Studies in Second Language Acquisition **0** (2019), 1–29
doi:10.1017/S0272263119000470

Research Article

DYNAMIC INTERPLAY BETWEEN PRACTICE TYPE AND PRACTICE SCHEDULE IN A SECOND LANGUAGE

THE POTENTIAL AND LIMITS OF SKILL TRANSFER AND PRACTICE SCHEDULE

Yuichi Suzuki*

Kanagawa University, Japan

Midori Sunada

Nihon University, Japan

Abstract

To investigate the skill transfer and the effects of practice schedules in the learning of second language syntax, 129 intermediate-level English learners were divided into six groups, based on practice format (input vs. output practice) and practice schedule (blocked vs. interleaved vs. hybrid [blocked + interleaved]). Analyses revealed that the learners tested on the skill they had practiced outperformed those who were tested on the nonpracticed skill. This pattern was particularly pronounced in comprehension processing speed and production accuracy. Moreover, hybrid practice facilitated skill development more than blocked or interleaved practice alone. Furthermore, a dynamic interplay was detected among practice format, schedule, and learners' prior knowledge. Hybrid practice led to the least transfer from receptive skills (gained through input practice) to productive skills. Unlike interleaved practice effects, the effects of blocked practice on comprehension speed were more susceptible to learners' prior processing speed.

INTRODUCTION

Practice is an essential component of second language (L2) learning. From the perspective of skill acquisition theory, L2 knowledge and skills develop through deliberate,

The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at: https://www.iris-database.org/iris/app/home/detail?id=york%3a936411&ref=search.

We would like to show our gratitude to Mr. Atsushi Miura, Ms. Satoko Yokosawa, Dr. Baikuntha Bhatta, and Ms. Misaki Kuratsubo for their assistance throughout the study.

*Correspondence concerning this article should be addressed to Yuichi Suzuki, Faculty of Foreign Languages, Kanagawa University, 3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa, 221-8686, Japan. E-mail: szky819@kanagawa-u.ac.jp

systematic, and extensive practice (De Jong, 2005; DeKeyser, 1997; Elgort, 2011; Li & DeKeyser, 2017; Li & Taguchi, 2014). One of the primary objectives of skill acquisition theory research agenda is thus to explore "what practice activities are best for whom, for what structures, for which skill at what time and in what context" (DeKeyser, 2018, p. xv). Extending the growing body of research on L2 practice (see Suzuki, Nakata, & DeKeyser, 2019a for an overview), the work reported in this article focuses on two lines of research that have examined the nature of L2 practice: (a) practice types and (b) practice schedule.

The first line of research aims to identify the most effective practice format for L2 learners. One of the central issues of L2 practice concerns the relative effectiveness of input and output (receptive and productive) practice (see Shintani, Li, & Ellis, 2013). Research on the role of input and output has resulted in diverse findings, with obvious theoretical and pedagogical implications (e.g., DeKeyser & Botana, 2015; Krashen, 1985; Sakai & Moorman, 2018; Swain, 1985; VanPatten, 2002). For example, while Krashen (1985) and VanPatten (2002) argued for the primary or even sole role of input processing for L2 acquisition, DeKeyser (1997, 2015), drawing on the skill acquisition theory, posited that both input and output practice have a unique role in the development of comprehension and production skills, respectively.

The second strand of research investigates how practice should be scheduled to optimize L2 learning. A large body of cognitive psychology studies suggests that optimal temporal spacing can enhance learning and retention across diverse skill and knowledge types (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). This issue has recently attracted interest of L2 researchers, aiming to examine the extent to which practical and theoretical cognitive psychology findings apply to L2 learning (Bird, 2010; Miles, 2014; Nakata, 2015; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017). However, the findings of previous research on temporal spacing effects are mixed, and the generalization is far from straightforward due to the limited scope of investigations (e.g., adult L2 learners tested in laboratory contexts, but see Rogers & Cheung, 2018; Serrano & Huang, 2018) to provide reliable pedagogical principles for L2 instruction (see Suzuki, 2017 for the detailed discussion of mixed findings on temporal spacing effects in L2 grammar learning). Even less understood is a related issue of blocking and interleaving effects on L2 learning. In this context, blocking refers to the learning of one target feature at a time, whereas interleaving involves learning multiple types of features concurrently (Kang, 2016). For instance, when practicing the use of subject relative clauses (RCs) and object RCs under the blocked-practice condition, learners process 20 instances of subject RCs followed by 20 instances of object RCs. Conversely, under the interleaved-practice condition, learners are presented with a mix of instances of both RC types (e.g., no more than two instances of one RC type were presented sequentially). Because cognitive psychologists have revealed the benefits of interleaved practice in a variety of learning contexts (e.g., Kornell & Bjork, 2008; Rohrer & Taylor, 2007; Rohrer, Dedrik, & Burgess, 2014; Shea & Morgan, 1979), it may be possible to enhance L2 learning by manipulating practice schedule (Nakata & Suzuki, 2019). Furthermore, the effects of schedules that combined blocked and interleaved practice (i.e., hybrid practice) have recently been explored (Porter, Landin, Hebert, & Baum, 2007; Porter & Magill, 2010; Wong, Whitehill, Ma, & Masters, 2013; Yan, Soderstrom, Seneviratna, Bjork, & Bjork, 2017).

In extant research, the two aforementioned issues of L2 practice (practice format and schedule) have been examined independently, thus failing to elucidate how these two factors jointly influence L2 development. This shortcoming calls for empirical investigation that unites the two lines of research, as pedagogical decisions are usually made by taking more than one factor into consideration. The current article addresses the gap noted previously by presenting a first empirical study in which the differential effects of practice type (input and output) and schedule (blocked, interleaved, and hybrid [blocked + interleaved]) on L2 grammar learning are examined simultaneously.

PRACTICE TYPE: SKILL-SPECIFICITY HYPOTHESIS

A variety of aspects in L2 learning can be understood within the framework of skill acquisition theory (DeKeyser, 2015), which is based on the Anderson's adaptive control of thought-rational (ACT-R) model (Anderson et al., 2004). The theory accounts for how people acquire a set of skills in L2, as well as other motor and cognitive skills, and distinguishes declarative and procedural knowledge. In L2 classrooms, learners typically attain declarative knowledge (e.g., knowing about grammatical rules), and they use it for practicing a variety of L2 skills, which results in the independent development of procedural knowledge (e.g., comprehension and production in L2). Extensive practice in one skill leads to gradual improvement or automatization in performance (e.g., lower error rates and faster response time [RT]), which follows power-of-law function curves similar to those describing nonlanguage skill acquisition (DeKeyser, 1997; Ferman, Olshtain, Schechtman, & Karni, 2009; Robinson, 1997). Procedural and automatized knowledge[1] form a foundation for comprehending and producing L2 quickly and efficiently in real-life communication settings.

The skill acquisition theory stipulates that deliberate practice results in proceduralization/automatization in very specific ways. When leaners proceduralize one skill through practice (e.g., comprehension skill acquisition through input practice), this tends to result in a specific procedure that cannot be easily transferred to another skill (e.g., production skill). While procedural knowledge is skill specific, declarative knowledge—deployed for retrieval of facts and rules—can be shared and used across skills. In other words, the theory yields the skill-specificity hypothesis, purporting that, if one skill is proceduralized and serves highly specific purposes (e.g., for production), it cannot be directly used for other purposes pertaining to different skills (e.g., comprehension). To date, only four empirical studies grounded in the skill acquisition theory have been conducted, providing evidence supporting the skill specificity of L2 practice in the domain of grammar, pragmatics, and pronunciation (DeKeyser, 1998; DeKeyser & Sokalski, 1996; Li & DeKeyser, 2017; Li & Taguchi, 2014).

DeKeyser and Sokalski (1996) conducted the first empirical study demonstrating the skill-specific effects of practice in L2 grammar learning. In their classroom study, Spanish L2 learners engaged in written input[2] practice (multiple-choice task) and output practice (fill-in-the-blank, translation) to acquire direct object clitic and conditional construction. After the classroom instruction on both structures and subsequent practice session, skill acquisition was measured by immediate and delayed posttests on comprehension and production. The overall findings supported the skill-specificity hypothesis that, compared to output practice, input practice led to better performance on the

comprehension test, while higher scores on the production test were associated with output practice. In a follow-up study, DeKeyser (1998) used an artificial language and provided written input and output practice of morphosyntactic structures. His study provided further evidence supporting the skill-specificity hypothesis.

