APTARA	LANG	lang12433	**Dispatch:** September 11, 2020	**CE:** N/A
	Journal	**MSP No.**	**No. of pages:** 41	**PE:** Matthew

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Optimizing Fluency Training for Speaking Skills Transfer: Comparing the Effects of Blocked and Interleaved Task Repetition

Q1 Yuichi Suzuki

Q2 Faculty of Cross-Cultural and Japanese Studies, Kanagawa University, Kanagawa, Japan

Abstract: In an exploration of the effects of task-repetition practice on fluency development, English-as-a-foreign language learners performed three oral narrative tasks involving six-frame cartoons for 3 consecutive days. They engaged in task-repetition practice under either a blocked (Day 1: A-A-A; Day 2: B-B-B; Day 3: C-C-C) or an interleaved (Day 1: A-B-C; Day 2: A-B-C; Day 3: A-B-C) task repetition schedule. The results yielded by a posttest involving new six-frame cartoons indicated that blocked practice resulted in greater fluency development (faster articulation rate and shorter mid-clause pause duration) than did interleaved practice. Moreover, the learners in the blocked-practice group tended to pause more frequently at clause boundaries. Blocked practice also led to significantly longer mean length of run and higher phonation/time ratio during training, although this advantage failed to transfer to meaningful pretest–posttest changes. These dynamic fluency developmental patterns are discussed to elucidate the underlying proceduralization in L2 speech processes.

Keywords task repetition; blocked and interleaved practice; fluency training; task-based language teaching; skill transfer; speaking

---

This study was supported by a Grant-in-Aid for Scientific Research (KAKENHI) from the Japan Society for the Promotion of Science (JP18K12). I would like to express my gratitude to Nel de Jong, Shungo Suzuki, and Keiko Hanzawa for sharing their insight and expertise that greatly assisted this research. I am grateful to Atsushi Miura, Misaki Kuratsubo, Miyu Koyama, Taeko Hosaka, and Kazuma Arai for their dedicated assistance in data collection and coding. I also thank Yuri Hosoda, Harumi Akiba, Keita Kikuchi, and Takahiro Iwahata for their cooperation in participant recruitment. Last, I would like to show my gratitude to the anonymous reviewers and the Editors of *Language Learning* for their insightful feedback and suggestions on earlier drafts of this article.

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

The handling editor for this article was Kara Morgan-Short.

DOI: 10.1111/lang.12433

Introduction

Speaking a second language (L2) fluently is becoming increasingly important. In task-based language teaching and learning, for instance, fluency has been a subject of extensive research as one of the major constructs, alongside complexity and accuracy (Skehan, 2009). Although fluency in a L2 can be interpreted differently in a variety of contexts (Tavakoli & Hunter, 2018), the study reported here focused on utterance fluency (e.g., Skehan, 2003), defined as objective features of utterance such as speed (e.g., articulation rate), breakdown (e.g., pauses), and repair fluency (e.g., repetitions).

Utterance fluency is related to changes in underlying cognitive processes, that is, cognitive fluency (de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013). Cognitive fluency for L2 is defined as the efficiency of executing and integrating utterance planning and assembling linguistic knowledge, which supports—and can be inferred from—L2 utterance fluency (Segalowitz, 2010). Cognitive fluency is highly dependent on procedural knowledge (de Jong & Perfetti, 2011; DeKeyser, 2018; Kahng, 2014; Kormos, 2006). According to skill acquisition and information-processing theory (DeKeyser, 2015; Mclaughlin, 1987), procedural knowledge (knowing "how") contrasts with declarative knowledge (knowing "what"). Declarative knowledge refers to factual information stored in memory, such as lexical items, examples, and rules pertaining to L2. Procedural knowledge is exercised by encoding declarative representations when certain tasks are performed (DeKeyser, 2018). Unlike declarative knowledge, procedural knowledge is processed more rapidly and efficiently and consumes less cognitive recourse to working memory. It is thus more closely related to executing fluent speech in the context of L2 learning and use.

L2 fluency development has been previously examined using short-term interventions in classroom settings (e.g., Lambert, Kormos, & Minn, 2017; Tavakoli, Campbell, & McCormack, 2016) as well as in the study abroad context (e.g., Freed, Segalowitz, & Dewey, 2004). However, L2 utterance fluency is multifaceted. For instance, it may take 10 or more years for different aspects of fluency, such as articulation rate and pause phenomena, to be developed (Saito, Ilkan, Magne, Tran, & Suzuki, 2018), and it can involve multiple facets such as turn-taking behavior (van Os, de Jong, & Bosker, 2020), collocation (Saito, 2020), and multiword sequences (Tavakoli & Uchihara, 2020). In the current study, English-as-a-foreign-language (EFL) learners engaged in a 3-day fluency training program outside their regular English classes. The goal of this intervention was thus short-term development of L2 utterance fluency, a stepping stone toward long-term development that should occur over a more extended period of time.

One research-informed classroom activity to enhance fluency is task repetition (Tavakoli & Hunter, 2018), which is defined as "the repeated performance of tasks that share some of the same pragmatic purpose or purposes and some of the same content" (Bygate, 2018, p. 13). Evidence has provided support for task repetition's facilitating improvement in utterance fluency (e.g., Ahmadian, 2011; Ahmadian & Tavakoli, 2011; Bygate, 1996, 2001; de Jong & Perfetti, 2011; Lambert et al., 2017; Lynch & Maclean, 2000; Thai & Boers, 2016). Researchers have examined variables pertinent to systematic L2 practice that influence L2 fluency development through task repetition such as manipulating task type repetition (e.g., Bygate, 2001; Kim & Tracy-Ventura, 2013) and variations in the tasks to be repeated (de Jong & Perfetti, 2011). The aim of the current study was to extend and refine this research agenda in order to better understand the effects of the key task repetition-related variables by applying insights from the body of research into L2 practice that is informed by cognitive psychology (see Suzuki, Nakata, & DeKeyser, 2019b, for an overview). In this strand of L2-practice research, the optimal distribution of repeated practice, such as massed versus spaced practice and blocked versus interleaved practice (e.g., Y. Suzuki, in press), has recently been examined for L2 learning. This emerging line of investigations holds promise for providing useful insights into the most effective task-repetition schedule in the task-based language teaching and learning framework (DeKeyser, 2018). Yet, its potential has not been fully exploited in either research or practice. The current study aimed to address this gap in knowledge by integrating two independent streams of L2 research (task-based language teaching and optimal L2 practice) to determine the systematic task-repetition practice schedule that would result in optimal L2 fluency development.

