand, requires learners to perform practice activities concerning one single category of the linguistic target A for a certain amount of time (e.g., until they become accustomed to using it) and then to move on to another linguistic target B, C, D, etc. (e.g., the practice item sequence would be AAABBBCCCDDD). On the other hand, interleaved practice intermixes and presents multiple target categories; learners practice multiple categories all together (e.g., the practice item sequence would be ABCDBADCDACB). A body of cognitive research has examined the effectiveness of blocked and interleaved practice on learning (see Kang, 2016 for review). Given the potential pedagogical implications for L2 learning, cognitive psychologists and L2 researchers have started to direct attention to the potential benefits of interleaved practice in second language grammar acquisition (Nakata & Suzuki, 2019; Pan, Tajrana, Lovelett, Osuna, & Rickard, 2019; Suzuki & Sunada, 2019). Due to the paucity of research conducted to date, more empirical studies are needed to critically evaluate and further accumulate evidence for the benefits that different practice schedules may offer.

Another relevant area of study concerns how L2 teaching/practice activities should be tailored based on individual differences. Increasing interest has been paid to aptitude-treatment interaction in L2 learning (e.g., Granena, Jackson, & Yilmaz, 2016; Gurzynski-Weiss, 2017; Wen, Biedroń, & Skehan, 2017). In the framework of aptitude-treatment interaction, the effects of different practice schedules can be maximized by capitalizing on learners' cognitive aptitude, such as language analytic abilities and working memory (e.g., Erlam, 2005; Li, 2013; Sanz et al., 2016; Suzuki & DeKeyser, 2017b; Yilmaz, 2013). Identifying moderating individual difference factors in different practice schedules informs the optimization of learning conditions by taking into account learner characteristics.

The current study first aimed at replicating a part of Suzuki and Sunada's (2019) experiment wherein blocked and interleaved schedules of practice through oral picture description were compared in order to ascertain the degree of acquisition of English syntactic structures. The study further examined whether the effectiveness of either blocked or interleaved practice would be moderated by learners' working memory capacity.

Literature Review

Effectiveness of Blocked Practice and Interleaved Practice

A growing amount of evidence from cognitive psychology research has suggested that interleaved practice leads to better learning than blocked practice. The benefits of interleaved practice have been found in a variety of domains such as motor skills (e.g., Hall, Domingues, & Cavazos, 1994; Shea & Morgan, 1979), category learning (e.g., Kang & Pashler, 2012; Kornell & Bjork, 2008), chemistry study (e.g., Eglington & Kang, 2017), and mathematics problems solution (e.g., Rohrer & Taylor, 2007; Rohrer, Dedrik, & Burgess, 2014). For instance, Taylor and Rohrer (2010) examined the effects of interleaved practice using four types of mathematical formulae to solve for four parts of a prism: face, corner, edge or angle. Participants received a training session and learned how to apply the appropriate formula to solve each aspect of a prism. The result of the posttest, which was conducted one day after the training session, showed that the participants under the interleaved condition gained higher scores compared to those under the blocked-practice condition.

Interleaved practice seems to be a panacea that can be applied to a wide range of knowledge and skills; however, interleaved and blocked practice induce different cognitive processes, and each offers distinct benefits. According to the discriminative-contrast hypothesis (P. F. Carvalho & Goldstone, 2014; Zulkiply & Burt, 2013), interleaved practice encourages learners to allocate attention to the differences between (frequently-altered) categories and it facilitates comparison between exemplars from similar categories (e.g., seal vs. sea lion). Put differently, when between-category discriminability is low, interleaved practice can be more beneficial than blocked practice. In contrast, when categories to be learned are dissimilar to the extent that learners can distinguish them easily (e.g., dog vs. cat), learners may be able to learn the categories well even under blocked-practice conditions. In other words, when between-category discriminability is high, the benefits of interleaving may be restricted.

As a case in point, Carpenter and Muller (2013) demonstrated the advantage of blocked practice over interleaved practice for learning materials with high between-category discriminability. They examined the acquisition of French pronunciation by native speakers of English who had no exposure to French before the experiment. Participants saw and heard French words on a computer. Then they were asked to figure out a certain rule from a different combination of words and associate the words with the pronunciation (e.g., *eau* is pronounced /o/, as in words such as *bateau*, *fardeau*, and *rameau*). The results of this experiment revealed that blocked practice was more beneficial than interleaved practice, probably because the target pronunciation rules were very different from each other (e.g., *eau*, *ch*, *s*, *t*), and the between-category discriminability was conceivably high. In sum, the effectiveness of blocked and interleaved practice seems to vary depending on the features of the target skills or knowledge that learners acquire.

Blocked and Interleaved Practice in L2 Grammar Learning: Skill Acquisition Perspectives

Given the pedagogical importance of practice schedules, cognitive psychologists and L2 researchers have started investigating how interleaved and blocked practice influence L2 grammar learning (Nakata & Suzuki, 2019; Pan et al., 2019; Suzuki & Sunada, 2019). Before delving into these three studies, a theoretical background of L2 grammar learning and knowledge is delineated from a skill acquisition perspective (Dekeyser, 2015). According to the skill acquisition theory, L2 learners first acquire declarative knowledge (e.g., knowledge *about* grammatical rules) and use it to attain procedural knowledge (e.g., knowledge of *how* to use grammatical rules). From behavioral evidence (e.g., DeKeyser, 1997), it is stipulated that "proceduralization can become complete after just a few trials/instances." (DeKeyser, 2015, p. 95, see Ullman, 2015, for a different neurobiological view of proceduralization). Procedural knowledge still needs to be fine-tuned or automatized for even faster and stable linguistic processing.

The declarative-procedural-automatization dimensions are typically assessed by accuracy and reaction-time data in L2 behavioral research. These indices were not pure measurements exclusively for tapping into one type of knowledge, and they should be considered to tap into different knowledge dimensions to different degrees. While accuracy measures were expected to indicate both declarative and procedural knowledge, speed measures were considered to represent to a larger extent the procedural (more arguably automatized) dimension of L2 knowledge (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013; Suzuki & Dekeyser, 2017a). Research suggests that deliberate and systematic practice leads to lower error rates and shorter reaction time of target skill executions, reflecting proceduralization and potentially initial automatization of linguistic knowledge (Dekeyser, 1997; M. Li & Dekeyser, 2017; Shuai Li & Taguchi, 2014; Suzuki, 2017).

A series of experiments reported by Pan et al. (2019) investigated the effects of blocked and interleaved practice for the acquisition of preterite and imperfect past tenses in Spanish. Participants were undergraduate students without any prior knowledge of Spanish. They studied the target forms in a fill-in-the-blank format and were tested on retention of declarative knowledge, measured by a written multiple-choice test. The findings (experiments 3 and 4) suggest that the interleaved practice was more effective than blocked practice for the one-week retention of knowledge. This corroborated the advantages of interleaved practice that have been found for a variety of skills (e.g., motor skills, category and mathematics learning) and extends them to L2 grammar acquisition.

To the best of our knowledge, only two other empirical studies have examined to date the effects of blocked and interleaved practice on L2 grammar acquisition (Nakata & Suzuki, 2019; Suzuki & Sunada, 2019).[1] Unlike Pan et al. (2019), these studies focused on

[1] Both studies (Nakata & Suzuki; Suzuki & Sunada) examined the effects of a hybrid schedule as well as blocked and interleaved schedules. In the hybrid schedule, learners engaged in blocked practice first, followed by interleaved practice. Since the study only focuses on interleaved and blocked practice, the literature review here only discusses those

L2 learners who had prior learning experience (6 years or longer) of English as a foreign language. Nakata and Suzuki (2019) compared blocked and interleaved schedules for the acquisition of the English tense-aspect-mood system. Five grammatical structures (i.e., simple past, present perfect, first conditional, second conditional, and third conditional) were chosen as they are similar and often confusing for L2 learners. Specifically, simple past and present perfect are often confused (e.g., _My parents lived* in this house since 1976_), and the three types of conditionals are also similar and confusing (e.g., _If they *missed the train, they would have called_). Based on the discriminative-contrast hypothesis (see above), since between-category discriminability was expected to be low, Nakata and Suzuki predicted that interleaved practice would be more effective than blocked practice.