More recently, Li and Taguchi (2014) examined the skill-specificity hypothesis in the domain of pragmatics (Chinese request-making forms). In the input-practice condition, L2 Chinese learners were instructed to read a series of request-making scenarios, after which they performed a judgment task. In the output-practice condition, the learners read the same scenarios as the ones in the input-practice condition, but subsequently performed a written translation and fill-in-the-blank task. Two outcome measures were employed (a receptive, listening judgment test and a productive, oral discourse completion test) and the resulting accuracy and speed measures (response time [RT]) were subjected to further analyses. The skill-specificity hypothesis was supported only for speed measures, not for accuracy measures. Discussing these divergent findings, Li and Taguchi (2014) argued that the two indices tapped into relatively different types of knowledge. In particular, in their view, accuracy scores primarily represented declarative knowledge, which was used for understanding the association among form, function, and context, whereas speed scores manifested procedural knowledge more strongly, as they indicate how quickly and efficiently pragmatic knowledge can be deployed.

Li and DeKeyser (2016) extended the inquiry into L2 pronunciation training. Their research focused on L2 Mandarin word tone learning, aiming to compare the effects of input training (sound-to-pinyin mapping and sound-to-meaning mapping) and output training (pinyin-to-sound mapping and meaning-to-sound mapping). The results indicate that more accurate and faster production performance was attained through output practice than input practice, while the significant advantage of input practice over output practice was observed in perception accuracy, not in perception speed. These results contrast those obtained by Li and Taguchi (2014), who noted that the effect of practice was more skill specific when learners' performance was assessed in terms of performance speed rather than accuracy. However, as Li and DeKeyser (2016) suspected, speed exhibited by their study participants might have reached a ceiling partly because their target skill was word learning. Hence, a difference in processing speed was less likely to emerge in this relatively simple word-learning task, compared to linguistic processing at the level of pragmatics targeted by Li and Taguchi (2014).

Available evidence lends support for skill-specificity hypothesis in different language domains (pronunciation, morphosyntax, and pragmatics) and in different L2 types (artificial language, Spanish, Chinese). Yet, pertinent research also suggests the presence of complex interactions among characteristics of target structures, skills, assessment measurements, and testing time (DeKeyser & Sokalski, 1996). Authors of previous studies in this domain primarily examined written practice of reading and writing skills (see preceding text). Hence, the generalizability of the findings through aural-oral practice format remains insufficiently explored. An exception is Li and DeKeyser's (2016) study on aural-oral modality, which nonetheless focused only on word-level pronunciation. This limitation is addressed in the current study, the aim of which is to extend the scope to the domain of syntax and examine the acquisition of English RC constructions through aural picture-matching and oral picture-description tasks.

PRACTICE SCHEDULE: BLOCKED, INTERLEAVED, AND HYBRID PRACTICE

In the cognitive psychology field, optimal practice scheduling has been extensively studied (Kang, 2016; Schmidt & Bjork, 1992; Toppino & Gerbier, 2014). This issue is particularly important for L2 instruction, if the findings yielded indicate that learners should focus on one element at a time (blocking) or should practice multiple elements simultaneously (interleaving). Cognitive psychologists have found that interleaved practice (e.g., ABCACBCAB) is more effective for learning and retention of knowledge and skills than blocked practice (e.g., AAABBBCCC). The advantage of interleaved practice has been documented in a variety of domains, such as motor skill acquisition (e.g., Hall, Domingues, & Cavazos, 1994; Shea & Morgan, 1979), category learning (e.g., Kang & Pashler, 2012; Kornell & Bjork, 2008), and mathematics (e.g., Rohrer & Taylor, 2007; Rohrer et al., 2014).

The superiority of interleaved relative to blocked practice is explained by the discriminative contrast hypothesis, which stipulates that interleaved practice facilitates discrimination among similar concepts and skills (Kang & Pashler, 2012; Kornell & Bjork, 2008). When learners encounter exemplars from different categories, they are more likely to attend to differences between these categories. For example, when L2 learners practice producing RC constructions, they may be able to better learn different types of RCs (e.g., *who*, *which*, *whom*). In addition to the discriminative advantage, the benefits of interleaved practice can also be explained by the distributed practice effect (i.e., enhancement of retention of learning through temporal spacing between exemplars compared to massing exemplars of the same category, Cepeda et al., 2006). Interleaved practice naturally introduces this spacing between exemplars from different categories. For example, when learning to use subject-relative pronouns *who* and *which*, blocked practice presents 20 exemplars of *who* followed by 20 exemplars of *which*, leading to no spacing between the exemplars of a given structure. In contrast, interleaved practice presents exemplars with *who* and *which* randomly, generating some spacing throughout the treatment. This means that interleaving usually corresponds to spaced learning, and blocking corresponds to massed learning. The distributed practice effect thus supports the benefits of interleaved practice (Kang, 2016).

Although the interleaving effects are replicated in a variety of learning materials and contexts (see Kang, 2016 for review), blocking has some advantage in certain contexts. In the study conducted by Carpenter and Mueller (2013), English-speaking college students were given pronunciation rules for French (e.g., *eau* is pronounced as a long *o* sound as in *cadeau* or *tableau*). In the blocked-practice condition, participants studied pronunciation rules by simply seeing and listening to words to which the same pronunciation rule applied in sequence (e.g., *bateau*, *fardeau*, *rameau*, . . . *tandis*, *brebis*, *vernis*, . . . *darder*, *combler*, *valser*). In the interleaved-practice condition, the words to which different rules applied were mixed (e.g., *bateau*, *tandis*, *darder*, *fardeau*, *brebis*, *combler*, *rameau*, *vernis*, *valser*). Challenging the extant research findings, Carpenter and Mueller found that blocking yielded better retention, measured by the receptive multiple-choice test, than interleaving. This observation may be in part explained by low similarity among practice items. Because the target pronunciation rules for specific words were very different (e.g., *eau*, *ch*, *s*, *t*), blocking helped learners to find

commonalities within each pronunciation category more easily than could be achieved by interleaving.

As demonstrated in the preceding text, both blocked and interleaved practice have their advantages; while blocking facilitates identifying the commonalities within each category, interleaving allows for discriminating similar features across different categories (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2013). These unique advantages of both practice schedules can be exploited further by combining them, which is a common classroom instruction strategy. For instance, L2 English learners first practice subject RCs and then object RCs in blocked schedule. When learners can use each structure with a certain level of confidence, they often engage in interleaved practice. This may also be explained by the desirable difficulty framework (Bjork, 1994). Specifically, blocking practice is likely to impose an appropriate level of difficulty (not too demanding) on learners in the early phase of learning, whereas interleaved practice may optimize the learning by providing more difficult learning conditions as the learners' skills improve (Porter et al., 2007; Porter & Magill, 2010; Wong et al., 2013; Yan et al., 2017).

Although in most prior psychology research purely interleaved and blocked practice schedules were compared, a few researchers have examined the effects of schedules that combined blocked and interleaved practice (i.e., hybrid practice). The target skill in prior research had some relevance for L2 learning in one study (speech motor learning, Wong et al., 2013), while other studies involve very different materials (artists' painting styles, Yan et al., 2017) and skills (basketball passes and golf putting, Porter et al., 2007; Porter & Magill, 2010) from L2 learning. Findings yielded by such investigations indicate that hybrid practice is as beneficial (Wong et al., 2013; Yan et al., 2017) or even more beneficial (Porter et al., 2007; Porter & Magill, 2010) than interleaved practice alone.

To our knowledge, the work by Nakata and Suzuki (2019) is the only empirical research conducted to date in which the effects of blocked, interleaved, and hybrid practice on L2 grammar acquisition were examined and compared. In their computerized experiment in the classroom, English-as-a-foreign-language learners studied five structures from the English tense-aspect-mood system (simple past, present perfect, first conditional, second conditional, and third conditional) using a written, multiple-choice fill-in-blank question format. They read a sentence where a verb phrase was omitted (e.g., *I _____ a car for my daughter last Christmas*) and were required to select the correct verb form among the four available options (e.g., *will buy*, *have bought*, *buy*, *bought*). The training materials consisted of 50 multiple-choice questions (10 questions × 5 constructions). Their acquisition of L2 knowledge was assessed by the grammaticality judgment test they took immediately upon training completion and one week after the treatment. Analysis of their test results showed that the interleaved practice led to significantly superior performance on the 1-week delayed posttest relative to the blocked practice. However, no significant differences were noted when comparing hybrid practice with blocked or interleaved practice. Although there are multiple factors that may account for these intriguing findings (e.g., a relatively high level of prior knowledge of the target structures), Nakata and Suzuki's (2019) study was an initial attempt to investigate the effects of three schedules on L2 grammar acquisition, thus leaving many questions open for further research. This was the motivation behind the current study, in

which the outcomes of hybrid practice relative to the interleaved and blocked modes were explored in more depth.