Background Literature

Optimizing Learning Through Systematic Repeated Practice

Cognitive psychology research has yielded a substantial body of knowledge regarding the benefits of systematic repeated practice for diverse forms of learning in a variety of subject areas such as mathematics, verbal memory, reading skills, and motor skills (e.g., Hattie, 2009; Horvath, Lodge, & Hattie, 2016). One of the topics that has been extensively studied is the comparison between massed practice (no spacing is provided between different categories of stimuli, as in AAA) and spaced practice (a temporal lag is included between practice tasks, as in A___A___A; e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006;

Donovan & Radosevich, 1999; Janiszewski, Noel, & Sawyer, 2003). The last decade has seen a surge of L2 studies seeking to identify the optimal temporal distribution of repeated practice for the acquisition of different linguistic domains such as vocabulary (Nakata, 2015; Nakata & Suzuki, 2019a; Serrano & Huang, 2018), pronunciation (Li & DeKeyser, 2019), and grammar (Bird, 2010; Kasprowicz, Marsden, & Sephton, 2019; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017). Evidence yielded by these investigations has allowed researchers to elucidate the effects of different practice distributions and indicated that optimal distribution varies depending on key moderating variables such as the number of practice sessions, complexity of target skills, cognitive aptitudes, skill/knowledge type, and the time between training and test (see Y. Suzuki, in press, for details).

A related stream of research concerns blocked and interleaved practice, the main focus of the current investigation. In a blocked schedule, each of the practice tasks is repeated in a sequence before moving on to another task (e.g., AAABBBCCC). In an interleaved schedule, practice tasks of different types are interspersed (e.g., ABCABCABC).[1] The advantage of interleaved practice over blocked practice has been found in prior research involving a single practice session for conceptual and category learning such as painting styles, diagnosis of disorder, and mathematics (see Brunmair & Richter, 2019; Kang, 2016 for review). This learning advantage is often referred to as the interleaving effect, which is "a positive effect of an interleaved compared with a blocked inductive learning condition on the performance in a subsequent category discrimination or classification task" (Brunmair & Richter, 2019, p. 2).

Researchers have attempted to account for interleaving effects in different ways. Some have based their study designs on the discriminative contrast hypothesis (Kang & Pashler, 2012), which states that a sequence of mixed exemplars from different categories highlights the differences between categories. This hypothesis was subsequently refined, giving rise to the sequential attention theory (Carvalho & Goldstone, 2017), which offers a more detailed theoretical explanation for interleaving effects. According to the attentional framework, interleaved practice can shift learners' attention to different properties of consecutive items, thereby promoting the discrimination of similar categories.

Although the interleaving effect has mostly been studied in the context of category/classification learning (Brunmair & Richter, 2019; Rohrer, 2012), different types of learning have been examined in the motor-skill acquisition research field. Many of the skills studied (e.g., baseball batting, badminton shots, and golf shots) comprise multiple components and involve little classification or discrimination. Yet, interleaved practice has been found to be more conducive to learning than has blocked practice (Brady, 2004). This advantage of interleaving in motor skill acquisition is often attributed to contextual interference. Contextual interference refers to the phenomenon whereby a practice condition that causes higher interference of skills eventually leads to better retention and transfer of acquired knowledge than does a practice condition with lower interference (see Merbah & Meulemans, 2011, for more detailed theoretical accounts of contextual interference). This can be linked to the desirable difficulty hypothesis, which posits that a practice condition that is optimally challenging for learners can enhance the acquisition of knowledge and skills (Bjork, 2018; Suzuki, Nakata, & DeKeyser, 2019a). Although the effects of blocked and interleaved practice have been primarily studied in non-L2 contexts, cognitive psychologists and L2 researchers have recently started to investigate the extent to which interleaved practice supports L2 learning.

Blocked and Interleaved Practice in L2 Research

Accumulated evidence has suggested that the interleaving effect extends to L2 grammar learning (Nakata & Suzuki, 2019b; Pan, Tajran, Lovelett, Osuna, & Rickard, 2019; Suzuki & Sunada, 2019; Suzuki, Yokosawa, & Aline, 2020). The advantage of interleaved practice was found in the acquisition of the English tense–aspect–mood distinction (Nakata & Suzuki, 2019b) and Spanish past-tense morphology (Pan, Lovelett, Phun, & Rickard, 2019), as well as in English subject/object relative clause constructions (Suzuki & Sunada, 2019; Suzuki et al., 2020). For instance, in Nakata and Suzuki's (2019b) study, EFL learners were introduced to five structures from the English tense–aspect–mood system: simple past, present perfect, first conditional, second conditional, and third conditional. They practiced the target structures in a written multiple-choice fill-in-the-blank question response format under interleaved-practice and blocked-practice conditions. Specifically, a total of 50 multiple-choice questions from each of the five structures were either blocked by category (e.g., simple past, 10 items; present perfect, 10 items; first conditional, 10 items…) or interleaved from different categories (e.g., simple past, first conditional, present perfect, third conditional, simple past…). They were subsequently tested using a written grammaticality judgment task. The results yielded by the 1-week delayed posttest showed that students in the interleaved-practice group significantly outperformed those in the blocked-practice group. This finding can be explained by the sequential attention theory (Carvalho & Goldstone, 2017) in that practice involving a sequence of interleaved exemplars is beneficial for facilitating the discrimination of similar grammatical categories.

The implications of these studies for the acquisition of different L2 skills have been difficult to establish, given that their authors mainly examined form-focused grammar practice targeting the ability to distinguish similar types of grammatical rules. In order to be able to speak a L2 fluently, for example, practice must go beyond such a narrow focus as in the domains of classification and discrimination tasks. Consequently, the current study broadened the scope of blocked versus interleaved practice to investigate the effects of task-based oral narrative practice on fluency development. This research objective responded to the recent call (e.g., Rohrer, 2012) for extending research into interleaving to identify the potential and the limits of interleaved practice for the development of less similar concepts and skills. Furthermore, blocked and interleaved practice were manipulated within a single session in all prior L2 studies, and the current study marked the first attempt to investigate the effects of blocked versus interleaved practice extending over multiple sessions.

Task Repetition for L2 Fluency Development and Proceduralization: Systematic Task Repetition and Practiced Skill Transfer

Task repetition is an effective teaching technique for enhancing L2 fluency among different types of learners using a wide variety of task types (e.g., Ahmadian, 2011; Ahmadian & Tavakoli, 2011; Bygate, 1996, 2001; de Jong & Perfetti, 2011; Lambert et al., 2017; Lynch & Maclean, 2000; Thai & Boers, 2016).

The benefits for speech of repeating the same task can be explained by Levelt's (1989) speech production model and the bilingual production model (Kormos, 2006), both of which comprise three major components: (a) the conceptualizer (e.g., creating a preverbal message), (b) the formulator (e.g., encoding of lexical and grammatical knowledge), and (c) the articulator. It is argued that, when L2 learners perform a task for the first time, they allocate their attentional resources in the working memory for conceptualization, for example, creating preverbal intention and thought. Consequently, a limited number of attentional resources are available for formulation and articulation. In the subsequent performance of the same task because L2 learners are already familiar with the task content, they can free up their attentional resources, some of which can be used for linguistic formulation (e.g., lexical and grammatical encoding). These speech production models (Kormos, 2006; Levelt, 1989) can offer a prediction that task repetition may be particularly beneficial when the content of tasks used in practice is identical (Bygate, 1996; Fukuta, 2016).