English-as-a-foreign-language (EFL) learners practiced five target grammatical structures in a fill-in-the-blank multiple-choice format. They were instructed to fill in the blank in a sentence (e.g., *I* _____ *a car for my daughter last Christmas*) by selecting an appropriate verb form from four options (e.g., *will buy, have bought, buy, bought*). Untimed written grammaticality judgement tests were administered as pretest, immediate, and 1-week delayed posttests to assess the acquisition and retention of declarative knowledge. The results showed that the learners who practiced under the interleaved schedule significantly outperformed those who practiced under the blocked schedule on both posttests.

Suzuki and Sunada (2019) further examined the effectiveness of blocked and interleaved practice for the acquisition of relative-clause (RC) structures. The four RC constructions were chosen (subject RC *who*, subject RC *which*, object RC *whom*, and object RC *which*). They are similar particularly in their surface form (e.g., *That is the cat which is watching the bird* vs. *That is the cat which the bird is watching*). Unlike the written practice and outcome test formats employed in Nakata and Suzuki (2019), Suzuki and Sunada (2019) trained EFL learners on an oral production (picture description) task and used the same picture description task for the pretest, immediate and 1-week delayed posttests. This oral production test can be analyzed for accuracy and speed of use of RC constructions, allowing for potentially tapping into proceduralization of grammatical knowledge (i.e., being able to use the grammatical knowledge more quickly and spontaneously). Results showed no significant difference in accuracy and speed measures between interleaved and blocked practice on the 1-week delayed posttest, although there was an advantage of interleaved practice at the descriptive level for the accuracy measure on the immediate posttest.

The equivocal findings in prior research warrant further investigation into the effects of L2 practice schedules. These two studies were different in several methodological points: target grammatical structure (tense-aspect-mood system vs. relative clause), practice task (written fill-in-the-blank task vs. oral picture description task), and outcome tests (grammaticality judgment test vs. oral picture description task). For the outcome tests, while

two schedules.

Nakata and Suzuki focused on the acquisition of declarative knowledge, Suzuki and Sunada used both accuracy and reaction time measurements to capture L2 grammatical development more systematically from a skill acquisition perspective. The need for a conceptual replication study is thus warranted to better understand the effects of different practice schedules on the development of declarative and/or procedural grammatical knowledge. Furthermore, learners' individual differences in working memory may be an important variable that influences the effects of practice (Sana, Yan, & Kim, 2017), which will be considered in the next section.

Role of Working Memory on Effectiveness of Blocked Versus Interleaved Practice

Among various individual difference factors, aptitude has recently attracted renewed attention from L2 researchers (e.g., Granena et al., 2016; Wen et al., 2017). Aptitude is defined as "a complex construct that comprises cognitive and perceptual abilities that predispose individuals to learn well or rapidly." (Granena, 2016, p. 577) One active area of research is an investigation into how aptitude moderates the effects of instruction; researchers have attempted to identify aptitude-treatment interaction patterns (Cronbach & Snow, 1977), which informs how we can cater treatment (instruction) to different types of learners. In addition to the experiments comparing the effects of interleaved versus blocked practice, an intriguing question is whether individual characteristics such as cognitive aptitude for L2 learning moderates the effectiveness of blocked and interleaved practice.

From the multi-componential view of aptitude, working memory (WM) capacity has been conceptualized as a major component of cognitive aptitude (Linck et al., 2013; Wen & Skehan, 2011). WM is a complex system whereby information is temporarily stored and manipulated simultaneously (Baddeley, 2012), and due to its nature, it has obvious links to different stages of L2 learning (Skehan, 2002, 2016). A number of studies examined the role of WM capacity under different L2 learning treatments such as deductive and inductive grammar instruction (Erlam, 2005), metalinguistic explanation (Sanz et al., 2016), types of corrective feedback (Li, 2013; Yilmaz, 2013) and practice distribution (Suzuki & Dekeyser, 2017b). A study by Suzuki and Dekeyser (2017b) is relevant to the current study because they explored the role of WM capacity (as well as language-analytic ability) for the acquisition of L2 morphology under short-interval (2 training sessions were separated over 1 day) and long-interval learning (2 training sessions were separated over 7 days) conditions. They found that learning gains were predicted by WM capacity only under the short-interval condition. This suggests that different amounts of spacing—practice schedule in a broad sense—places different levels of demands on WM. In the field of L2 research, there are no studies that we are aware of that have explored the potential link between WM capacity and blocked-interleaved practice schedules.

In the field of cognitive psychology, however, one study explored the role of WM capacity for inductive learning of statistical concepts under blocked- and interleaved-practice conditions (Sana et al., 2017). Undergraduate students in the United States studied the differences between three nonparametric statistical concepts (chi-square test, Kruskal–Wallis test, and Wilcoxon signed-ranks test) under blocked and interleaved conditions. A complex WM task (operation-span task) was used to measure participants' WM capacity. Results showed that WM capacity predicted test performance under the blocked-practice condition, but not under the interleaved-practice condition. Sana et al. (2017) suggested that the participants with higher WM capacity might have been more efficient in retrieving relevant information from long-term memory (Unsworth & Engle, 2007) in the blocked-practice condition because they were able to retain the previously learned information longer to compare with a new piece of information. In contrast, interleaved practice facilitated the retrieval of relevant information because participants were constantly required to distinguish similar concepts. Put differently, interleaved practice leveled the playing field for learners, regardless of WM capacity. The current study extends these findings to investigate the role of WM for L2 grammar learning under two different practice schedules.

Research Questions and Hypotheses

The present study investigated the effectiveness of blocked and interleaved practice for developing L2 grammatical knowledge. While Nakata and Suzuki (2019) found a clear advantage for interleaved practice, Suzuki and Sunada (2019) failed to find a significant difference between blocked and interleaved practice on a delayed posttest. This discrepancy was one of the motivations for the current replication study. The current study employed a similar research design as Suzuki and Sunada's study to further retest the effectiveness of the two practice schedules. The first research question thus focused on the replication of findings of Suzuki and Sunada's study:

1. Is interleaved practice more effective than blocked practice for the acquisition of relative clauses?

The second research question addressed the role of WM capacity in blocked and interleaved grammar practice:

2. Is the effectiveness of blocked and interleaved practice moderated by learners' WM capacity?

As the study by Sana et al. (2017) suggested, an interaction between practice schedules and WM capacity was expected. That is, learners who engage in blocked practice may be more susceptible to effects of WM capacity compared to those who engage in interleaved practice.

In blocked practice, where practice items of the same category are presented in block and items from different categories appear after a time, learners with higher WM capacity may be better able to retrieve relevant information than those with lower WM capacity. In contrast, in interleaved practice where practice items from different categories are immediately presented in sequence (e.g., a subject RC construction is followed by an object RC construction), learners may find it easier to compare and distinguish similar exemplars in their WM. This can lead to a lesser effect of WM capacity in the interleaved-practice condition.

Methods

Participants

Participants were 60 Japanese university students in three intact EFL classes (Economics, Law, and Foreign Language majors) offered at a Japanese university. Participants within each class were randomly assigned to either a blocked-practice (n = 29) or an interleaved-practice condition (n = 31). Note that the number of participants was increased for the current study (n = 18 and 19, respectively, in Suzuki and Sunada, 2019).

Prior knowledge of the linguistic target (i.e., five RC constructions) was assessed with a sorting-questions test (see Appendix A). The test consisted of eight sorting problems regarding subjective relatives, eight problems regarding objective relatives, and eight problems regarding relative adverbs. The maximum score of the sorting-questions test was 24, and the internal consistency was high (Cronbach's alpha = .90). This test was given before the experiment in order to assign participants into two groups by controlling for prior declarative knowledge of relative clauses. The average score for the blocked group was 12.90 (*SD* = 4.55), which was similar to that for the interleaved-practice group at 13.68 (*SD* = 7.03), $t(51.80) = - 0.51, p = .61$.