THE CURRENT STUDY

The goal of the current study was to advance our understanding of the complex interplay between practice format and schedule by merging these two isolated strands of research. The effects of practice format and practice schedule in the acquisition of RC constructions in L2 English by Japanese speakers were investigated. Participants performed either aural-input practice (picture-matching task) or oral-output practice (picture-description task) under one of the three (blocked, interleaved, and hybrid)[3] practice schedules. Assessment tasks were matched to the practice format (thus comprising of comprehension and production tests), and a detailed analysis of accuracy and speed was conducted to tap into relatively different dimensions of L2 declarative and procedural knowledge. Three research questions (RQs) were addressed:

1. How do different practice formats (input vs. output practice) contribute to the comprehension/production skill development?
2. How do different practice schedules (blocked vs. interleaved vs. hybrid) influence the comprehension/production skill development?
3. How do practice format and schedule interact and influence the comprehension/production skill development?

In answering the RQ1, the aim was to obtain further evidence of skill specificity in L2 practice. It was stipulated, based on the skill-specificity hypothesis, that input practice would lead to a greater progress in comprehension skill acquisition, whereas output practice would facilitate the development of production skills (DeKeyser, 1998; DeKeyser & Sokalski, 1996; Li & DeKeyser, 2017; Li & Taguchi, 2014). Furthermore, it was anticipated that the skill-specificity hypothesis would be supported more strongly for speed measures (Li & Taguchi, 2014).

Regarding RQ2, interleaved practice was predicted to be more effective than blocked practice. Because the target structures involved similar surface features as in subject and object RCs (e.g., *the boy who hugs the girl* vs. *the boy whom the girl hugs*), interleaved practice would highlight the subtle differences and facilitate learning to discriminate them more effectively than blocked practice would (Kang & Pashler, 2012; Kornell & Bjork, 2008). Furthermore, learners in the hybrid practice may take advantage of the benefits of both blocking and interleaving practice. According to the desirable difficulty framework, the level of learning difficulty should be desirable to maximize learning (Bjork, 1994). Blocking practice, due to the nature of the relatively lower cognitive demands, may be more suitable for learners in the early phases of learning, whereas interleaved practice may optimize learning by providing more difficult learning conditions in the later phases of learning (Porter et al., 2007; Porter & Magill, 2010; Wong et al., 2013; Yan et al., 2017). In sum, hybrid practice may be predicted to be more beneficial than either blocked or interleaved practice alone.

Answering RQ3 was of particular importance, as the interaction between practice format and schedule has never been examined within a single study. In this context, two possible

guiding questions, rather than predictions, emerged. First, we explored whether the skill specificity effect would vary depending on the chosen learning schedule. On the one hand, if three practice schedule conditions yielded the same level of learning (proceduralization), the amount of skill transfer would not differ across the three schedule conditions. If different schedules, on the other hand, resulted in different levels of proceduralization, the amount of skill transfer may differ among the three schedule conditions. For instance, suppose that proceduralization takes place for input practice under one schedule (e.g., interleaved practice) to a greater extent than noted under a different schedule (e.g., blocked practice). Because more fine-tuned proceduralization leads to less transfer, less transfer would be observed from comprehension skill to production skill in interleaved-practice condition than in blocked-practice condition. The second issue that had to be examined was whether the effects of different schedules would be dependent on practice format. For instance, because the acquisition of comprehension skills is easier than that of production in RC construction by L2 learners (Izumi, 2003), a more demanding schedule (i.e., interleaving) may be more beneficial in input practice for comprehension skills than a less demanding schedule (i.e., blocking). By contrast, for output practice aimed at productive skills, which are more difficult to acquire, hybrid practice may scaffold heavy-demand learning and lead to better learning outcomes than would blocked and interleaved practice.

METHOD

PARTICIPANTS

The study sample comprised of 155 Japanese speakers who were studying English in seven English classes held at two Japanese universities. Prior to the experiment, students were randomly assigned to six groups characterized by different practice formats (input or output) and practice schedules (blocked, interleaved, or hybrid). The data pertaining to 26 participants were subsequently excluded, as they indicated that they had studied the target grammatical structures outside the experiment between the immediate and delayed posttest (see "Procedure"). Consequently, the data submitted for analysis related to the remaining 129 participants (male = 56, female = 51) in different academic years: first year (11), second year (98), third year (16), and fourth year (4). They formed input-blocked ($n = 22$), input-interleaved ($n = 18$), input-hybrid ($n = 31$), output-blocked ($n = 18$), output-interleaved ($n = 19$), and output-hybrid ($n = 21$) groups.

TARGET STRUCTURE

Target syntactic structures in this study were relative clause (RC) constructions. RCs were chosen as a target structure because Japanese learners have difficulty in fully mastering these structures, although they are typically taught explicitly during junior and high school (Mochizuki & Ortega, 2008). Classroom instruction and practice, however, has been found to influence the acquisition of RC constructions (Doughty, 1991). Thus, the training session adopted in the present study was designed to facilitate the acquisition of learners' knowledge of the following RCs:

(a) Subjective RC who (e.g., That is the girl who is washing the bird.)

(b) Subject RC which (e.g., That is the cat which is watching the bird.)
(c) Objective RC whom (e.g., That is the girl whom the cat is watching.)
(d) Object RC which (e.g., That is the dog which the woman is carrying.)

INSTRUMENTS

Training Materials

Because RCs can be avoided in L2 production, particularly by Japanese speakers (Schachter, 1974), controlled, form-focused practice format was employed for treatment. Controlled practice alone is insufficient for L2 acquisition; however, it can be an efficient technique for developing declarative and procedural knowledge about specific target structures (Ellis & Shintani, 2014).

Participants in the output-practice condition orally described the pictures on the computer screen using appropriate relative pronouns. As shown in the left panel in Figure 1, they were first presented with a prompt, which was accompanied by the first part of the sentence indicated at the top (e.g., *That is the boy…*) and lexical items necessary for oral description (i.e., boy, dog, kiss) to help participants concentrate on practicing using RC structures. The participants were given 12 seconds to describe the picture (e.g., *That is the boy who is kissing the dog*). A correct answer was provided both visually and aurally and it remained on the screen for 8 seconds.

Participants in the input-practice condition listened to a sentence once and were required to select the matching picture. As shown in the right panel in Figure 1, they were first presented with the two pictures side by side and listened to an audio sentence (e.g., *That is the boy who is kissing the dog*). They were given 12 seconds to choose the matching picture by pressing a corresponding button. Once the response was given, learners were presented with the screen accompanied by the correct answer, which remained displayed for 8 seconds.

The treatment comprised of 64 stimulus sentences, which were used for both input- and output-practice conditions. Sixteen sentences, each pertaining to four structures, contained one of the eight verbs (i.e., carry, hug, kick, kiss, massage, push, wash, watch) with human or nonhuman nouns as an antecedent. All verbs were familiar to participants because they are taught in junior high school and/or are loan words in Japanese (see the list of stimulus sentences in Appendix A in the online supplementary file).

Outcome Tests

Production and comprehension tests were designed to assess the development of accuracy and speed in the use of both skills. The tests were identical in format to the training material; however, participants were not provided any feedback. In the production test, the participants were required to describe the picture within 12 seconds. In the comprehension test, the participants listened to a sentence once and were given 12 seconds to select the matching picture.

Each test consisted of 16 items (see Appendix B in the online supplementary file). Four items were created for each of the four target structures (subject RC *who*, subject RC

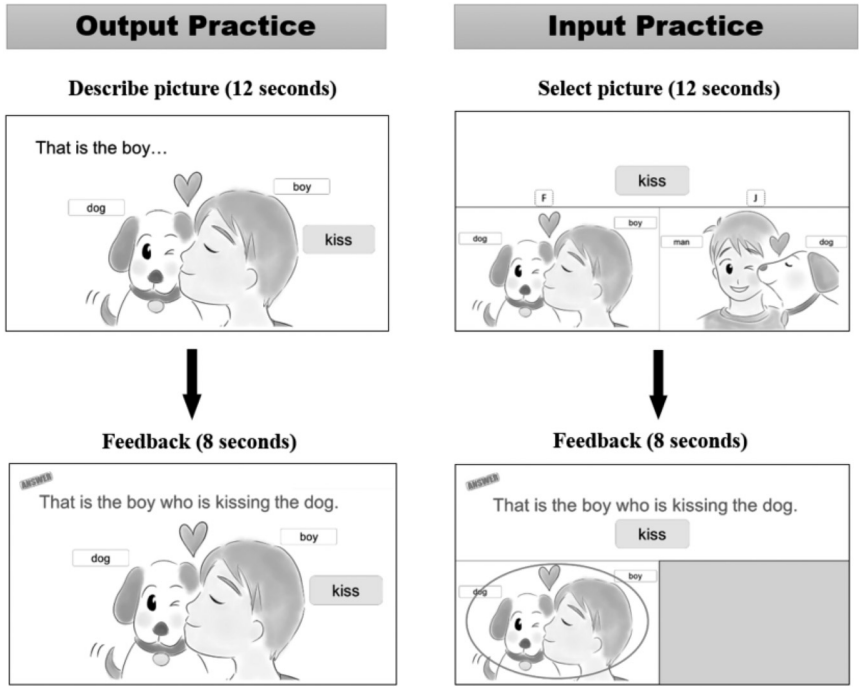

FIGURE 1. (color online) A sample display of output and input practice.

which, object RC *whom*, object RC *which*). The eight verbs used in the training session were also employed in the tests. To reduce the practice effect among the pretest and two posttests, three equivalent test forms were created for comprehension and production tests in which the same action verbs were used, but the combinations of action doer and the recipient were varied.