As a case in point, Lambert et al. (2017) documented L2 fluency development through repeated speaking practice based on the same task. In a classroom-based study, Japanese university EFL learners engaged in a paired speaking task (instruction task, narration task, or opinion task) six times. The performance changes across the six repetitions were analyzed in terms of speech rate (the number of pruned syllables), mid-clause and clause-final filled pauses, self-repetition, and the number of self-repairs. Analysis results showed that there was a steady, significant increase in speech rate and a decrease in mid-clause pauses until the fifth performance. The number of self-repairs also decreased significantly in the fifth performance in comparison to those observed in the first two performances. The fluency development found by Lambert et al. (2017) may reflect potential proceduralization or automatization of underlying linguistic knowledge. However, the authors noted that there was no significant effect of task repetition on the changes in clause-final pauses or in self-repetition, where results were less closely related to L2 cognitive fluency (de Jong et al., 2013; Kahng, 2014).

Although Lambert et al. (2017) clearly demonstrated that repeating the same task several times improved fluency, an interesting question arises regarding the optimal degree of similarity among the repeated tasks for maximizing fluency development, for example, narrative AAAAAA (same task repetition) versus narrative ABCDEF (repeated practice involving different tasks). de Jong and Perfetti (2011) examined the effects of content similarity (variability) of repeated tasks on L2 fluency development and proceduralization. As a part of their investigation, students of English as a L2 at a university in the United States engaged in nine monologue 4/3/2-minute speech tasks, where the time allocated for each task progressively decreased over a period of 2 weeks. The no-repetition group (nine learners) performed the speech task on nine different topics (Day 1: ABC; Day 2: DEF; Day 3: GHI), but the two repetition groups (10 learners in the first experiment and five in the second experiment) performed the speech task on the same topic three times (Day 1: AAA; Day 2: DDD; Day 3: GGG). One-week and 4-week delayed posttests were administered to examine the fluency gains due to repeated practice transfer to speech performance on a different topic. Analysis of the delayed posttest results indicated that only the repetition groups demonstrated meaningful fluency development related to proceduralization (e.g., mean length of run, pause length, and phonation/time ratio). The authors argued that the repetition of the same task induced the changes underlying cognitive fluency or proceduralization, which led to enhanced utterance fluency.

de Jong and Perfetti's (2011) study is noteworthy because its findings suggested that repetition of the same task is more effective than repeated practice involving different tasks for transferring the effects of fluency training to performance on a different task. Because this practice transfer issue in the task-based language teaching framework has rarely been investigated (for the rare examples, see Ahmadian, 2011; Bygate, 2001; de Jong & Perfetti, 2011; Kim & Tracy-Ventura, 2013), de Jong and Perfetti's (2011) findings are particularly valuable because they indicated that systematically manipulating the sequence of practice tasks can enhance L2 proceduralization that can be transferred to a new task of the same type (e.g., narration of a different story). However, de Jong and Perfetti's (2011) study was not without limitations. Because the participants in their study were learners of English as a L2 living in the United States, they probably had many opportunities to speak English outside the classroom during the experimental period, inducing potential sources of confounding variables. Additionally, only 24 participants were involved in the experiment, which clearly indicated the need for further research with a larger sample of EFL leaners from contexts where opportunities for speaking English outside the study setting are relatively limited.

The Current Study

The current study lies at the interdisciplinary nexus of task repetition research in task-based language learning (Bygate, 2018) and L2 practice research informed by cognitive psychology (Suzuki et al., 2019b). The study sample included 68 English learners at a Japanese university who engaged in oral narrative tasks using six-frame cartoons three times a day for three consecutive days. Specifically, the sequence of three types of cartoons was manipulated while keeping the task variation equal, allowing the effects of the blocked-practice condition (e.g., Day 1: AAA; Day 2: BBB; Day 3: CCC) and interleaved-practice condition (e.g., Day 1: ABC; Day 2: ABC; Day 3: ABC) to be compared.[2] A pretest and a posttest adopting different stories from those used for the practice tasks were administered to measure the transfer of fluency improvement to different content of the same task type. The following two research questions were addressed:

1. How does L2 fluency change during the task-repetition practice phase under blocked- and interleaved-practice conditions?
2. To what extent does L2 fluency training through repeated narrative tasks under blocked- and interleaved-practice conditions lead to fluency gains measured by performance on new narrative tasks?

For Research Question 1, it was hypothesized that, during the training phase, fluency would increase more steadily in the blocked-practice condition (e.g., AAA) than in the interleaved-practice condition (e.g., ABC) within each day. Previous research on L2 grammar practice (Nakata & Suzuki, 2019b; Suzuki & Sunada, 2019) had shown that learners in the blocked-practice condition outperformed those in the interleaved-practice condition, and this was assumed to be the case for L2 fluency training.

Research Question 2 addressed the extent to which blocked and interleaved practice lead to fluency development, as measured by performance on a new narrative task. This issue concerns the transfer of practiced skills, which is often described as the Holy Grail of education (Haskell, 2001). The hypothesized answer to this key research question was left open, and two possible scenarios were envisaged. On the one hand, interleaved practice might better facilitate transfer than would blocked practice. This potential advantage of interleaving may be accounted for by the transfer-appropriate processing model (Lightbown, 2008; Morris, Bransford, & Franks, 1977). This model stipulates that, when a training condition and a test condition are similar, learners are more likely to apply their learned skills effectively in the test phase. When presented with different interleaving narrative tasks, learners need to deal with a variety of tasks within the same day. Because interleaving might create practice situations similar to the testing condition where a less familiar task must be performed, it may equip learners with the ability to cope with a novel narrative task that necessitates a strategic use of new linguistic resources. This prediction might also be supported by the contextual interference effect and the desirable difficulty hypothesis (see the Background Literature section), because interleaving may prompt learners to engage in speaking practice under a more desirably difficult condition though alternating narrative stories. Thus, in this scenario, interleaved practice involving multiple narrative tasks performed on the same day was predicted to increase the transferability of fluency training.

On the other hand, although interleaving has been shown to be beneficial for learning of similar grammatical features for relatively simple and decontextualized skills (Nakata & Suzuki, 2019b; Pan et al., 2019; Suzuki & Sunada, 2019), this might not always be the case for more complex and challenging activities, such as speaking practice. Learners might find it too hard to engage in fluency training effectively through interleaving because they are less likely to use newly acquired linguistic and cognitive resources immediately on the same day. It was thus argued that blocking might facilitate fine-tuning or proceduralizing complex skills through repetition of the same tasks. If this were the case, the advantage of blocking might be observed in fluency indices such as articulation rate, mean length of run, and pauses within the clause, which presumably reflect cognitive fluency and proceduralization (e.g., Kahng, 2017; Kormos, 2006; Saito et al., 2018).