Target Structure

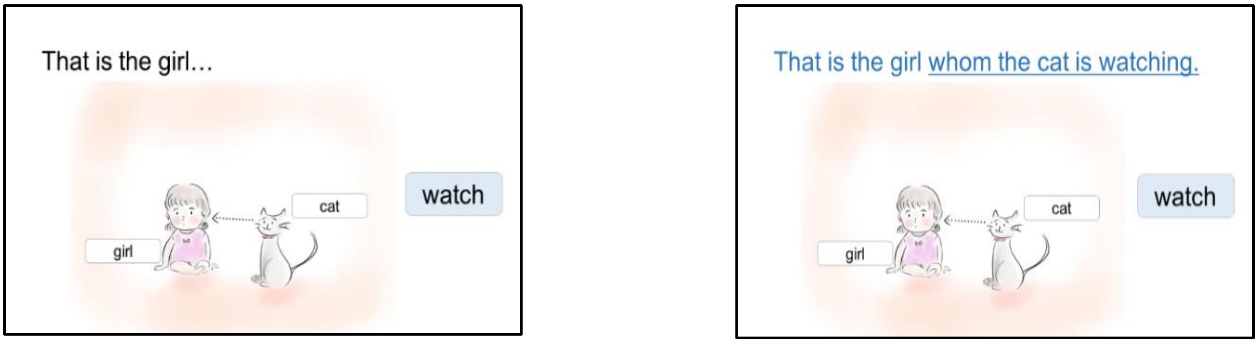
Target syntactic structures in the current study were RC constructions. RCs were chosen as the target structure because they are typically taught explicitly in junior and high school English classes, but Japanese learners often display difficulty in full control of these structures (Izumi, 2003; Mochizuki & Ortega, 2008). Although RCs are acquired relatively late, classroom instruction and practice leads to steady development (Doughty, 1991; Pavesi, 1986). Five different types of RCs were used as the target structures in this experiment:

(a) Subject relative pronoun *who* (e.g., *That is the boy who is washing the dog.*)
(b) Subject relative pronoun *which* (e.g., *That is the cat which is watching the bird.*)
(c) Object relative pronoun *whom* (e.g., *That is the girl whom the cat is watching*.)
(d) Object relative pronoun *which* (e.g., *That is the dog which the woman is carrying*.)
(e) Relative adverb *where* (e.g., *That is the park where the boy is watching the bird*.)

While only the first four types of relative pronouns were used in Suzuki and Sunada (2019), the relative adverb *where* was added to the current study. This additional structure was included here for practical reasons concerned with course management, thus maintaining the ecological validity of the learning environment.

Instruments

Training Materials. The participants were asked to describe pictures that appeared on a computer screen using appropriate relative pronouns or the relative adverb *where* (see Figure 1). All lexical items necessary for oral description (i.e., the action doer, recipient, and verbs) were shown in the picture so that participants could focus on practicing relative clauses. The first part of the sentence (the subject, *be* verb, and the antecedent) was provided both visually and aurally. The participants were given 12 seconds to respond. After that, a correct answer was provided both visually and aurally and the example sentence remained on the screen for 8 seconds (See the right panel in Figure 1). In sum, the practice trial lasted 20 seconds across all trials so that the time on training task was equal between the groups. This training procedure was identical to Suzuki and Sunada's (2019), except that the time-on task was controlled in this study.[2]



That is the girl…

That is the girl whom the cat is watching.

Figure 1. Training material

The training session consisted of 50 instances (see Appendix B). Ten different verbs were used for the relative clauses so that the participants were able to practice ten times for each target structure, which means 50 sentences were used in the training session (10 verbs x 5 structures). The ten verbs used in this experiment were the following: *carry*, *hit*, *hug*, *kick*, *kiss*, *massage*, *push*, *touch*, *wash*, and *watch*. These ten words were familiar to the participants because they are taught at the very early stage of English education in Japan or adopted into the Japanese language as loan words. In addition, these verbs require two objects so that they can be used for both subjective relative pronouns and objective relative pronouns.

[2] In Suzuki and Sunada (2019), participants were allowed to proceed to the next item before the 12-second time limit by pressing the button.

The verbs used within the relative clause were specified to allow the participants to focus on producing relative clauses.

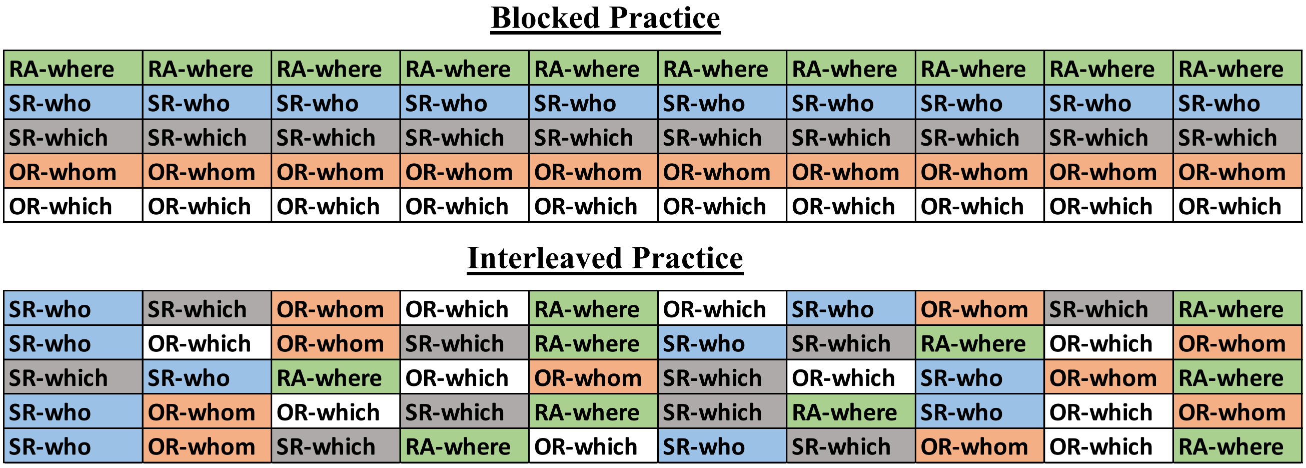
Outcome Tests. The outcome tests were designed and analyzed in the same way as in Suzuki and Sunada's (2019) study. The tests consisted of 20 items each. Four questions were created for each target structure: five target structures x four questions = 20 questions. Each test took approximately five minutes to administer. As in the training session, the participants were given 12 seconds to respond. Unlike the training session, however, no feedback was provided during any of the three tests. In order to reduce any practice effect, three equivalent versions of the tests were created (see Appendix C). The same verbs were equally employed across the three versions, while different action doers and recipients were assigned. The questions from different grammatical categories were randomly presented so that the participants were required by themselves to use the appropriate relative pronoun or adverb that would most accurately describe the picture displayed on the screen. Two measures (accuracy and speed) were derived from the performance on the outcome tests to tap into declarative and procedural knowledge to varying degrees (see the Literature Review section).

Listening-span Task. A complex WM task, the listening-span (L-span) task, adapted from a study by Osaka et al. (2003), was used to assess WM capacity. In this paper-and-pencil WM task, the participants (a) listened to a set of Japanese sentences, (b) made plausibility judgments of the sentences on a sheet of paper, and (c) recalled and wrote down on the sheet the first word of each sentence after hearing each set. They were required to comprehend the sentence while retaining the first word of each sentence for subsequent recall; the task assessed the participants' capacity to process and store information efficiently. One set of sentences gradually increases from two, to three, four and then five sentences. There were five sets for each sentence condition, resulting in a total of 70 sentences (i.e., 2 sentences x 5 sets, 3 sentences x 5 sets, 4 sentences x 5 sets, 5 sentences x 5 sets). Two sets of practice items were administered to familiarize the participants with the test procedures. Each test item was scored correct only when the word was recalled in the correct position and the plausibility judgment was correct for the sentence containing the recall word. Internal consistency indexed by Cronbach's alpha was .89.