TRAINING SCHEDULE

Participants engaged in systematic input or production practice under one of the three practice schedules (blocked, interleaved, or hybrid). Figure 2 presents the sequence of practice instances for the three schedules. In the blocked-practice condition, all instances of each grammatical category were studied and sequenced as a block. Specifically, the participants encountered sentences using a subject relative pronoun *who* 16 times, a subject relative *which* 16 times, an object relative *whom* 16 times, and then an object relative *which* 16 times. In the interleaved-practice condition, instances from the four grammatical categories were intermixed and were presented in a randomized order. No items from the same category were presented twice in a row. In the hybrid-practice condition, the participants were presented with the first half of the items (32 items) in blocked practice format, followed by the remaining 32 items through interleaved schedule.

Blocked practice

SR-who	SR-who	SR-who	SR-who	SR-who	SR-who	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	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-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	OR-which	OR-which	OR-which	OR-which	OR-which	OR-which

Interleaved practice

SR-who	SR-which	OR-whom	OR-which	OR-whom	SR-which	OR-which	SR-who	SR-which	OR-whom	SR-who	OR-which	OR-whom	SR-which	SR-who	OR-which
SR-who	SR-which	OR-which	OR-whom	SR-which	SR-who	OR-which	OR-whom	SR-who	OR-which	SR-which	SR-who	OR-whom	SR-which	SR-who	OR-whom	OR-which
SR-who	SR-which	OR-whom	OR-which	SR-who	OR-which	OR-whom	SR-which	OR-which	SR-which	SR-who	OR-whom	SR-which	OR-whom	OR-which	SR-who
SR-which	OR-whom	OR-which	SR-who	OR-whom	SR-which	SR-who	OR-which	OR-whom	SR-which	SR-who	OR-which	SR-who	SR-which	OR-whom	OR-which

Hybrid practice

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
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
SR-who	SR-which	OR-whom	OR-which	SR-who	OR-which	OR-whom	SR-which	OR-which	SR-which	SR-who	OR-whom	SR-which	OR-whom	OR-which	SR-who
SR-which	OR-whom	OR-which	SR-who	OR-whom	SR-which	SR-who	OR-which	OR-whom	SR-which	SR-who	OR-which	SR-who	SR-which	OR-whom	OR-which

FIGURE 2. Practice schedule.
Note. SR = Subject RC, OR = Object RC.

PROCEDURES

The study was conducted during two regular class hours. In Session 1, participants performed a pretest (5 minutes), a training task (20 minutes), and an immediate posttest (5 minutes) individually in a computer room. In Session 2, which took place a week after the first session, the delayed posttest was administrated to assess participants' retention of acquired grammatical knowledge. In each test session, the production test, which provides no input regarding RC constructions, was administered first, followed by the comprehension test, to minimize the influence of the first test on the outcome of the second. All training and test materials were administered through the DMDX software program (Forster & Forster, 2003).

DATA CODING

Comprehension Test

Accuracy on the comprehension tests was scored as 1 or 0. For the speed measure, response time (RT) was measured from the onset of the prompt to the button press. Only RTs of correct test items were submitted for analysis. In addition, participants' data was included in the speed analysis only if their mean accuracy rates were at least 65% (15% above the level that could be attained by chance) on the pretest and two posttests. This stringent screening procedure was adopted to retain sufficient test items for analyzing RTs reliably for each participant (see, for instance, Hulstijn, Van Gelderen, & Schoonen, 2009 for a similar approach). Setting the cutoff value was necessary to prevent the underrepresentation of learners' procedural knowledge (from lower cutoff values) due to the low accuracy of comprehension and production (e.g., lack of declarative knowledge). Furthermore, outlying responses were identified and treated as missing values. The lower cutoff was set to a point where a relative pronoun was pronounced in each stimulus sentence (this resulted in 0.7%, 1.3%, and 0.9% of the data being excluded at pretest,

immediate posttest, and delayed posttest, respectively), whereas the higher cutoff was set at 2 SD or higher than the group mean for each test item (3.8%, 4%, and 4.8%, were thus excluded from pretest, immediate posttest, and delayed posttest data sets, respectively). The internal consistency indexed by Cronbach's alpha was sufficient (.70−.90) for all comprehension measures (see Appendix C in the online supplementary file), with the exception of the accuracy measure of the pretest score (alpha = .58), probably due to the lower accuracy rates prior to the treatment session.

Production Test

Two raters first coded the outcome tests for accuracy and speed using the sound analysis software Praat (Boersma & Weenink, 2016). Both individuals coded a subset of data independently (12.4% of the production tests) and discussed inconsistencies until their coding matched. After that, the same raters coded the remaining data independently.

Accuracy was scored as 1 or 0 for each test item. When assigning scores, if the word order and relative pronoun were used correctly, a credit was given to utterances with incorrect (non-)use of articles (e.g., *That is boy who is kissing dog*) and/or wrong tense and aspect (e.g., *That is the boy who kiss the dog*). For the speed measures, RT was measured from the onset of the prompt to the end of the utterance. RT data was excluded if (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*), (c) a content word different from the specified word was used (e.g., using "man" instead of "grandfather"), and/or (d) participants did not follow the instructions (e.g., not uttering the first phrase "That is . . . "). As in comprehension tests, to retain sufficient test items for analysis, RT was computed for students whose accuracy rate was ≥ 50% on the immediate and delayed posttests to retain sufficient test items for RT analysis. The chosen cutoff accuracy rate was lower than that in comprehension tests, as in this case, participants could not respond correctly by chance (see Hulstijn et al., 2009 for a similar approach). Once again, outlying responses were identified and treated as missing values, with 2 SD below the group mean serving as the lower cutoff (Pretest: 0%; Immediate posttest: 0.04%; Delayed posttest: 0.1%) and 2 SD above the group mean serving as the lower cutoff (Pretest: 0%; Immediate posttest: 1.2%; Delayed posttest: 0.6%) for each test item. The internal consistency indexed by Cronbach's alpha was sufficient (.70−.92) for all production measures (see Appendix C in the online supplementary file). Due to the less constrained nature of production tests than comprehension tests, the aforementioned RT cleaning procedure resulted in a very small data set suitable for production test analysis (see "Results" section). Given the small number of participants in each group, the results based on speed analysis should be interpreted with caution and regarded as a supplementary analysis to accuracy measures.

STATISTICAL ANALYSIS

To examine the effects of practice and schedule, the accuracy scores achieved on the comprehension and production tests were analyzed separately for immediate and delayed posttests using a logistic mixed-effects model (mixed logit model). For the speed measures, a linear mixed-effects model was used to analyze the RT on the two

comprehension and production posttests. The only difference between logistic and linear mixed-effects models pertained to the dependent variable, with the former employing binary accuracy response (correct/incorrect) and the latter using RT data (Jaeger, 2008). Both models were implemented through the lme4 software package in R (Bates, Mächler, Bolker, & Walker, 2014). The fixed effects were practice (input vs. output) and schedule (blocked vs. interleaved vs. hybrid). The fixed effect of practice was centered using deviation coding, while the effect of schedule was dummy coded with blocked practice as a reference group (Linck & Cunnings, 2015). Accuracy scores on the pretest, which were scaled (standardized) to reduce collinearity, were included as a covariate in both models. For speed measures, RT on the pretest (standardized) as well as accuracy scores were included as covariates. Note that the pretest RTs were available for the comprehension tests only because production pretest RT data of only few participants were retained for analysis (see Table 2).

All the mixed-effects models were incrementally developed using a maximum likelihood technique with forward-selection procedure (Cunnings, 2012; see, e.g., Yi, 2018 for a similar approach). First, the initial model started with random-intercept-only models, with practice and schedule as fixed effects and pretest as a covariate. Second, interactions among the two fixed effects and one covariate were added to the initial model, followed by random slopes. These models with additional interaction term(s) were compared with the initial model, using the *anova* function in the lme4 package (Bates et al., 2014). The forward-model selection procedure was conducted with alpha levels of .05, and the best-fitting models are discussed in the "Results" section.[4] Post-hoc comparisons, whenever necessary, were conducted using the R package, lsmeans (Lenth, 2016).