Method

Participants

The participants (aged 18−22 years) were recruited for the current study through announcements in regular EFL classes at a university in Japan. Their English proficiency was estimated to fall between the A2 (elementary) and the B1 (intermediate) levels in the Common European Framework of Reference for Languages benchmarks.[3] They were randomly assigned to either a blocked-practice condition ($n = 24$) or an interleaved-practice condition ($n = 26$). The learners in the control group ($n = 18$), who took the pretest and posttest only, were recruited from the same classes as those assigned to the two experimental groups.[4]

In order to identify the presence of potential proficiency differences among the three groups at the outset of the study, the participants' L2 English proficiency was assessed using an objective proficiency test titled the junior English Minimal Test, which was developed for research purposes. This short dictation test has been found to positively correlate with general L2 English proficiency as measured by scores on the reading and listening sections of Japanese university entrance exams (Goto, Maki, & Kasai, 2010). The means (standard deviations) were 47.96 (8.00), 44.62 (9.58), and 46.93 (6.70) for the blocked-practice, interleaved-practice, and control groups, respectively. A one-way ANOVA revealed that there were no significant differences among the three groups, $F(2, 61) = 1.02$, $p = .37$, $\eta^2 = .03$.

Instruments

Training Materials

Three prompts (bicycle, tiger, and race), each consisting of six-panel picture stories, were used for the fluency training. They had been adopted from Heaton (1996) by de Jong and Vercellotti (2016) and de Jong and Tillman (2018) in their research on L2 oral production. All three picture stories and the guiding questions for them (available in the IRIS digital repository of data collection instruments, see Marsden, Mackey, & Plonsky, 2016) were adopted from their studies. These three stories had a tight sequential structure with similar narrative structure involving little causal reasoning, that is, the main character

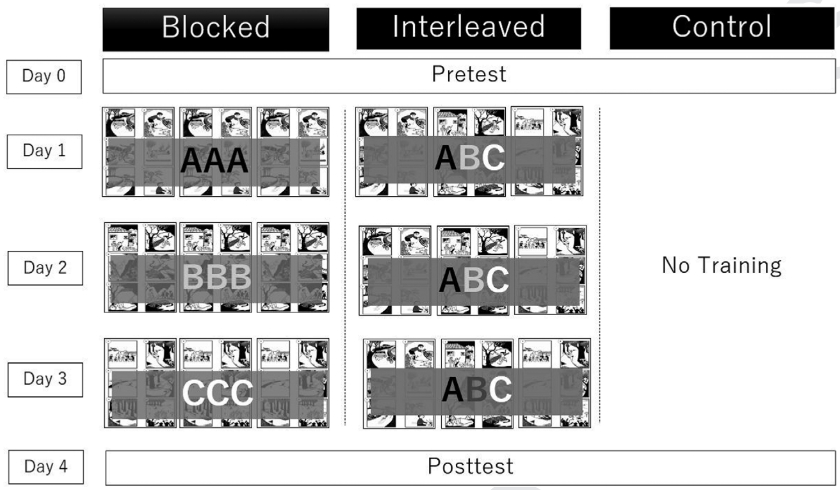

Figure 1 Schedule of pretest, training sessions, and posttest. The three six-picture prompt series used in the training sessions are indicated by A, B, and C. Because the order of these prompts was counterbalanced, individual letters do not refer to a specific prompt (see Appendix S1 in the online Supporting Information for more details).

experiences one surprising event. Each of the prompts also involved two main characters in an outside location (road, mountain, and forests).

Pretest and Posttest

A different set of two prompts—street (Oba, 2018) and airport (Suzuki, 2011)—was used for the pretest and posttest. As in the training prompts, the pretest and posttest prompts also consisted of six-panel stories. These two stories also had a tight sequential structure with a similar narrative structure involving little causal reasoning (i.e., a thief steals the main character's purse/suitcase, and another main character helps to catch the thief). Each of the prompts involved three main characters (thief, victim, helper) in a different location (street and airport). All instruments are available in the IRIS digital repository of data collection instruments (Marsden et al., 2016).

Procedure

Overview

Figure 1 shows the pretest–training–posttest design adopted in the present study. In the pretest, all participants were tested individually in a computer lab 1 week prior to the training session. After the participants in the experimental conditions had been randomly assigned either to the blocked condition or to the interleaved condition, they received an experiment packet containing three booklets with training materials, a timer, and a digital recorder, and the participants were instructed to follow the prescribed 3-day fluency training program outside of the lab (e.g., in a quiet place such as at home). The three booklets included the instructions and the necessary information to complete the training individually (e.g., instructions on how to use the timer and record their speech using the digital recorder). These booklets were created for both intervention groups. However, the participants assigned to the blocked-practice condition performed the same narrative three times a day (e.g., AAA on the first day, followed by BBB on the next day, and finally CCC), whereas those in the interleaved-practice condition performed three different narratives on each of the three days (e.g., ABC, ABC, ABC).[5] To ensure that the participants performed the fluency training as indicated, a research assistant sent a daily reminder to them using a chat application on their smart phones. The participants had to report to the research assistant every day when they had finished their training. All the participants returned to the computer lab for the posttest 1 day after the last training session.

Pretest and Posttest Sessions

The pretest and posttest procedure included: (a) a 3-minute preparation with the cartoon (i.e., the prompt), the guiding questions, and a list of useful vocabulary; (b) a 3-minute oral narration while the participants viewed the prompt; and (c) two questionnaires about the participants' planning and task performance behaviors. The test booklet instructed that "Yesterday, you saw an event depicted in the six-frame cartoon on the next page. You are going to explain the story to a friend who doesn't know the story in three minutes." The participants were allowed to plan their narration for 3 minutes, aided by the picture prompt, a list of useful vocabulary (13 English words, along with their Japanese translation equivalents), and several guiding questions. The guiding questions were included "to provide the participants with additional suggestions for content, potentially reducing individual differences" (de Jong & Vercellotti, 2016, p. 393). The participants were told that they would not be allowed to take notes or read the guiding questions during the test. After the preparation phase, the participants narrated the story for three minutes aided only by the prompt (i.e., without access to the vocabulary list or questions). Finally, after each narration practice, the participants completed two questionnaires inquiring about their behaviors during the preparation time and about their performance of the task. Due to space limitations, the survey results have not been reported in this paper. The order of the two test prompts (Street and Airport) was counterbalanced across the participants at pretest and posttest to minimize task effects (see Appendix S1 in the online Supporting Information).

Fluency Training Sessions

The participants followed the instructions for each day provided in the booklet and performed the three narratives based on the designated schedule (see Figure 1). They were told that they could undertake the session at any time during the day and that the next session should occur after the nighttime sleep. The training procedure was identical to that adopted in the English pretest and posttest sessions with the guiding questions and the vocabulary list. As instructed in the booklet, the participants prepared their speech for three minutes, narrated the story for three minutes using the timer and digital recorder in their packet, and completed the questionnaires. Each training session lasted about 30 minutes. Although authors of some previous studies have gradually imposed time pressure on repeated task performance using the 4/3/2 procedure (e.g., de Jong & Perfetti, 2011; de Jong & Tillman, 2018), the participants in the current study were allowed the same amount of time (i.e., 3 minutes) for each narration throughout the experiment. This schedule was adopted because imposing time pressure sometimes reduces the amount of repetition in subsequent performances (in part due to learners' tendency to leave out information in order to meet the time requirements), which may lead to less ideal conditions for proceduralization (N. de Jong, personal communication, October 16, 2018).