Training Schedules

Figure 2 illustrates the practice sequences for the blocked- and interleaved-practice conditions. For the blocked-practice group, each grammatical item was designed to be studied as a set. For example, if a participant practiced producing sentences using the relative pronoun *who* 10 times first, they then practiced using the relative pronoun *which* 10 times, and then practiced using the relative pronoun *whom* 10 times. In order to counterbalance the practice order effect in the blocked-practice group, a total of three versions of the blocked-practice sequence were created. It is generally considered to be the case that subjective relative pronouns are easier to learn than objective relative pronouns (Izumi, 2003). For this reason, the order of subjective relative pronouns and objective pronouns was fixed in all the groups (i.e., subjective relative pronouns always appeared before objective relative pronouns). The only difference between the three groups was the placement of the relative adverb *where*. In Version A (illustrated in the upper panel in Figure 2), 10 practice items on the relative adverb *where* were first presented. In Version B, 10 practice items on the relative adverb *where* were presented after the 10 items on the subject relative pronoun *who* and the 10 items on the subject relative pronoun *which.* In Version C, 10 practice items on the relative adverb *where* were presented after all the other 40 items were presented.

In contrast, the categories of target grammar were intermixed for the interleaved-practice group. No items using the same grammatical structure appeared in a row. For instance, if a participant encountered a question for which they had to use the relative pronoun *who*, they were required to use one of the other structures (e.g., *which*) for the next question. (see the lower panel in Figure 2).



Blocked Practice

RA-where	RA-where	RA-where	RA-where	RA-where	RA-where	RA-where	RA-where	RA-where	RA-where
SR-who	SR-who	SR-who	SR-who	SR-who	SR-who	SR-who	SR-who	SR-who	SR-who
SR-which	SR-which	SR-which	SR-which	SR-which	SR-which	SR-which	SR-which	SR-which	SR-which
OR-whom	OR-whom	OR-whom	OR-whom	OR-whom	OR-whom	OR-whom	OR-whom	OR-whom	OR-whom
OR-which	OR-which	OR-which	OR-which	OR-which	OR-which	OR-which	OR-which	OR-which	OR-which

Interleaved Practice

SR-who	SR-which	OR-whom	OR-which	RA-where	OR-which	SR-who	OR-whom	SR-which	RA-where
SR-who	OR-which	OR-whom	SR-which	RA-where	SR-who	SR-which	RA-where	OR-which	OR-whom
SR-which	SR-who	RA-where	OR-which	OR-whom	SR-which	OR-which	SR-who	OR-whom	RA-where
SR-who	OR-whom	OR-which	SR-which	RA-where	SR-which	RA-where	SR-who	OR-which	OR-whom
SR-who	OR-whom	SR-which	RA-where	OR-which	SR-who	SR-which	OR-whom	OR-which	RA-where

Figure 2. Practice schedule

Note. SR = Subject Relative, OR = Object Relative, RA = Relative Adverb

Procedures

Participants performed the pretest, training task, and posttest individually in a computer room using the presentation software, DMDX (Forster & Forster, 2003). The pretest and the training session followed by the posttest were administered in the first session, and one week later, the delayed posttest was administrated to measure retention of the participants' grammatical knowledge. Each test contained 20 items, which lasted approximately five minutes. The training session consisted of 50 items, which took roughly 20 minutes. The test with sorting questions (see Participants section) and the listening-span task were administered one week and two weeks, respectively, prior to the first session of the experiment.

Data Coding

Three trained raters first coded the outcome tests for accuracy and speed using the sound analysis software Praat (Boersma & Weenink, 2016). They independently coded the same subset of data, and 90% of the coding matched in the initial data coding (378 out of 420 trials). After the raters discussed and resolved any discrepancies in the initial data coding, the rest of the data was divided and coded by the same raters.

Accuracy was scored as all or nothing for each test item. A credit was given to each utterance with correct word order and relative pronoun. Minor errors unrelated to the relative clauses, such as (non-)use of articles (e.g., *That is boy who is kissing dog*) and incorrect tense and aspect (e.g., *That is the boy who kissed the dog*), were ignored during coding.

Speed was coded by measuring the RT from the onset of the prompt to the end of the utterance. The data exclusion criteria and cleaning procedures followed the procedures used in Suzuki and Sunada (2019). In order to retain sufficient test items for RT analysis, the speed measure was used for analysis among the participants with accuracy rates of 50% or higher on the outcome tests.[3] In the pretest, only 15 participants scored 50% or higher, which made it impossible to use the speed measure on the pretest. A total of 47 participants scored 50% or higher on both immediate and delayed posttests, so they were included in the speed analysis (see Appendix D for mean accuracy rates of these 47 participants). All test items were retained for the speed analysis because all 20 test items were above 50% in accuracy on the immediate posttest. The RT data was cleaned by excluding cases in which (a) the response was incorrect, (b) the response included repairs and/or rephrasing (e.g., *That is the man which… who is kissing the dog*), and/or (c) a content word which was different from the specified word was used (e.g., using *man* instead of *grandfather*). Outlying item responses were also identified and treated as missing values: the group mean minus 2SD for each test item as the lower cutoff (Immediate posttest: 0%; Delayed posttest: 0.1%) and the group mean plus 2SD as the higher cutoff (Immediate posttest: 3.4%; Delayed posttest: 2.7%). The internal consistency of 20 test items as indexed by Cronbach's alpha was above .80 for all test measures (For accuracy measure: pretest = .81, immediate posttest = .86, and delayed posttest = .86. For speed measure: immediate posttest = .89 and delayed posttest = .82).

Statistical Analysis

In order to examine the effects of learning conditions, the accuracy scores were analyzed for the immediate and delayed posttests using a logistic mixed-effects model,

[3] In recognition task (e.g., lexical decision task), a more stringent criterion (e.g., 75% or higher accuracy rate) is usually applied, in order to take into account the chance level for guessing (e.g., Hulstijn et al., 2009).

implemented through the lme4 software package in R (Bates, Mächler, Bolker, & Walker, 2014). The dependent variable was a binary response (correct/incorrect), and a fixed effect was Condition. Two variables, pretest score and sorting test (see Participants section), were included as covariates in the mixed-effects model to control for the level of prior knowledge on the relative clauses. The fixed-effect variable was centered using deviation coding (blocked = -0.5, interleaved = 0.5) in order to match the inferences drawn from analyses of variance, and the covariates were scaled to standardize the scores for facilitating the score interpretation (Linck, 2016). Participant and item were treated as random effects. The maximal random-effect structure was built justified by the design (Barr, Levy, Scheepers, & Tily, 2013); converged model specifications are presented in the notes of each table. The assumption of collinearity among variables were met for all models. The effect sizes were interpreted using the benchmark in L2 research proposed by Plonsky and Oswald (2014): small (d = 0.4), medium (d = 0.7), and large (d = 1.0).[4] After statistically testing the overall between-group effect (i.e., Condition), a logistic mixed-effects model was built by adding the individual difference variable (i.e., L-span score) and the interaction term (i.e., Condition x L-span score). This second step allows for further testing the effect of a potential moderating variable on learning and the interaction with learning condition (Linck, 2016).

A similar approach to the accuracy analysis was employed to analyze the effects of learning condition on the speed measure. A linear mixed-effects model, which is used for continuous dependent variables (RT data in this case), was built for the speed measures on the immediate and delayed posttests. All fixed- and random-effect variables were identical to those for the logistic mixed-effects models for accuracy measures.

Results

Performance Change During the Training Session

Although the analysis on performance changes during the training session is not directly related to the research questions, the mean accuracy scores from the training session may aid in interpreting the test results (Figure 3, see also Appendix E for numerical data). The accuracy score of the first practice opportunity was 36.55 % and 39.35% for the blocked-practice group and the interleaved-practice group, respectively. However, the mean accuracy score of the second practice opportunity presented a significant difference between the practice conditions; the blocked-practice group achieved 76.55% accuracy while the interleaved-practice group was 41.29% and did not show much progress from the first practice opportunity. The mean accuracy score of the blocked-practice group reached beyond

[4] One of the reviewers pointed out that Cohen’s d may not be a good representation of group difference, especially considering the inclusion of the complex random effects in the mixed-effects models. Cohen's d was reported in this paper to make the results easier to be compared with previous research in SLA, but interpretation based on this statistic should be treated with caution.