RESULTS

ACCURACY MEASURES

Comprehension Test

Table 1 presents means, SDs and 95% confidence intervals (CIs) of comprehension test accuracy rates in all six conditions. Mean accuracy rates on the pretest ranged from 62.22% to 77.08%, indicating that learners' performance was above a mere 50% chance level (see Appendix D in the online supplementary material). After the treatment, accuracy rates increased to well above 80% (range: 84.21−92.26%) across groups. As predicted by the skill-specificity hypothesis, the average gains (calculated as means of averages pertaining to all three schedule groups) from the pretest to the immediate posttest in the input-practice condition (20.34%) appeared to be greater than those in the output-practice condition (13.55%). In contrast, the score gains from the pretest to the immediate posttest among the three schedule conditions seemed very similar (blocked = 22.73%; interleaved = 18.75%, hybrid = 19.56%). A consistent pattern in the results was also noted in the average gains from the pretest to the delayed posttest.

Although the input-practice condition yielded higher accuracy rates than the output-practice condition at the descriptive level, the logit model results for comprehension tests showed no significant fixed effects of practice for either immediate or delayed posttest ($ps > .10$, see Appendix E). Similarly, the fixed effect of schedule was not significant on either posttest ($ps > .10$).

TABLE 1. Descriptive statistics for the accuracy measures

		Comprehension Test				Production Test			
				95% CI				95% CI	
		M	*SD*	Lower	Upper	*M*	*SD*	Lower	Upper
Input_Blocked (n = 22)	Pretest	62.22	16.65	55.40	69.03	23.86	16.66	17.33	30.68
	Immediate Post	84.94	15.15	78.42	90.63	73.58	26.51	62.22	84.38
	Delayed Post	81.82	20.94	73.02	89.77	58.81	29.80	46.31	70.45
	Mean Difference (Immediate Post – Pre)	22.73				49.72			
	Mean Difference (Delayed Post – Pre)	19.60				34.94			
Input_Interleaved (n = 18)	Pretest	71.88	16.78	63.54	79.51	29.51	19.15	20.84	38.19
	Immediate Post	90.63	10.56	85.76	94.79	70.14	27.50	55.90	81.60
	Delayed Post	93.75	6.78	90.63	96.53	61.46	26.36	48.96	73.60
	Mean Difference (Immediate Post – Pre)	18.75				40.63			
	Mean Difference (Delayed Post – Pre)	21.88				31.94			
Input_Hybrid (n = 31)	Pretest	68.35	14.25	63.51	73.39	28.43	17.52	22.18	34.07
	Immediate Post	87.90	13.59	82.67	92.54	65.12	25.30	56.25	73.58
	Delayed Post	88.71	12.95	83.87	93.14	64.92	26.65	55.44	73.18
	Mean Difference (Immediate Post – Pre)	19.56				36.69			
	Mean Difference (Delayed Post – Pre)	20.36				36.49			
Output_Blocked (n = 18)	Pretest	72.22	14.42	65.28	78.13	26.39	13.31	21.18	32.99
	Immediate Post	87.15	11.64	81.60	92.01	80.56	18.92	72.22	88.19
	Delayed Post	90.63	11.19	85.07	95.14	68.06	23.76	57.99	78.13
	Mean Difference (Immediate Post – Pre)	14.93				54.17			
	Mean Difference (Delayed Post – Pre)	18.40				41.67			
Output_Interleaved (n = 19)	Pretest	73.68	16.87	65.46	81.25	29.61	13.31	24.34	35.53
	Immediate Post	84.21	16.97	76.32	91.45	89.80	13.05	83.55	95.07
	Delayed Post	90.79	10.90	85.53	95.39	73.36	17.90	65.46	81.58
	Mean Difference (Immediate Post – Pre)	10.53				60.20			
	Mean Difference (Delayed Post – Pre)	17.11				43.75			

(continued on following page)

TABLE 1. Descriptive statistics for the accuracy measures (continued)

		Comprehension Test				Production Test			
				95% CI				95% CI	
		M	*SD*	Lower	Upper	*M*	*SD*	Lower	Upper
Output_Hybrid (n = 21)	Pretest	77.08	14.02	71.13	83.03	25.00	15.44	18.75	31.25
	Immediate Post	92.26	11.68	86.61	96.73	89.29	17.91	80.95	95.24
	Delayed Post	89.29	12.21	83.33	94.05	69.05	29.41	55.96	80.65
	Mean Difference (Immediate Post – Pre)	15.18				64.29			
	Mean Difference (Delayed Post – Pre)	12.20				44.05			

Production Test

Production accuracy rates on the pretest were below 30% across all groups (see Table 1 and Appendix D). All three output-practice groups achieved 80% on the immediate posttest, whereas the three input-practice groups scored around 70%. As predicted by the skill-specificity hypothesis, greater mean gains (calculated as means of averages pertaining to all three schedule groups) were observed in the output-practice condition (59.55%) than in the input-practice condition (42.34%). The mean gains from the pretest to the delayed posttest seemed to have remained consistently higher for the output-practice condition (43.15%) than for the input-practice condition (34.46%).

When comparing the score gains across the three schedule groups, an interesting asymmetry appeared in the immediate posttest. In the output-practice condition, the descending order of the three schedule conditions in terms of score improvements was blocked practice (54.17%), interleaved practice (60.20%), and hybrid practice (64.29%). For the input-practice condition, the score gains were reversed, as hybrid practice (36.69%) was followed by interleaved practice (40.63%) and finally blocked practice (49.72%). On the delayed posttest, negligible differences were noted across the three schedule conditions.

The results of logit models for production tests are presented in Appendix E. For the immediate posttest, the effects of schedule (hybrid practice) was significant in comparison to the reference group (blocked practice), $z = 2.13$, $p = .03$. The other two comparisons did not yield significant differences: interleaved versus blocked ($z = 1.63$, $p = .10$) and interleaved versus hybrid ($z = -0.44$, $p = .66$). While the fixed effect of practice was not significant, $z = -0.56$, $p = .58$, a significant interaction was found between practice and schedule (interleaved and hybrid), $z = -1.98, -3.05$, $p = .002$. These significant interactions are illustrated in Figure 3, revealing virtually no difference between the effects of input and output practice under the blocked-practice condition.

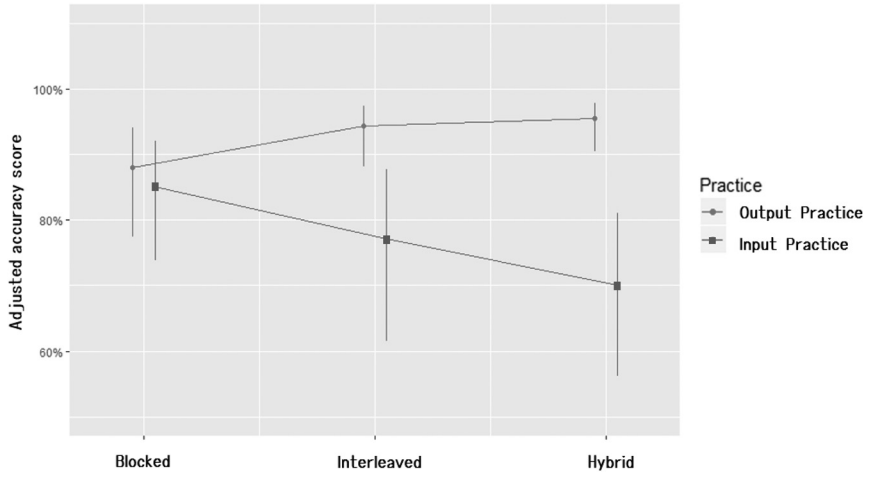

FIGURE 3. (color online) Significant interaction between practice and schedule on accuracy scores of production test (immediate posttest).
Note. The accuracy scores were adjusted for the pretest scores. The error bars indicate 95% confidence intervals.

Conversely, output practice was more effective than input practice in the interleaved- and hybrid-practice conditions. The advantage of output practice was more pronounced in the hybrid-practice condition ($d = 0.93$, 95% CI [0.23, 1.59]) than in the interleaved-practice condition ($d = 1.24$, 95% CI [0.62, 1.83]).

On the delayed posttest, the fixed effects of schedule or significant interactions in the immediate posttest were no longer significant. However, the fixed effect of practice remained significant, $z = -2.10$, $p = .04$, indicating that the production condition yielded better performance than the comprehension condition ($M = 43.15\%$ vs. 34.46%).

SPEED MEASURES

Comprehension Test

Table 2 presents means, SDs and 95% CIs of the comprehension test speed measures in all six groups. Mean RTs on the pretest were around 4,800 ± 300 ms (see Appendix D). On the immediate posttest, the mean RT decrease was greater in the comprehension practice condition (1,093 ms) than in the production condition (521 ms). In the input-practice condition, interleaved and hybrid practice (1,423 and 1,291 ms) appeared to have improved more than blocked practice (564 ms). In the output-practice condition, RTs in the hybrid-practice group improved the most (815 ms); the blocked and interleaved practice groups showed smaller, similar levels of practice effects (344 ms and 404 ms). A consistent pattern in the results was also noted in the improvement from the pretest to the delayed posttest.