Data Coding

Using the free sound-analysis software PRAAT (Boersma & Weenink, 2016), three trained coders first identified the filled and unfilled (silent) pauses of at least 200 milliseconds duration (de Jong & Perfetti, 2011) with the help of a PRAAT script (de Jong & Wempe, 2009). After the three coders familiarized themselves with the coding scheme in a training session, they independently coded 10% of the pretest and posttest datasets. Pearson's correlation coefficients between the coders were above .90 for all fluency measures. They also transcribed the utterances into analysis-of-speech units (broadly similar to clauses; Foster, Tonkyn, & Wigglesworth, 2000). Their work was subsequently assessed by another coder to ensure accuracy of all transcriptions.

Choosing relevant fluency measures in L2 fluency research is difficult. For the present study, the following nine measures were selected to ensure compatibility with pertinent prior research (de Jong & Perfetti, 2011; Kahng, 2014; Lambert et al., 2017):

1. mean length of run (the number of syllables between pauses including false starts and repetition but excluding fillers),
2. articulation rate (the number of syllables per minute of speech, excluding pauses),
3. phonation/time ratio (utterance duration divided by the total duration),
4. mid-clause pause duration (mean duration of mid-clause filled and unfilled pauses),
5. clause-final pause duration (mean duration of clause-final filled and unfilled pauses),
6. mid-clause pause frequency (the number of mid-clause filled and unfilled pauses per minute),
7. clause-final pause frequency (the number of clause-final filled and unfilled pauses per minute),
8. repetition frequency (the number of repetitions per minute), and
9. repair frequency (the number of self-repairs per minute).

These nine variables were computed to capture different aspects of utterance fluency (Skehan, 2003), namely speed fluency (1−3), breakdown (4−7), and repair fluency (8−9).

Before proceeding with the analysis of the obtained results, three important characteristics of these fluency measurements should be noted. First, mean length of run and phonation/time ratio may be conceptualized as a composite index of speed and breakdown fluency because they are dependent on the number of pauses as well as the speed aspect of fluency (Bosker, Pinget, Quené, Sanders, & de Jong, 2013). Despite this relatively global nature of these two measures, both were retained to facilitate comparisons with the results obtained by de Jong and Perfetti's (2011) earlier study. Second, the distinction between mid-clause pauses (i.e., within the analysis-of-speech unit) and clause-final pauses (i.e., at the boundary of the analysis-of-speech unit) reflects different L2 speech processes. The former is more strongly related to cognitive and perceived fluency than the latter (Kahng, 2014, 2017; Saito et al., 2018). The pauses within clauses presumably indicate linguistic breakdowns such as lexical and syntactic ones and are disrupting to listeners, that is they are stronger indicators of cognitive fluency, but pauses between clauses tend to reflect conceptualization including planning of content (de Jong, 2016; Kahng, 2017; Lambert et al., 2017; Skehan, Foster, & Shum, 2016). Third, repetition and repair frequency may not be a strong indicator of L2 cognitive fluency or perceived fluency (de Jong et al., 2013; Saito et al., 2018) because repetitions and repairs stem not only from a variety of linguistic (e.g., weak representation of linguistic knowledge) but also from non-linguistic variables (e.g., L1 speaking styles, lack of attention, and anxiety; Derwing, Munro, Thomson, & Rossiter, 2009; Zuniga & Simard, 2019). The benefits of task repetition may be less likely to be observed for these repair fluency aspects compared to what has been observed for the other fluency measures.

Statistical Analysis

Analysis of Training Data

In order to compare the fluency changes between blocked and interleaved conditions during the training period, a series of two-way mixed ANCOVAs were conducted for each training day. Each of the fluency measures on each training day (9 measures × 3 days) was used as a dependent variable. Time (second and third performance) was the within-subject variable, and practice condition (blocked and interleaved) was the between-subject variable. The fluency measure from the first performance of each training day (i.e., Day 1–1 [Day 1−performance 1], Day 2–1 [Day 2−performance 1], and Day 3–1 [Day 3−performance 1]) served as a covariate for the analyses for that day. This strategy allowed the second and third performances to be compared between the two practice conditions, after controlling for potential individual differences in the first performance of each training day (an analytical design recommended by Dimitrov & Rumrill, 2003). In other words, changes from the first performance to the subsequent performance within each day were compared between the two conditions. The interaction between condition and time was also included in the model to identify any group differences at different time points. No more than two participants were identified in some of the ANCOVAs as outliers ($z > 3.29$; Tabachnick & Fidell, 2013, see Appendix S2 in the online Supporting Information) and were thus excluded from those analyses. Because the dependent variable repetitions was not normally distributed for all training sessions, a log-transformation was performed to correct the distributions for repetitions. The assumption of homogeneity of variance was met for all analyses. The alpha level for statistical significance was set at less than .05. Because there is no specific benchmark for effect size partial eta squared for the interleaving effect, the effect size magnitudes were also interpreted based on the educational research benchmark for partial eta squared (Richardson, 2011; small: $\eta_p^2 = .0099$; medium: $\eta_p^2 = .0588$; and large: $\eta_p^2 = .1379$). Although it was difficult to establish a meaningful effect size for the ANCOVA results, the medium effect size of .0588 was selected as the minimum effect size for a meaningful main effect, which can be regarded as not too liberal, or at least not conservative.

When a main effect or interaction was significant in the two-way mixed ANCOVAs, follow-up univariate ANCOVAs were conducted for each practice performance with condition (blocked and interleaved) as the between-subject variable for those fluency measures that were significant. The score on Day 1−1 was used as the covariate for estimating Day 1−2 and Day 1−3 means, whereas the score obtained on Day 2−1 served as the covariate for estimating Day 2−2 and Day 2−3 means, and the score on Day 3−1 was used as the covariate for estimating Day 3−2 and Day 3−3 means. These adjustments were useful for controlling for the potential differences in the first performance on each day. The effect size of group difference—Cohen's *d*—was computed using the adjusted posttest scores. Its magnitude was interpreted based on a specific benchmark (see Suzuki, 2017 for a similar approach). In a meta-analysis of 59 studies in the field of psychology (Brunmair & Richter, 2019), the overall effect size of interleaving effect was $d = 0.42$. A similar mean effect size of 0.45 was obtained by synthesizing the results reported in three recent L2 studies on interleaved grammar practice among EFL learners[6] (Nakata & Suzuki, 2019b; Suzuki & Sunada, 2019; Suzuki et al., 2020). Consequently, in the current intervention, the treatment-specific effect size of approximately 0.40, corresponding to a small between-subject effect size according to a L2 field-general benchmark (Plonsky & Oswald, 2014), was used as the minimum effect size of meaningful difference. It should be noted that this treatment-specific effect size concerned the pretest–posttest changes for which

Q3

Analysis of Pretest and Posttest Data

A series of univariate ANCOVAs were conducted on the posttest fluency measure scores with condition (blocked, interleaved, control) as a between-subject variable, as well as with the corresponding pretest as a covariate for each fluency measure.[7] The main effect of condition was interpreted in the same way as it was for the training data results. When the main effect of condition was statistically significant, multiple comparisons with Bonferroni corrections were conducted to compare the differences among the three groups. Effect size magnitudes were interpreted in the same way as for the training data results (i.e., Cohen's $d =$ approximately 0.40 was regarded as the minimum effect size of meaningful difference). According to the *z*-scores for pretest scores ($z > 3.29$, Tabachnick & Fidell, 2013), one to three participants were identified as outliers for a few fluency measures on which ANCOVAs were conducted and were thus excluded from analysis (see Appendix S2 in the online Supporting

Information). The assumption of homogeneity of variance was met for all analyses. However, one of the dependent variables (repetitions) was not normally distributed on the posttest. Because the data for repetitions were not adequately corrected after a log or a square root transformation, a rank ANCOVA was conducted using the ranks of the pretest and posttest repetition scores.