90% quickly on the third opportunity. The final accuracy rate was 97.24% on the 10th practice opportunity. The performance change for the interleaved-practice group was slower than that for the blocked-practice condition. There was a spike from the second (41.29%) to the third practice opportunity (63.87%), and after that, the improvement was gradual but consistent. The final accuracy rate was 87.74% for the interleaved-practice group, which is approximately 10% lower than that for the blocked-practice group. The blocked-practice group consistently outperformed the interleaved-practice group throughout the training phase (no overlapping 95% confidence intervals, except for the 7th practice opportunity).

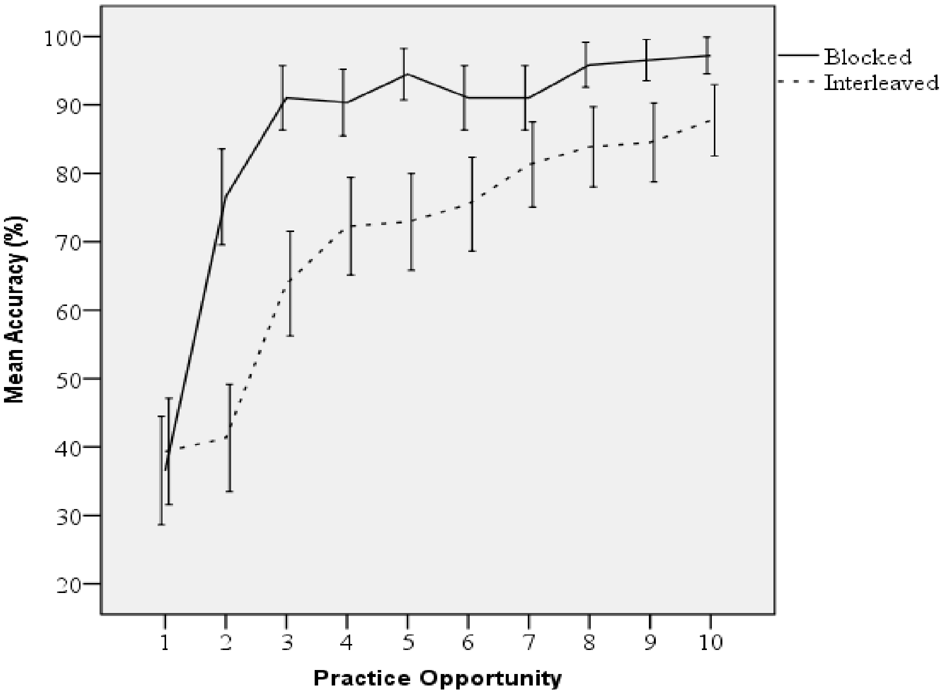


Figure 3. The mean accuracy score of the blocked practice group and the interleaved practice group during the training session

Notes. Practice opportunity represents the order of practice items within each RC type. For instance, mean accuracy rate at the 1st practice opportunity was computed by averaging the accuracy scores of the first item across the five RC constructions (SR-who, SR-which, OR-whom, OR-which, RA-where).

Learning Outcome: Accuracy Measures

As shown in Table 1, the learners in both conditions showed approximately 30% accuracy on the pretest. On the immediate posttest, the interleaved-practice group outperformed the blocked-practice group. The mean accuracy score of the delayed posttest, conducted one week after the treatment, still showed the advantage of the interleaved practice.

Table 1

Descriptive Statistics of Accuracy Scores in Two Conditions

	Blocked				Interleaved			
	n	*M*	*SD*	95% CI	*n*	*M*	*SD*	95% CI
Accuracy Measures (percentage)								
Pretest	29	32.04	20.08	[24.4,39.68]	31	31.32	22.02	[23.24,39.39]
Immediate	29	78.74	21.34	[70.62,86.85]	31	90.86	13.66	[86.37,95.35]
Delayed	29	62.64	23.77	[53.6,71.68]	31	72.45	22.95	[64.58,80.32]
Speed Measures (milliseconds)								
Pretest	-	-	-	-	-	-	-	-
Immediate	21	5828	966	[5389,6268]	26	5522	5522	[5102,5943]
Delayed	21	6034	825	[5659,6410]	26	6151	6151	[5708,6595]

The results of the logistic mixed-effects model on accuracy scores are reported in Table 2. On the immediate posttest, the fixed effect of Condition was significant with a small-medium effect size (*z* = 3.07, *p* < .01, *d* = 0.58 [0.06, 1.09]). Similarly, Condition was also significant on the delayed posttest with a small effect size (*z* = 2.09, *p* = .04, *d* = 0.37 [-0.14, 0.88]). One covariate (pretest) was significant in both posttests (*p* < .01). Overall, the interleaved practice was more effective than the blocked practice for the acquisition and retention of knowledge of relative clauses, despite the fact that the blocked practice showed better performance during the treatment.

Table 2

Results of Logistic Mixed-Effects Models (Accuracy Scores)

	Immediate posttest				Delayed posttest			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	2.17	0.41	5.25	.00	0.65	0.33	1.96	.05
Condition	1.35	0.44	3.07	.00	0.75	0.36	2.09	.04
Pretest	1.28	0.31	4.09	.00	0.73	0.23	3.12	.00
Sorting Test	-0.03	0.27	-0.12	.91	-0.01	0.22	-0.03	.98

Note. Converged model formula for the immediate posttest: Accuracy ~ Condition + Pretest + SortingTest + (Pretest + SortingTest | Subject) + (Condition + Pretest + SortingTest | Item), Converged model formula for the delayed posttest: Accuracy ~ Condition + Pretest + SortingTest + (Pretest + SortingTest | Subject) + (Condition + Pretest + SortingTest | Item)

Learning Outcome: Speed Measures

The results from the speed measures also showed an advantage (about 300 milliseconds) for the interleaved-practice condition over the blocked-practice condition on the immediate posttest (Table 1). On the delayed posttest, this difference disappeared as both conditions showed slower performance.

Table 3 presents the results of the linear mixed-effects model. On the immediate posttest, the fixed effect of Condition was marginally significant with a small effect size (*t* = - 1.86, *p* = .07, *d* = 0.48 [-0.11, 1.06]). The advantage of interleaved practice over blocked practice was not significant, possibly due to the relatively smaller sample size for the speed analyses. Its magnitude of effect size (0.48) fell in between that found in the accuracy measures of immediate posttest (0.58) and delayed posttest (0.37). On the delayed posttest, however, there was no significant difference between the two conditions with null effect size (*t* = - 0.04, *p* = .97, *d* = 0.01 [-0.57, 0.58]).

Table 3

Results of Linear Mixed-Effects Models (Speed Measures)

	Immediate posttest				Delayed posttest			
	Estimate	*SE*	*t*	*p*	Estimate	*SE*	*t*	*p*
Intercept	5913.02	223.00	26.52	.00	6342.62	297.64	21.31	.00
Condition	-498.93	268.89	-1.86	.07	-11.50	265.06	-0.04	.97
Pretest	-419.25	172.26	-2.43	.03	-403.35	151.88	-2.66	.02
Sorting Test	-13.63	165.94	-0.08	.94	-71.11	152.40	-0.47	.64

Note. Converged model formula for the immediate posttest: RT ~ Condition + Pretest + SortingTest + (Pretest + SortingTest | Subject) + (Condition + Pretest | Item), Converged model formula for the delayed posttest: rt ~ Condition + Pretest + SortingTest + (Pretest + SortingTest | Subject) + (Condition + Pretest | Item)

Role of Working Memory Capacity

The role of working memory capacity was further examined, and a potential interaction between working memory capacity and learning condition was explored. The average listening-span test score was 50.56 (*SD* = 11.63) and 52.36 (*SD* = 7.82) in the blocked- and interleaved-practice conditions, respectively.[5] The logistic mixed-effects model was built with L-span scores. Here, we focus on the effects of L-span score and the interaction between Condition and L-span score (see Appendices F and G for the comprehensive results of the model).