The results of linear mixed-effects models for comprehension tests are presented in Appendix F. In the immediate posttest, the effect of practice was significant ($t = -4.24$, $p < .001$), supporting the hypothesis that input practice improved RT more than output practice did. The fixed effect of schedule was also significant in the comparison between hybrid practice and blocked practice ($t = -2.43$, $p = .02$), while no significant difference was found between interleaved practice and blocked practice ($t = -0.58$, $p = .57$). The advantage of hybrid practice relative to interleaved practice was marginally significant ($t = 1.87$, $p = .07$). Furthermore, a significant three-way interaction was detected among practice (input practice), schedule (interleaved), and pretest RT ($t = -2.43$, $p = .02$). This interaction was further examined by applying a linear mixed-effects model with schedule (blocked vs. interleaved practice) as a fixed effect, focusing solely on input-practice condition (see Appendix G in the online supplementary material). The model yielded a significant interaction between schedule and pretest RT ($t = -2.92$, $p = .01$). This interaction is illustrated in Figure 4, where it can be seen that the participants with greater RT (slower processing speed) on the pretest tended to perform better in interleaved practice compared to blocked practice. Conversely, only marginal differences between two practice schedules were noted among learners with shorter RT (faster processing speed).

The same pattern was observed in the delayed posttest results. Identical significant fixed effects and interaction term were obtained for the immediate and delayed posttests (see Appendix G in the online supplementary file). Similarly to the immediate posttest, hybrid practice was more beneficial in enhancing processing speed than interleaved practice ($t = 2.48$, $p = .02$). The significant three-way interaction among practice (input practice), schedule (interleaved), and pretest RT for the delayed posttest was, once again, examined further. As shown in Figure 4, a very similar interaction pattern was found for

TABLE 2. Descriptive statistics for the reaction time measures

		Comprehension Test					Production Test			
					95% CI				95% CI	
		n	*M*	*SD*	Lower	Upper	*n*	*M*	*SD*	Lower	Upper
Input_Blocked	Pretest	9	4430	753	3948	4894	2	–	–	–	–
	Immediate post	9	3866	665	3490	4327	7	6844	1123	6101	7615
	Delayed post	9	3684	504	3390	4002	7	6017	759	5488	6511
	Mean Difference (Pre – Immediate Post)		564								
	Mean Difference (Pre – Delayed Post)		746								
Input_Interleaved	Pretest	13	5162	1140	4653	5819	3	–	–	–	–
	Immediate post	13	3740	458	3511	3992	7	6959	1778	5755	8118
	Delayed post	13	3927	416	3724	4136	6	6368	1339	5393	7343
	Mean Difference (Pre – Immediate Post)		1423								
	Mean Difference (Pre – Delayed Post)		1235								
Input_Hybrid	Pretest	14	4887	1000	4396	5392	3	–	–	–	–
	Immediate post	14	3596	778	3210	4014	15	6484	962	6018	6951
	Delayed post	14	3585	673	3262	3966	13	6226	969	5707	6730
	Mean Difference (Pre – Immediate Post)		1291								
	Mean Difference (Pre – Delayed Post)		1303								
Output_Blocked	Pretest	12	4724	802	4314	5207	1	–	–	–	–
	Immediate post	12	4381	626	4022	4704	9	6350	1126	5709	7109
	Delayed post	12	4422	957	3944	4931	9	6525	670	6136	6942
	Mean Difference (Pre – Immediate Post)		344								
	Mean Difference (Pre – Delayed Post)		303								
Output_Interleaved	Pretest	10	4540	754	4094	4995	2	–	–	–	–
	Immediate post	10	4136	909	3617	4683	14	5738	962	5237	6221
	Delayed post	10	4086	883	3593	4601	17	6204	1037	5742	6675
	Mean Difference (Pre – Immediate P)		404								
	Mean Difference (Pre – Delayed P)		454								

(continued on following page)

TABLE 2. Descriptive statistics for the reaction time measures (continued)

		Comprehension Test					Production Test			
					95% CI				95% CI	
		n	*M*	*SD*	Lower	Upper	*n*	*M*	*SD*	Lower	Upper
Output_Hybrid	Pretest	16	4812	894	4423	5259	3	–	–	–	–
	Immediate post	16	3997	825	3628	4386	12	5502	853	5092	5976
	Delayed post	16	3921	835	3529	4350	13	6241	1025	5747	6792
	Mean Difference (Pre – Immediate P)		815								
	Mean Difference (Pre – Delayed P)		891								

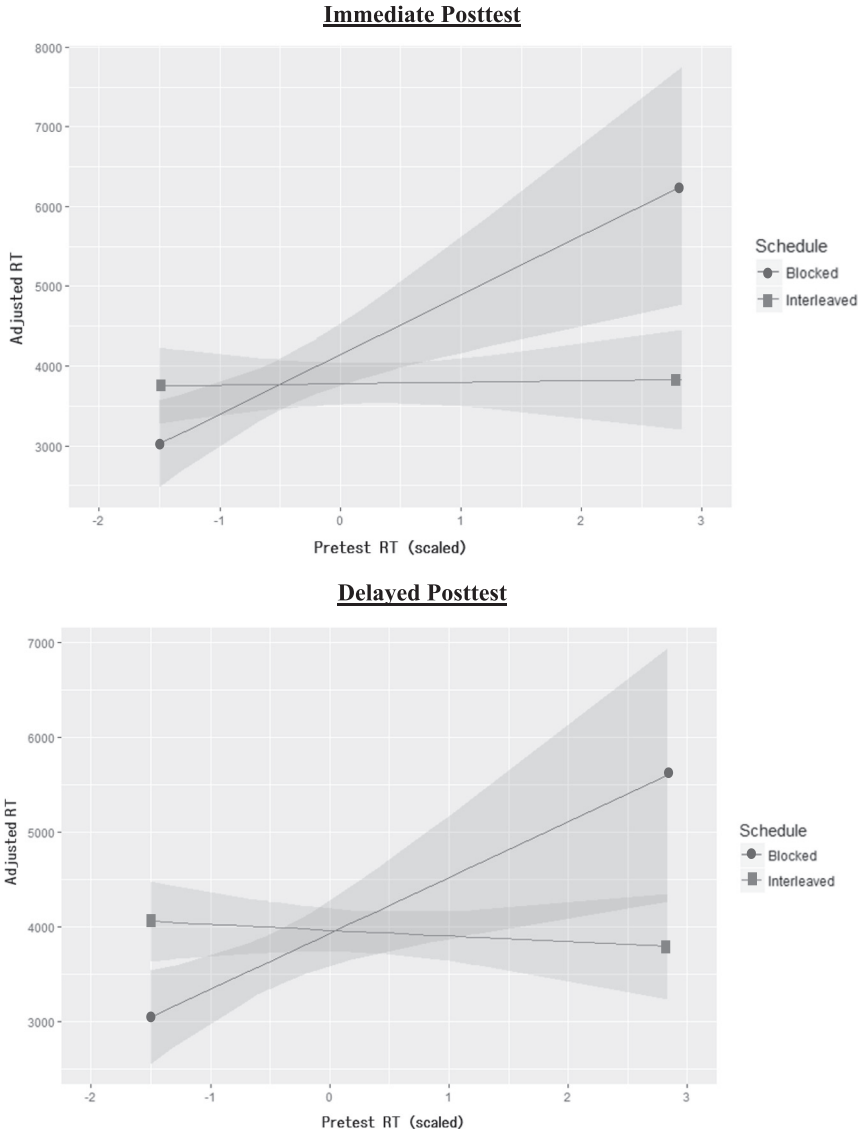

FIGURE 4. (color online) Significant interaction between schedule (blocked vs. interleaved practice) and pretest RT in comprehension test performance by input-practice groups.
Note. The shaded areas indicate 95% confidence intervals.

the delayed posttest, suggesting that the participants with greater RT on the pretest were more likely to benefit less from blocked practice than from interleaved practice.

Production Test

As shown in Table 2, the mean RTs (calculated as means of averages pertaining to all three schedule groups) on the immediate posttest were shorter in the output-practice condition

(5,863 ms) than in the input-practice condition (6,763 ms). In the output-practice condition, hybrid practice (5,502 ms) appear to have resulted in the shortest RT, followed by interleaved (5,738 ms) and blocked practice (6,350 ms). Similarly, in the input-practice condition, hybrid practice (6,484 ms) exhibited shorter RT than blocked and interleaved practice (6,844 and 6,959 ms). In the delayed posttest, the six groups performed similarly in terms of RT, with differences not exceeding 500 ms (6,017–6,525 ms).