Results

Training Performance Change

Table 1 summarizes the mixed ANCOVA results for the training data. Significant main and/or interaction effects were found for most fluency measures; two of the nine measures (mean length of run and clause-final pause frequency) were significant on all 3 training days, four (articulation rate, phonation/time ratio, mid-clause pause duration, clause-final pause duration) were significant on 2 training days, and one (repetitions) was significant on only 1 training day. The size of partial eta squared for the main or interaction effects presented in Table 1 (.06–.37) was medium to large and considered meaningful. The interaction effects suggested that the blocked and interleaved practice groups differed at different time points, which was further analyzed in follow-up univariate ANCOVAs. No significant main effect or interaction effect was found in two fluency measures (mid-clause pause frequency and repairs), indicating that no meaningful difference existed between the blocked- and interleaved-practice conditions for these measures.

The results yielded by follow-up univariate ANCOVAs are presented in Figure 2, where mean scores—adjusted for the scores on the first practice performance of each day—are graphically depicted. Overall, during the training phase, the participants assigned to the blocked-practice condition exhibited superior performance when compared to those in the interleaved-practice condition. Table 2 summarizes the effect sizes of group difference for each fluency measure, indicating that the Cohen’s *d*s for these significant effects (0.50–1.52) were all above the treatment-specific benchmark of 0.40. (The other effect sizes can be found in Appendix S5 in Supporting Information online.)

The participants from the blocked-practice group extended their mean length of run, accelerated their articulation rate, increased their phonation/time ratio, and decreased their mid-clause duration, their clause-final pause duration, and their repetitions. Interestingly, however, on each day, clause-final frequency significantly increased in the blocked-practice group compared to that noted for the interleaved-practice group. This divergent trend in the obtained findings also merited attention in the pretest–posttest score analyses.

Table 1 Summary of the effect sizes (η_p^2) and significance for the two-way mixed ANCOVA results (training data)

	Day 1		Day 2		Day 3	
Fluency measure	Condition	Time × Condition	Condition	Time × Condition	Condition	Time × Condition
Mean length of run	.10*	.09*	.21*	—	.13*	.06+
Articulation rate	.14*	.08+	.09*	.13*	—	—
Phonation/time ratio		.07+	.37*	.06+	.22*	.08*
Mid-clause pause duration	—	—	.16*	.14*	.16*	—
Clause-final pause duration	.09*	—	—	—	.23*	.09*
Mid-clause pause frequency	—	—	—	—	—	—
Clause-final pause frequency	.10*	—	.10*	.09*	.20*	.13*
Repetitions	—	—	.10*	—	—	—
Repairs	—	—	—	—	—	—

Note. To enhance interpretations, effect sizes are provided only for cells that yielded a significant or marginally significant main effect or interaction. The p values and other effect sizes that are not reported in this table are available in Appendix S3 in the online Supporting Information.
$^{+}p < .10$. $^{*}p < .05$.

Table 2 Summary of effect sizes (Cohen's *d*) and significance for pairwise comparisons of blocked and interleaved conditions (training data, compared to the first performance in each day's session)

Fluency measure	Day 1–2	Day 1–3	Day 2–2	Day 2–3	Day 3–2	Day 3–3
Mean length of run	—	0.75*	0.80*	1.01*	0.52+	0.80*
Articulation rate	—	0.91*	—	0.91*	—	—
Phonation/time ratio	—	0.60*	1.02*	1.52*	0.79*	1.13*
Mid-clause pause duration	—	—	—	−1.05*	−0.50+	−0.95*
Clause-final pause duration	—	−0.62*	—	—	−0.74*	−1.03*
Mid-clause pause frequency	—	—	—	—	—	—
Clause-final pause frequency	0.55+	0.53+	—	0.83*	0.51+	1.08*
Repetitions	—	—	−0.84*	—	—	—
Repairs	—	—	—	—	—	—

Note. Effect sizes are presented only for cells that yielded a significant main effect of condition in the univariate ANCOVAs. A positive effect size indicates a higher value in blocked practice, whereas a negative effect size indicates a higher value in interleaved practice. The *p* values and other effect sizes that are not reported in this table, along with 95% confidence intervals for all effect sizes, are available in Appendix S5 in the online Supporting Information.

$^{+}p < .10$. $^{*}p < .05$.

lang12433 W3G-lang.cls September 11, 2020 17:39

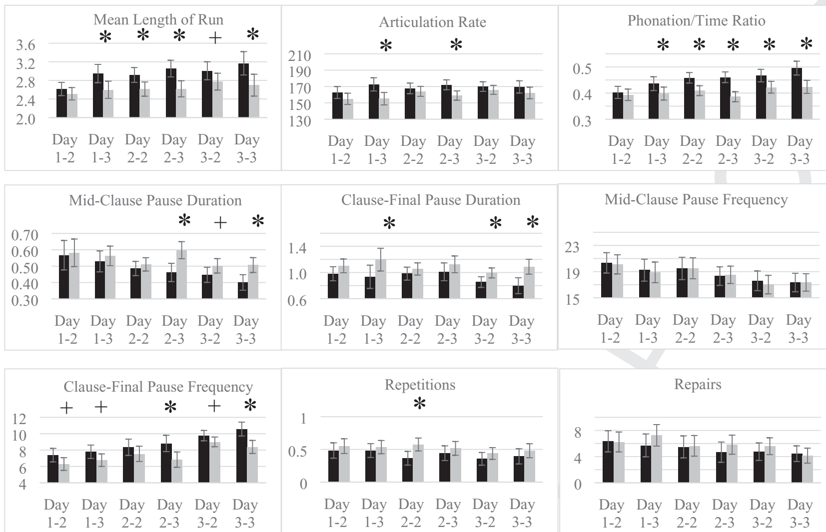

Figure 2 Adjusted mean scores for blocked and interleaved conditions during the training phase. The black bars represent blocked-practice condition, and the gray bars represent interleaved-practice condition. Mean scores were adjusted for the scores on the first practice performance of each day (i.e., covariate). The error bars represent 95% confidence intervals. All raw scores during the training performance are presented in Appendix S4 in the online supplementary file.
$+p < .10$. $*p < .05$.

Furthermore, as suggested by the Time × Condition interaction effects yielded by the mixed ANCOVA, the group differences were more pronounced for the third performance (i.e., Day 1–3, Day 2–3, Day 3–3) than for the second performance (i.e., Day 1–2, Day 2–2, Day 3–2) on all fluency measures except repetitions. For instance, significant group differences were found only on the third performance for articulation rate, mid-clause pause duration, clause-final pause frequency, suggesting that the cumulative blocked practice effects appeared in the third performance.