For the accuracy data analysis, the effect of L-span score was significant for both the immediate and delayed posttests (*z* = 2.52, *p* = .01; *z* = 1.99, *p* = .046). However, no significant interaction between Condition and L-span was found on either immediate or

[5] Five participants (two in the blocked-practice and three in the interleaved-practice groups) did not take the listening-span task, so they were excluded from further analysis.

delayed posttest (z = - 0.27, p = .79; z = 0.86, p = .39). For the speed measure analysis, the effect of L-span score was significant again for both posttests (t = - 3.08, p = .01; t = - 2.77, p = .02). Critically, the interaction between Condition and L-span score was significant in both immediate and delayed posttests (t = 2.93, p = .01; t = 2.80, p = .01).[6] These significant interactions are illustrated in and seem to be driven by performance in the blocked-practice group. That is, RT decreased as a function of L-span score only in the blocked-practice condition, while there seems to be no systematic relationship between the posttest performance and L-span score in the interleaved-practice condition. This suggests that the learners' production speed on the posttests attained through blocked practice is influenced by working memory capacity, while such effect of working memory is negligible among the learners in the interleaved-practice condition.

[6] Likelihood ratio tests also confirmed that the models with the interaction term was better than the models without it (chi-square difference = 3.83, p = .0503; chi-square difference = 5.42, p = .02, for immediate and delayed posttests, respectively).

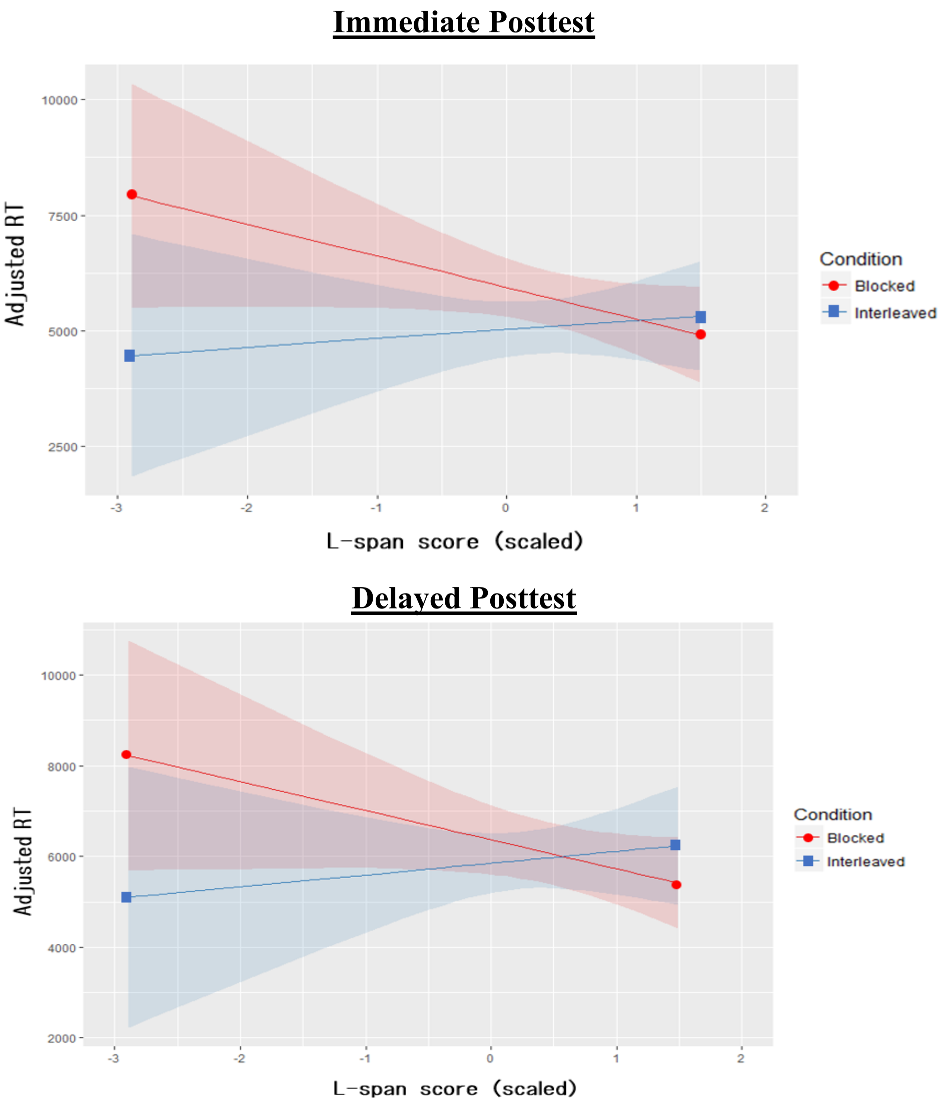


Figure 4. Significant interactions between practice schedule (blocked vs. interleaved) and working memory capacity in speed analysis of immediate and delayed posttests
Note. The shaded areas indicate 95% confidence intervals.

Discussion

Effects of Blocked and Interleaved Practice for Proceduralization

The present study examined whether interleaved practice is more effective than blocked practice for the acquisition of L2 syntax. The results of this study showed an advantage for interleaved practice over blocked practice in L2 grammar learning, which corroborates many of the previous studies in cognitive psychology (Mayfield & Chase, 2002; Rohrer, 2012; Rohrer & Taylor, 2007; Rohrer et al., 2014; Taylor & Rohrer, 2010). Since the conceptual replication of previous research was one of the chief objectives of this study, the methodological differences and findings of pertinent research targeting Japanese EFL learners (Suzuki & Sunada, 2019 and Nakata and Suzuki, 2019) were summarized to facilitate the interpretations (Table 4).

Table 4

Summary of Prior Research Methodology and Findings on Blocked and Interleaved Practice Among English L2 Learners

	Current Study	**Suzuki and Sunada (2019)**	**Nakata and Suzuki (2019)**
Participants	Japanese university students (blocked=29, interleaved=31)	Japanese university students (blocked=18, interleaved=19)	Japanese university students (blocked=39, interleaved=40)
Target Structure	English Relative Clause Constructions (subject RC [*who*, *which*], object RC [*whom*, *which*] and relative adverb *where*)	English Relative Clause Constructions (subject RC [*who*, *which*] and object RC [*whom*, *which*])	English Tense Systems (simple past, present perfect, 1st, 2nd and 3rd conditionals)
Practice Tasks	Oral picture description	Oral picture description	Multiple-choice format
Practice Items	50 items	64 items	50 items
Outcome Tasks	Oral picture description	Oral picture description	Written grammaticality judgement
Timing of Outcome Tasks	Pretest, immediate posttest, 1-week delayed posttset	Pretest, immediate posttest, 1-week delayed posttset	Pretest, immediate posttest, 1-week delayed posttset
Findings	<u>Accuracy</u> Immediate: Interleaved > Blocked (*d* = 0.58) Delayed: Interleaved = Blocked (*d* = 0.37) <u>Speed</u> Immediate: Interleaved > Blocked (*d* = 0.48) Delayed: Interleaved = Blocked (*d* = 0.01)	<u>Accuracy</u> Immediate: Interleaved > Blocked (*d* = 0.47) Delayed: Interleaved = Blocked (*d* = 0.01) <u>Speed</u> Immediate: Interleaved > Blocked (*d* = 0.19) Delayed: Interleaved = Blocked (*d* = 0.05)	<u>Accuracy</u> Immediate: Interleaved > Blocked (*d* = 0.51) Delayed: Interleaved > Blocked (*d* = 0.64)

In light of the findings from Suzuki and Sunada's experiment (2019), the current findings seem somewhat inconsistent. Overall, the benefits found for interleaved practice seem to be larger in the current study than the previous one. Yet, the direction of effects did not diverge. In terms of accuracy, Suzuki and Sunada (2019) found null-medium effect sizes (*d* = 0.47 and 0.01 for the immediate and delayed posttests, respectively), while the current study revealed slightly larger effect sizes (*d* = 0.58 and 0.37). For the speed measures, Cohen's *d* was also lower in Suzuki and Sunada (0.19 and 0.05) than in the current study (*d* = 0.48 and 0.01).