The results of linear mixed-effects models for production tests are presented in Appendix F. For the immediate posttest, practice was a significant fixed effect ($t = 3.78$, $p < .001$), suggesting that output practice was more beneficial than input practice for production test performance. The effect of schedule was marginally significant ($t = -1.85$, $p = .07$), whereby learners in the hybrid-practice group outperformed those assigned to the blocked-practice condition. On the delayed posttest, none of the fixed effects were significant ($ps > .10$).

DISCUSSION

SKILL-SPECIFICITY HYPOTHESIS

The first research question pertained to the skill specificity of practice. The results reported in this work indicate presence of skill-specificity effects for both practice modalities (input and output practice) in the receptive and productive use of English RC constructions. These findings corroborate the previously observed skill-specific effects of L2 practice (DeKeyser, 1998; DeKeyser & Sokalski, 1996; Li & DeKeyser, 2017; Li & Taguchi, 2014) and further demonstrate that they extend to aural-oral practice in L2 acquisition of syntax. The skill specificity effects were systematically observed particularly for the speed measures of the comprehension tests and the accuracy measures of the production tests.

The accuracy rates achieved in the comprehension tests did not provide any evidence of skill-specificity effects. This finding may be due to the nature of the output training task in the current study. In particular, the correct answer in a form of visual and aural input was provided (i.e., serving as partial input practice) throughout the output-practice treatment, which might have led to the nonsignificant effect of practice type on the comprehension accuracy score. Another explanation may pertain to the relatively higher prior level of comprehension skills than that of production skills in this group of L2 learners, which may also be the case for many Japanese learners in general (e.g., Izumi, 2003). In contrast, the processing speed of learners that took part in this study was more variable and was thus more likely to benefit from practice, which might have contributed to a clearer pattern in the specificity effects. Theoretically, greater proceduralized (as opposed to declarative) knowledge becomes more fine-tuned when specific skills are practiced, leading to less transfer from output practice to comprehension skill (Anderson, 1993). This theoretical prediction may be supported by the findings reported by Li and Taguchi (2014). These authors found skill-specific practice effects only on the speed measures (i.e., how quickly grammatical knowledge can be used), which represent procedural knowledge more strongly than do accuracy measures. The results yielded by current study corroborate Li and Taguchi's (2014) findings. Procedural grammatical knowledge is skill specific; it is thus reasonable to assume that the skill-specific effects of input practice manifested most strongly in the speed dimension of comprehension skills.

In the production tests, by contrast, skill-specificity effects were found for the accuracy measures on both posttests, as well as for the speed measure of the immediate posttest. It appears somewhat inconsistent that the effect was found particularly for accuracy measures, given that response speed is a purer indicator of procedural knowledge. However, accuracy scores also reflect procedural knowledge (as well as declarative knowledge) because production test performance involves integration of multiple skills (e.g., lexical and grammatical processing, articulation) for uttering a sentence under moderate time pressure (DeKeyser, 2015; Kormos, 2006). Accuracy measures may better reflect earlier stages of proceduralization than speed measures. Of course, behavioral tests cannot completely distinguish procedural from declarative knowledge. Consequently, skill-specificity effects are most likely to manifest in the acquisition stage of procedural knowledge, which is assessed by some indicators of some tests, depending on the learners' L2 proficiency level (Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter, & Wong, 2014). The learners that took part in the present study were probably still in the earlier stages of productive skill proceduralization, compared to comprehension skills. Hence, they might have acquired accurate productive skills more efficiently from output practice than input practice.

PRACTICE SCHEDULE EFFECTS

The second research question probed into the relative effectiveness of three practice schedules. The significant effects of practice schedule emerged in the same subskills (i.e., comprehension speed and production accuracy) as those found for the effects of practice format. These findings indicate that the hybrid-practice group significantly outperformed the blocked- and interleaved-practice groups in terms of comprehension speed (both posttests) and production accuracy (immediate posttest). This observation is consistent with some of prior research findings, suggesting that combining blocked and interleaved practice is more beneficial than interleaved practice alone (Porter et al., 2007; Porter & Magill, 2010). It is noteworthy, however, that the present finding diverges from those obtained in the Nakata and Suzuki's (2019) prior research, where interleaved practice yielded significantly better performance than hybrid or blocked practice.

The advantage of hybrid practice may be explained by drawing on the desirable difficulty framework (Bjork, 1994), which postulates that knowledge/skill acquisition is most enhanced when learners engage in the tasks with the appropriate level of difficulty. When task difficulty is relatively high (e.g., early in the training), the least demanding, blocked-practice schedule may scaffold learning, thus optimizing the level of learning difficulty relative to the learners' initial skill level. As learners' skills improve through blocked practice, a more demanding, interleaved-practice schedule can then offer an ideal learning condition, where learners' skill level matches task demands. The learning trajectory during the treatment may lend some support for this interpretation (see Appendix H for performance during the treatment). In the hybrid-practice schedule, the earlier blocked practice efficiently improved production accuracy and comprehension speed, and then the latter interleaved practice introduced difficulty in performance (indicated by lower production accuracy and slower comprehension speed). In sum, the hybrid schedule gradually increased task demands from less taxing, blocked practice to

more demanding, interleaved practice, thus challenging learners optimally throughout the training session (Guadagnoli & Lee, 2004; Suzuki et al., 2019b).

In the current study, the training/outcome tasks (aural picture-matching and oral picture-description tasks) were probably more difficult than those employed by Nakata and Suzuki (2019) (a written fill-in-the-blank task/a written GJT that provided accuracy scores only). First, the speed measures in the current comprehension tests tap into more procedural knowledge, which requires more efficient skill to successfully perform the task. Second, learners need to execute multiple cognitive processes simultaneously for production (e.g., lexical retrieval, grammatical processing, and articulation) under moderate time pressure. These factors might have contributed to the apparent higher training task difficulty. The gradual scaffolding throughout the training phase (hybrid schedule) might have optimized the difficulty levels of consecutive tasks for the learners that took part in this investigation. In contrast, when task difficulty is relatively lower, as was the case in the Nakata and Suzuki's (2019) prior study, the learners might be able to take advantage of the effects of interleaved practice early in the training, rendering this practice schedule more advantageous. In contrast, the learners that took part in the present study were less likely to benefit from the effects of interleaved practice early in the training process due to the difficulty surpassing learners' skill level. Thus, its diminished effects might have led to a nonsignificant difference between interleaved and blocked practice formats.

DYNAMIC INTERPLAY AMONG PRACTICE TYPE, SCHEDULE, AND PRIOR KNOWLEDGE

The results reported in this work reveal two significant interactions pertaining to comprehension and production test, respectively. In the comprehension test, a significant two-way interaction between schedule (blocked vs. interleaved) and pretest RT was revealed in the input-practice condition. This pattern was consistently found on both immediate and delayed posttests (speed performance), which is characterized as aptitude-treatment interaction[5] in a broad sense (Cronbach & Snow, 1977). It suggests that, while interleaved practice may neutralize the effects of prior knowledge, blocked practice is susceptible to learners' prior processing speed. Critically, this aptitude-treatment interaction (depicted in Figure 4) is primarily driven by the data related to less skilled learners (i.e., those whose processing speed was slower). These less skilled learners might have been pushed to discriminate subject and object RC constructions from similar instances presented randomly, that is, interleaved presentation (Carvalho & Goldstone, 2014; Zulkiply & Burt, 2013). However, they had difficulty in learning to distinguish similar grammatical categories in blocked-practice schedule, possibly because their processing skill was insufficient. The more skilled learners, in contrast, may have been able to utilize their more efficient processing abilities to accelerate discrimination of similar syntactic structures regardless of practice schedules (Sana, Yan, & Kim, 2017).

An intriguing aspect of dynamic interplay between practice and schedule was found for the accuracy measures on the immediate production posttest. While no significant difference was found between input-practice and output-practice groups under the blocked-practice condition, the output-practice groups outperformed the input-practice groups under both interleaved- and hybrid-practice conditions. As illustrated in Figure 3,

this interaction is characterized by the declining performance in the input-practice condition. Specifically, the amount of transfer from the receptive to productive skill seems to be decreasing from blocked to interleaved to hybrid practice. This pattern in production test performance is reversed in comprehension test performance. Recall that, in the comprehension test (speed measure), hybrid practice was the optimal practice schedule, followed by the interleaved practice and the blocked practice, as the least effective method of accelerating processing comprehension speed. In other words, the greatest proceduralization of comprehension skills presumably occurred in the hybrid-practice condition, followed by interleaved-practice and blocked-practice condition. The greater proceduralization of comprehension skills under the hybrid-practice condition may make it difficult for learners to transfer their comprehension skills to production skills, however. In contrast, learners in the blocked-practice condition, who exhibited the least proceduralized knowledge, may be able to perform relatively better in production test (greater transfer) possibly because they would rely more on declarative knowledge and less on procedural knowledge (Bavelier, Bediou, & Green, 2018). It must be noted that this interpretation is inconclusive, given that this interaction was found only in production accuracy on the immediate posttest. Yet, this finding is worth exploring further in future research because it precisely fits the prediction of the skill acquisition theory (Anderson, 1993). Moreover, it demonstrates how practice format and schedule influence L2 learning from declarative-procedural knowledge perspectives.