Pretest–Posttest Score Changes

Table 3 summarizes the ANCOVA results for posttest scores adjusted for pretest scores. A significant main effect of condition was found for the articulation rate with a meaningful effect size. A marginally significant main effect (i.e., $p < .10$, but $> .05$) was observed in the following four breakdown fluency measures with meaningful effect sizes (all $\eta_p{}^2 > .06$): mid-clause pause

Table 3 Summary of the univariate ANCOVA results for the posttests adjusted for the corresponding pretest

Fluency measure	*F*	*df*	*p*	η_p^2
Mean length of run	0.66	2, 64	.52	.02
Articulation rate	3.47	2, 64	.04*	.10
Phonation/time ratio	0.59	2, 64	.56	.02
Mid-clause pause duration	2.83	2, 63	.07+	.08
Clause-final pause duration	3.07	2, 61	.05+	.09
Mid-clause pause frequency	2.72	2, 64	.07+	.08
Clause-final pause frequency	2.51	2, 63	.09+	.07
Repetitions	0.43	2, 64	.65	.01
Repairs	1.18	2, 63	.32	.04

Note. See Appendix S6 in the online Supporting Information for the full results.
$^+p < .10$. $^*p < .05$.

duration, clause-final pause duration, mid-clause pause frequency, and clause-final pause frequency. No significant main effect was noted for mean length of run, phonation/time ratio, repetitions, and repairs.

Figure 3 presents posttest scores adjusted for pretest scores for all nine fluency measures for each group. A summary of multiple pairwise comparison results for the (marginally) significant five fluency measures is provided in Table 4, each of which will be discussed in detail. Instead of using *p* values as a dichotomous cutoff for significant and non-significant results, the effect sizes along with their 95% confidence intervals have primarily been reported to provide more nuanced interpretations of the findings. Specifically, the treatment-specific effect size (Cohen's *d*) of approximately 0.40 was used as the benchmark for interpreting meaningful differences. It is worth noting that some results from the multiple pairwise comparisons were not significant. The results should thus be interpreted with caution particularly when 95% confidence intervals include 0.

Articulation Rate

Multiple pairwise comparisons in Table 4 showed that the blocked-practice group improved their articulation rate relative to both the interleaved-practice group and the control group with meaningful effect sizes, although no meaningful difference was detected between the interleaved-practice group and the control group.

Table 4 Summary of effect sizes for the multiple pairwise comparison results by fluency measure and condition (blocked compared to interleaved and to control) for the posttests adjusted for the corresponding pretest

	Condition					
	Interleaved			Control		
Fluency measure / Condition	*d*	95% CI	*p*	*d*	95% CI	*p*
Articulation rate						
Blocked	0.66+	[0.08, 1.22]	.07	0.68+	[0.04, 1.30]	.10
Interleaved	—	—	—	0.03	[−0.58, 0.63]	1.00
Mid-clause pause duration						
Blocked	−0.60	[−1.16, −0.03]	.12	−0.64	[−1.27, 0.00]	.15
Interleaved	—	—	—	−0.05	[−0.66, 0.57]	1.00
Clause-final pause duration						
Blocked	−0.37	[−0.94, 0.21]	.62	−0.78*	[−1.4, −0.13]	.05
Interleaved	—	—	—	−0.41	[−1.02, 0.22]	.58
Mid-clause pause frequency						
Blocked	0.31	[−0.25, 0.86]	.84	0.73+	[0.09, 1.35]	.07
Interleaved	—	—	—	0.42	[−0.19, 1.02]	.51
Clause-final pause frequency						
Blocked	0.56	[−0.02, 1.13]	.16	0.61	[−0.03, 1.23]	.18
Interleaved	—	—	—	0.05	[−0.55, 0.65]	1.00

Note. A positive effect size indicates a higher value in the condition in the row, whereas a negative effect size indicates a higher value in the condition in the column.
$^{+}p < .10$. $^{*}p < .05$. (Bonferroni-corrected *p* values)

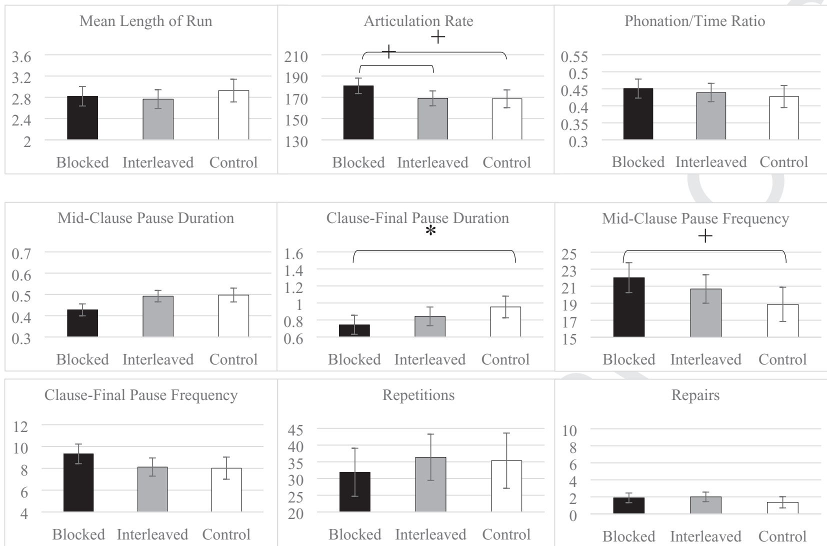

Figure 3 Adjusted mean posttest scores for blocked, interleaved and control conditions. Mean posttest scores were adjusted for the pretest scores (i.e., covariate). The error bars represent 95% confidence intervals. All raw pretest and posttest scores are presented in Appendix S7 in the online supplementary file.
$+p < .10$. $* \, p < .05$.

Mid-Clause Pause Duration

Multiple pairwise comparisons (Table 4) showed that the blocked-practice group decreased their mid-clause pause duration more than both the interleaved-practice group and the control group did, all with meaningful effect sizes. No meaningful difference was noted between the interleaved-practice and the control groups.

Clause-Final Pause Duration

Similarly, the clause-final pause duration decreased more in the blocked-practice group relative to the interleaved-practice group, with a slightly smaller effect size than the benchmark (and it included 0 in the 95% CI), and relative to the control group with a meaningful effect size. Moreover, the interleaved-practice group made shorter pauses than the control group did with a meaningful effect size (albeit 0 was included in the 95% CI).

Mid-Clause Pause Frequency

According to the multiple pairwise comparisons, the blocked-practice group paused more frequently mid-clause than both the interleaved-practice group with a smaller effect size than the benchmark (which included 0 in the 95% CI) and the control group with a meaningful effect size. In addition, the interleaved-practice group paused more frequently than the control group with a meaningful effect size (although 0 was included in the 95% CI).