While the findings from these studies are not completely inconsistent, it may be worth speculating on some factors that may account for the relatively small discrepancy. First, the relative adverb *where* was added in the current study (hence, five types of structures rather than four), which might have conferred a slight advantage on the interleaved-practice condition possibly due to the higher demands of discriminating more structures (e.g., Carvalho & Goldstone, 2017). Second, the increased number of participants probably aided in manifesting a clearer pattern in the findings. Last but not least, there were more training items in Suzuki and Sunada (2019) than in the current study. When comparing the accuracy performance during the training between the two studies (compare Figures 3 and 5), the gap between the blocked-practice and the interleaved-practice conditions seem to have been smaller in the latter training phase (between the 8th and 14th practice opportunities) in Suzuki and Sunada's training data. Possibly, the more practice opportunities might have diminished the difference between the two practice schedules.[7] These potential explanations are neither exhaustive nor conclusive; further research is needed to gather more empirical evidence with a larger sample.

[7] We thank an anonymous reviewer for suggesting the need to further interpret the training performance.

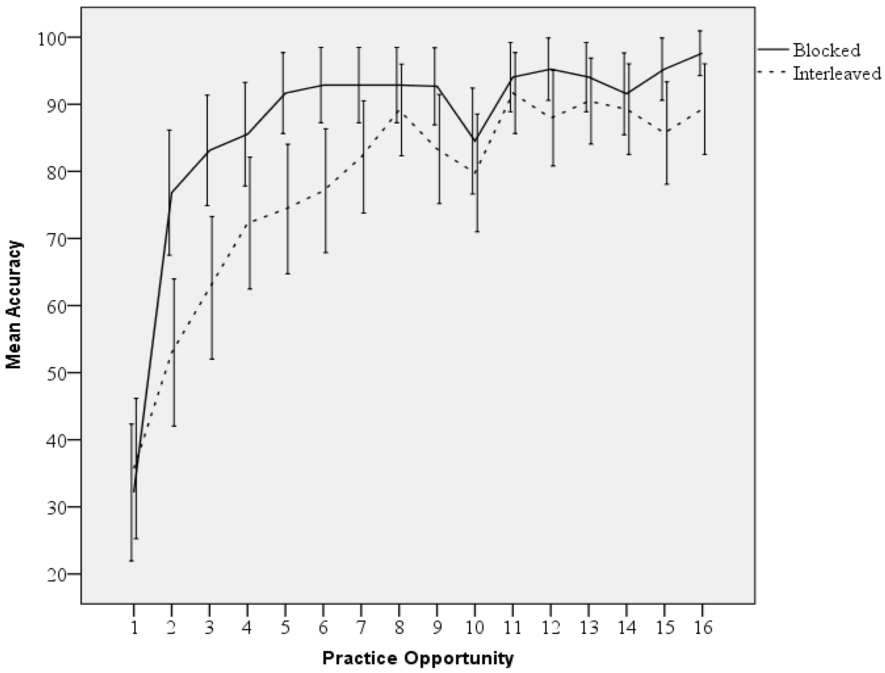


Figure 5. The mean accuracy score of the blocked practice group and the interleaved practice group during the training session reported in Suzuki & Sunada (2019)
Note. This figure was edited from the figure of Appendix H originally reported in Suzuki & Sunada (2019). Accuracy rates of the blocked- and interleaved-practice conditions were illustrated here (for the full results including the hybrid-practice condition, see Appendix H in Suzuki & Sunada, 2019).

If we bring our attention to the commonality of the current findings with other previous research, the current study corroborates two other previous studies on interleaved practice effect in L2 grammar learning (Nakata & Suzuki, 2019; Pan et al., 2019). In particular, Nakata and Suzuki (2019) targeted a similar population (English learners in Japan) and merits closer attention. There are major differences between Nakata and Suzuki's study and the current study, such as target grammatical structure (tense-aspect-mood system vs. relative clause), practice task (written fill-in-the-blank task vs. oral picture description task), and outcome tests (written grammaticality judgment test vs. oral picture description task). Despite these methodological differences, the current study successfully replicated the benefits of interleaved practice. This study found a small-medium effect size for the practice schedule on accuracy measures (d = 0.58 and 0.37 for the immediate and delayed posttests, respectively). This fell into a similar range to what Nakata and Suzuki (2019) found: 0.51 and 0.64 on the immediate and 1-week delayed posttest in accuracy (GJT) measures. In sum, our study is a stepping stone to accumulating more evidence for the benefits of interleaved practice for L2 grammar learning.

Interaction between Practice Schedule and WM Capacity

Regarding the second research question, the current findings showed that the effects of different practice schedules are susceptible to learners' individual differences in WM capacity. WM capacity plays a significant role only among learners in the blocked-practice condition. Specifically, while no significant difference was seen in terms of the practice condition for the participants with higher WM, participants with lower WM capacity showed slower performance under the blocked-practice condition compared to those who studied through the interleaved-practice condition. This pattern of results corroborates the findings in cognitive psychology reported by Sana et al. (2017) and demonstrates that the selective role of WM in blocked practice extends to L2 grammar learning.

As Sana et al. (2017) suggested, higher WM capacity supports efficient retrieval of relevant information from memory (Unsworth & Engle, 2007) and may facilitate blocked practice. In the blocked-practice condition, when learners need to compare one RC type (e.g., subject RC *who*) with other RC types, they had to deliberately retrieve a previously-encountered exemplar or syntactic rule of different RC types (e.g., object RC *which*). This cognitive process (retrieving declarative information that was encoded further apart [each RC type was presented 10 items apart]) might have taxed WM heavily and overloaded the blocked-practice learners. In contrast, interleaved practice, where practice items from different categories (e.g., subject RC and object RC) are presented *immediately*, might have prompted learners to compare these *recently-activated* exemplars and/or rules of different RC types in their WM, which presumably lessened the burden on WM. This interpretation may be in part supported by the lower accuracy rates during the training phase in the interleaved-practice condition, suggesting that learners were constantly challenged to produce different RC structures. Broadly, this type of difficulty imposed on learners' mental capacity (i.e., WM) could have induced desirable difficulty to enhance L2 grammar learning (Bjork, 2018; Suzuki, Nakata, & Dekeyser, 2019). In contrast, in the blocked-practice condition, learners were probably not challenged enough to accurately produce the same type of RC constructions during the training (90% in accuracy after the third practice opportunity).

An intriguing aspect of the current findings is that the significant interaction between practice schedule and WM was found only on the speed measure, not on the accuracy measure. Knowledge accumulation and control (speeded-up use) of knowledge progress simultaneously; distinguishing the two stages of skill acquisition is often hard to accomplish with behavioral tests (Hulstijn et al., 2009). The two behavioral measures taken from the picture description outcome test (i.e., correctness of utterance and the utterance speed in which the accurate speech is delivered) might have tapped into different stages of L2 development. Accuracy was most likely to index both declarative and initial proceduralization, whereas speed measures might have primarily tapped proceduralization and automatization. With regard to the latter proceduralization-automatization stage, Skehan

(2016) conjectures that working memory may be relevant in his aptitude-acquisition framework. The proceduralization of morphosyntactic processing may be facilitated by central executive operations in WM because more efficient access to long-term memory via WM supports proceduralization. Although Skehan (2016) acknowledges that there is little empirical evidence explicitly indicating the link between working memory and L2 proceduralization to date, the current findings seem to suggest this possibility. Yet, this interpretation should be taken with a grain of salt and attested with further investigation.

Limitations and Future Directions

Several limitations of the current study are highlighted that shed light on directions for future research. First, one reviewer pointed out that the tasks and procedures of the training and outcome tasks were identical and that there could have been practice effects. In order to minimize this potential confounding factor, an oral production task, such as an elicited imitation task targeting the RC construction knowledge (Suzuki & Sunada, 2018), may be useful.