CONCLUSIONS

The aim of the present study was to investigate how practice type (input and output) and schedule (blocked, interleaved, and hybrid) influence the acquisition of comprehension and productive L2 skills. The findings reported in this work support the skill-specificity hypothesis and add to the growing evidence indicating that input and output practice play specific roles in L2 skill acquisition (DeKeyser, 1998; DeKeyser & Sokalski, 1996; Li & DeKeyser, 2017; Li & Taguchi, 2014). Hybrid schedule, compared to blocked and interleaved schedule, resulted in greater processing speed improvements in comprehension, as well as led to more accurate production. A significant interaction between schedule (blocked and interleaved) and prior level of knowledge was found for the speed measures on the comprehension tests. These observations indicate that the effectiveness of blocked practice is contingent on learners' prior processing speed, while interleaved practice may be effective irrespective of prior knowledge levels. Another significant, more complex, interplay between practice format and schedule was detected in the production accuracy achieved on the immediate posttest, which suggests that more fine-tuned, procedural knowledge that is gained from systematic input-hybrid practice was less likely to transfer to production skill.

These findings have obvious implications for L2 classroom instruction, as they emphasize the importance of both input and output practice, particularly when the aim is L2 skill proceduralization. Furthermore, the effects of practice format and schedule on comprehension speed improvements were found to depend on the prior processing speed level that L2 learners bring to practice. Specifically, although learners with higher comprehension processing skills can benefit equally from both blocked and interleaved

practice, interleaved practice should be adopted for learners with lower processing skills. Although comprehension and speaking skills of the specific target structures (RC constructions) were examined in this study, more global L2 proficiency (e.g., listening and speaking abilities) was not assessed. In future research, the proficiency factor should be taken into account and may be examined as a potential moderating variable of training effects.

Given the small sample sizes employed in this work (especially for the speed analysis), our interpretations and implications of the findings reported here are tentative. Obviously, a replication study employing the same design is needed. This study focused solely on a specific element of syntactic structure (accuracy and speed of using subject and object RC constructions). Thus, generalizability of these findings should be further attested by examining different linguistic domains of a variety target structures aimed at assessing diverse L2 skills because different characteristics of grammatical structures may interact with the effects of blocked and interleaved practice. In addition, the practice format was controlled and form-focused, which limits our understanding of how practice format and schedules influence L2 learning through more meaning-focused tasks. Last but not least, the current research contributes to a growing body of L2 research on optimal schedules that are informed by cognitive psychology (e.g., Nakata, 2015; Suzuki, 2017; Suzuki et al., 2019a). Because a majority of previous psychology and L2 research focuses on adult L2 learners in laboratory contexts, it is imperative to examine to what extent the current results along with the existing lab-based findings are applicable for instructed settings (Küpper-Tetzel, Erdfelder, & Dickhäuser, 2014). The present study thus opens avenues for future lab-based and classroom-based investigations into complex effects of pertinent practice variables on L2 acquisition from skill acquisition perspectives (Suzuki et al., 2019a).

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263119000470

NOTES

[1]Procedural and automatized knowledge overlap to some extent. Procedural knowledge is used for specific behavior (as opposed to declarative knowledge) and primarily acquired in an earlier stage of skill learning while automatization takes place later and requires extensive practice for an extended period.

[2]In this article, for consistency, the terms "input" and "output" practice refer to the practice format, whereas "comprehension" and "production" tests refer to assessment tasks.

[3]The hybrid-practice schedule included the interleaved-practice schedule instead of a mini blocked-practice schedule (e.g., ABCDABCDABCD) so that the hybrid-practice schedule could be compared with the "pure" interleaved-practice schedule.

[4]Although the most appropriate modeling procedure for the mixed-effects models are currently debated in the literature, our approach is considered parsimonious and more sensible for the data set in this study (Bates, Kliegl, Vasishth, & Baayen, 2015). All the best-fitting models, except for one, included random intercepts only. The models with only random intercept tend to lead to higher Type I error (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), so the significant fixed effects should be interpreted with caution.

[5]While aptitude is often narrowly specified as cognitive aptitudes (i.e., cognitive abilities that predict the success of L2 learning) in the L2 acquisition literature (Doughty, 2018), the word "aptitude" is used in this

article to keep in line with the original conceptualization of aptitude-treatment interaction in the psychology field (Snow, 1994), which includes prior levels of knowledge among other individual difference factors (e.g., cognitive abilities, motivation, personality).

REFERENCES

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060. doi: 10.1037/0033-295X.111.4.1036.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. arXiv preprint arXiv: 1506.04967.

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.

Bavelier, D., Bediou, B., & Green, C. S. (2018). Expertise and generalization: Lessons from action video games. *Current Opinion in Behavioral Sciences*, *20*, 169–173. doi: 10.1016/j.cobeha.2018.01.012.

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

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

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

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

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.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. doi: 10.1037/0033-2909.132.3.354.

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

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369–382. doi: 10.1177/0267658312443651.

De Jong, N. (2005). Can second language grammar be learned through listening? An experimental study. *Studies in Second Language Acquisition*, *27*, 205–234. doi: 10.1017/S0272263105050114.

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

DeKeyser, R. M. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42–63). 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.

DeKeyser, R. M. (2018). Foreword. In C. Jones (Ed.), *Practice in second language learning* (pp. xiv–xviii). Cambridge, UK: Cambridge University Press.

DeKeyser, R. M., & Botana, P. G. (2015). The effectiveness of processing instruction in L2 grammar acquisition: A narrative review. *Applied Linguistics*, *36*, 290–305. doi: 10.1093/applin/amu071.

DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, *46*, 613–642. doi: 10.1111/j.1467-1770.1996.tb01354.x.

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.

Doughty, C. (2018). Cognitive language aptitude. *Language Learning*. Advance online publication. doi: 10.1111/lang.12322.

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

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

Ferman, S., Olshtain, E., Schechtman, E., & Karni, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics*, *22*, 384–412. doi: 10.1016/j.jneuroling.2008.12.002.

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.

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior*, *36*, 212–224. doi: 10.3200/JMBR.36.2.212-224.

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.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. doi: 10.1016/j.jml.2007.11.007.

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.

Kormos, J. (2006). *Speech production and second language acquisition*. New York, NY: Routledge.

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.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. New York, NY: Longman.

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

Lenth, R. V. (2016). Least-squares means: The r package lsmeans. *Journal of Statistical Software*, *69*, 1–33.

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

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type i error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. doi: 10.1016/j.jml.2017.01.001.

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

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.

Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K. A., Carpenter, H., & Wong, P. C. M. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition*, *17*, 56–72. doi: 10.1017/S1366728912000715.

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

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.

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. doi: 10.1260/174795407782233100.

Porter, J. M., & Magill, R. A. (2010). Systematically increasing contextual interference is beneficial for learning sport skills. *Journal of Sports Sciences*, *28*, 1277–1285. doi: 10.1080/02640414.2010.502946.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition*, *19*, 223–247.

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

Rogers, J., & Cheung, A. (2018). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*. Advance online publication. doi: 10.1177/1362168818805251.

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.

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*, 187–224. doi: 10.1017/S0142716417000418.

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.

Schachter, J. (1974). An error in error analysis. *Language Learning*, *24*, 205–214.

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

Serrano, R., & Huang, H. Y. (2018). Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? *TESOL Quarterly*, *52*, 971–994. doi: 10.1002/tesq.445.

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.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context: Interactionist perspectives on human intelligence* (pp. 3–37). Cambridge, UK: Cambridge University Press.

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. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, *21*, 166–188. doi: 10.1177/1362168815617334.

Suzuki, Y., Nakata, T., & DeKeyser, R. M. (2019a). Optimizing second language practice in the classroom: Perspectives from cognitive psychology. *Modern Language Journal*, *103*, 551–561. doi: 10.1111/modl.12582.

Suzuki, Y., Nakata, T., & DeKeyser, R. M. (2019b). 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.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

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

VanPatten, B. (2002). Processing instruction: An update. *TLanguage Learning*, *52*, 755–803. doi: 10.1111/1467-9922.00203.

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. doi: 10.3109/17549507.2012.761282.

Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How should exemplars be sequenced in inductive learning? Empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied*, *23*, 403–416. doi: 10.1037/xap0000139.

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition*, *40*, 831–856. doi: 10.1017/S0272263118000141.

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.