Second, the same reviewer further suggested making the more detailed interpretations of the information provided by the mixed models (e.g., log odds, the covariate of pretest score, and random-effect effect structures). Interpreting these rich and complex data was beyond the scope of this study but would allow for further interpretations that may deepen understanding of the results (see Appendix H for random-effect structures).

Third, since the current study utilized only one WM task (L-span) to tap WM capacity due to the practical constraints, multiple WM tasks should be used to tap into WM capacity more comprehensively. Additionally, a different scoring method may be useful to capture different aspects of WM capacity (Conway et al., 2005).

Fourth, although one of the key issues of this study was a replication of previous research, no prior power analysis was conducted. Future research that aims to replicate this study and other related studies should conduct a power analysis to collect the required number of participants judiciously.

Lastly, it is acknowledged that extra caution should be used in interpreting the results of RT measurements in this study because pretest RT was not available to control for the potential difference between the two groups. If learners with higher initial accuracy rates (before the treatment) are tested, we can elucidate the effects of blocked and interleaved practice for the later proceduralization-automatization stages of grammatical development. Furthermore, the effects of WM capacity could also be examined during the training.

Conclusion

The main objective of this study was to explore practice schedules which enhance proceduralization of English syntactic structures from the skill acquisition theory perspective

(Dekeyser, 2015). Also examined was how the effects of practice schedules vary depending on learners' WM capacity. The results of the experiment demonstrated that participants in the interleaved-practice group achieved approximately 10% higher scores over the blocked-practice group on both the posttest and the delayed posttest (small-medium effect sizes). Furthermore, the present study found a potential interaction between the practice schedule and WM. While the effects of blocked practice may depend on WM capacity for the later stage of proceduralization and automatization in particular, the effects of interleaved practice did not seem to be influenced by individual differences in WM capacity. These findings suggest that interleaved practice may neutralize the effects of WM capacity for practice. This adds further weight to the significance of interleaved practice for facilitating L2 grammar acquisition. One pedagogical implication of the present study is utilizing interleaved practice for isolated L2 grammar practice; this approach is particularly effective for L2 learners with lower WM capacity. This recommendation is tentative given that the nature of the L2 practice in this study is limited to form-focused practice. Further research is necessary to understand the applicability and limits of interleaving effects in aspects of L2 learning.

References

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1-29. doi:10.1146/annurev-psych-120710-100422

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

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

Bjork, R. A. (2018). Being suspicious of the sense of ease and undeterred by the sense of difficulty: Looking back at schmidt and Bjork (1992). *Perspectives on Psychological Science*, *13*, 146-148. doi:10.1177/1745691617690642

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

Carvalho, P., & Goldstone, R. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of experimental psychology. Learning, memory, and cognition*.

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*, *42*, 481-495. doi:10.3758/s13421-013-0371-0

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769-786. doi:10.3758/bf03196772

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington Pub.

De Jong, N. H., 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

Dekeyser, R. M. (1997). Beyond explicit rule learning. *Studies in Second Language Acquisition*, *19*, 195-221.

Dekeyser, R. M. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. New York, NY: Cambridge University Press.

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, NY: Routledge.

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of sl relativization. *Studies in Second Language Acquisition*, *13*, 431-469. doi:10.1017/S0272263100010287

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. doi:10.1016/j.jarmac.2017.07.005

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, *9*, 147-172. doi:10.1191/1362168805lr161oa

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

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition*, *34*, 445-474. doi:10.1017/S0272263112000149

Granena, G. (2016). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Psycholinguistics*, *37*, 577-600. doi:10.1017/S0142716415000120

Granena, G., Jackson, D. O., & Yilmaz, Y. (2016). *Cognitive individual differences in second language processing and acquisition*. Amsterdam, the Netherlands: John Benjamins.

Gurzynski-Weiss, L. (2017). *Expanding individual difference research in the interaction approach*. Amsterdam, the Netherlands: John Benjamins.

Hall, K. G., Domingues, D. A., & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Perceptual and Motor Skills*, *78*, 835-841. doi:10.2466/pms.1994.78.3.835

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*, 555-582. doi:10.1017/S0142716409990014

Izumi, S. (2003). Processing difficulty in comprehension and production of relative clauses by learners of English as a second language. *Language Learning*, *53*, 285-323. doi:10.1111/1467-9922.00218

Jones, C. (2018). *Practice in second language learning*. Cambridge: Cambridge University Press.

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, NY: Routledge.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*, 97-103. doi:10.1002/acp.1801

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

Li, M., & Dekeyser, R. M. (2017). Perception practice, production practice, and musical ability in L2 mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*, 593-620. doi:10.1017/S0272263116000358

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal*, *97*, 634-654. doi:10.1111/j.1540-4781.2013.12030.x

Li, S., & Taguchi, N. (2014). The effects of practice modality on pragmatic development in L2 chinese. *The Modern Language Journal*, *98*, 794-812. doi:10.1111/modl.12123

Linck, J. A. (2016). Analyzing individual differences in second language research: The benefits of mixed effects models. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 105-128). Amsterdam: John Benjamins.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., . . . Doughty, C. J. (2013). Hi-lab: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*, 530-566. doi:10.1111/lang.12011

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA. *Studies in Second Language Acquisition*, *32*, 265-302. doi:10.1017/s0272263109990520

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in SLA: A collection of empirical studies* (pp. 408–452). New York, NY: Oxford University Press.

Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativization. *Language Teaching Research*, *12*, 11-37. doi:10.1177/1362168807084492

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

Osaka, M., Osaka, N., Kondo, H., Morishita, M., Fukuyama, H., Aso, T., & Shibasaki, H. (2003). The neural basis of individual differences in working memory capacity: An fmri study. *NeuroImage*, *18*, 789-797. doi:10.1016/S1053-8119(02)00032-0

Pan, S. C., Tajrana, J., Lovelett, J., Osuna, J., & Rickard, T. (2019). 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:10.1037/edu0000336

Pavesi, M. (1986). Markedness, discoursal modes, and relative clause formation in a formal and an informal context. *Studies in Second Language Acquisition*, *8*, 38-55. doi:10.1017/S0272263100005829

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

Rohrer, D., Dedrik, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, *21*, 1323-1330. doi:10.3758/s13423-014-0588-3

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*, 481-498. doi:10.1007/s11251-007-9015-8

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*, 84-98. doi:10.1037/edu0000119

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. doi:10.1093/applin/amu058

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 179-187.

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, *63*, 296-329. doi:10.1111/lang.12001

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69-94). Philadelphia, PA: John Benjamins.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 17-40). Amsterdam: John Benjamins.

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

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

Suzuki, Y., & Dekeyser, R. M. (2017b). Exploratory research on L2 practice distribution: An aptitude x treatment interaction. *Applied Psycholinguistics*, *38*, 27-56. doi:10.1017/S0142716416000084

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

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. doi:10.1017/S1366728916000857

Suzuki, Y., & Sunada, M. (2019). Dynamic interplay between practice type and practice schedule in a second language: The potential and limits of skill transfer and practice schedule. *Studies in Second Language Acquisition*, *Online Advanced Access*.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471-483). Mahwah, NJ: Erlbaum.

Ullman, M. T. (2015). The declarative/procedural model: A neurobiologically-motivated theory of first and second language. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 135-158). New York, NY: Routledge.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104-132. doi:10.1037/0033-295X.114.1.104

Wen, Z., Biedroń, A., & Skehan, P. (2017). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching*, *50*, 1-31. doi:10.1017/S0261444816000276

Wen, Z., & Skehan, P. (2011). A new perspective on foreign language aptitude research: Building and supporting a case for "working memory as language aptitude". *A Journal of English Language, Literatures in English and Cultural Studies*, *60*, 15-44.

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied linguistics*, *34*, 344-368. doi:10.1093/applin/ams044

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. doi:10.3758/s13421-012-0238-9.