Clause-Final Pause Frequency

Similarly, higher clause-final pause frequency was noted for the blocked-practice group relative to both the interleaved-practice group and the control group. Although both effect sizes were considered meaningful, caution should be used for any interpretations given that the bounds of 95% confidence interval included 0. No meaningful difference was detected between the interleaved-practice and blocked-practice groups.

Discussion

In the training period, blocked practice was shown to be significantly superior to interleaved practice for six of the nine fluency measures, with meaningful effect sizes ($d > 0.40$) on at least 2 of the 3 days. Although five measures supported the greater effectiveness of blocked practice (longer mean length of run, faster articulation rate, higher phonation/time ratio, shorter mid-clause pause duration, and shorter clause-final pause duration), the clause-final pause frequency increased more in blocked practice than in interleaved practice during training.

In terms of the pretest–posttest fluency changes, blocked practice led to faster articulation rate and shorter mid-clause pause duration than did interleaved practice, both beyond the treatment-specific benchmark ($d = 0.66$ and -0.60, respectively). The shorter clause-final pause duration noted for blocked practice in comparison to interleaved practice was slightly below the benchmark ($d = -0.37$). Similar to the training data, clause-final pause frequency at the posttest was higher in the blocked-practice than in the interleaved-practice condition ($d = 0.56$). These results for clause-final pauses need to be interpreted with caution as the 95% confidence intervals of the effect sizes included 0. Although a similar pattern was also noted in mid-clause pause frequency, the group difference was not considered meaningful ($d = 0.31$) according to the previously established criterion and thus has not been considered further in the discussion. Virtually no meaningful differences were observed consistently in the repair fluency measures (repetition[8] and repairs) either during the training phase or between pretest and posttest, possibly because repair fluency is less closely related to L2 proceduralization of linguistic knowledge.

The remainder of this section is organized in six parts. First, an attempt is made to reconcile the divergence between the current findings and those yielded by previous L2 and psychology studies from several theoretical perspectives. Second, the advantage of blocked practice relative to interleaved practice, which was particularly pronounced for two fluency measures (articulation rate and mid-clause pause duration), is discussed. Third, some intriguing patterns in clause-final pause duration and frequency are highlighted. Fourth, the lack of transfer of a training effect to the posttest performance is evaluated in relation to two fluency measures, mean length of run and phonation/time ratio. Fifth, limitations of the current study are delineated, thus providing directions for future research. Last, tentative pedagogical implications of the current findings are presented.

Contrasting Current Findings With Those From Previous L2 and Psychology Research

In the present study, blocked practice was shown to be superior to interleaved practice with meaningful effect sizes for articulation rate and mid-clause pause duration, countering the findings reported by other researchers for L2 grammar learning (Nakata & Suzuki, 2019b; Pan et al., 2019; Suzuki & Sunada, 2019) as well as those yielded by most studies in the field of psychology (e.g., Brunmair & Richter, 2019; Kang, 2016). This incongruence is to be expected to some extent, given the differences between the nature of the practice in those previous studies and the fluency training in the current study. For L2 grammar learning (e.g., tense–aspect–mood distinction), the discriminative contrast hypothesis (Kang & Pashler, 2012) and the sequential attention theory (Carvalho & Goldstone, 2017) can explain the benefit of interleaved practice for classification and discrimination of similar categories. In contrast, fluency development does not necessarily involve such learning processes, possibly because it is less reliant on the ability to distinguish between similar skills and/or categories that are being interleaved.

The contextual interference effect (e.g., Brady, 2004) and the desirable difficulty hypothesis (Bjork, 2018; Suzuki et al., 2019a) may thus be more useful in the present context (as discussed in the Background Literature section). Both accounts stipulate that the quality of learning processes during a training phase determines the transferability of acquired skills. It is possible that the interleaving practice was less effective in this study because the training content was too difficult for the participating learners, and learning was suboptimal.

The narrative task might have already been challenging enough for the participants in this study. Therefore, a more difficult, interleaved-practice condition could have overwhelmed these participants, which would in turn have hindered development of their procedural knowledge. This may be in part due to insufficient practice opportunities because fluency training was provided over a very short period. Research on motor skills has demonstrated that a significant number of practice trials is necessary for the interleaving effects to emerge, but blocked practice is more advantageous when relatively fewer practice trials are performed (de Croock & Van Merriënboer, 2007; Shea, Kohl, & Indermill, 1990). If more opportunities to practice oral narrative tasks had been provided in this study, the advantage of interleaved practice might have been observed. In addition to the amount of practice, blocked and interleaved practice were implemented across multiple learning sessions in the current study, which might have contributed to the divergence from the previous L2 and psychology research findings from studies that were conducted in single sessions. Furthermore, many other variables may potentially moderate the interleaving effect, such as task complexity, skill types, and individual difference variables (Suzuki et al., 2020).

Blocked Practice May Facilitate Fluency Development: Articulation Rate and Mid-Clause Pause Duration

In the current investigation, the benefits of blocked practice exceeded the treatment-specific benchmark on the two fluency measures (articulation rate and mid-clause pause duration) both during training and in pretest–posttest comparisons. Because these two fluency measures presumably reflect underlying L2 proceduralization in the linguistic formulation phase of L2 speech processes (e.g., as suggested by Kahng, 2017; Kormos, 2006; Saito et al., 2018), blocked practice might have allowed the participants to effectively engage in proceduralization of linguistic knowledge and skills during training.

Although the exact nature of the cognitive processes conferred by blocked practice is presently unclear, it is speculated that certain linguistic constructions were activated during the first task performance and were then re-used in the subsequent performance to a greater extent through immediate blocked practice than through interleaved practice. Interestingly, as indicated by the significant Condition × Time interactions in the two-way mixed ANCOVAs for the training data, the advantages of blocked practice over interleaved practice on articulation rate and mid-clause pause duration tended to be more pronounced on the third performance than on the second performance within each session (see Table 2). It is thus likely that, through repeating the identical task three times during the same training session, blocked practice gradually allowed the participants to effectively engage in proceduralization of linguistic knowledge and skills in the formulator for fluent L2 speech production. Consequently, the effects of repeatedly practicing the same task might have carried over to the posttest task, resulting in improvements exhibited by the participants assigned to the blocked-practice condition. In contrast, the participants in the interleaved-practice condition could not have taken advantage of their previous task performance (e.g., recent activation of relevant linguistic knowledge) when engaging in the next narrative practice as it was based on a different prompt.

The current findings align with those reported by de Jong and Perfetti (2011), who demonstrated that the same repetition (AAA−BBB−CCC), which essentially corresponds to blocked practice in the current study, better facilitated L2 fluency development than did the repetition of different speech topics (ABC−DEF−GHI). However, there was an important difference between the two studies: In de Jong and Perfetti's study, all tasks in the latter condition were different to each other, but in the present study identical tasks were used and repeated the same number of times across both blocked- and interleaved-practice conditions. Hence, the findings yielded in the current study extend those reported by de Jong and Perfetti (2011) in that they indicate that merely changing the order of task repetition impacts the efficacy of task-repetition practice transfer. It can thus be tentatively postulated that repetition of an identical task (blocking) plays a crucial role in facilitating transfer of fluency skills acquired during training to a new narrative task performance. Given the small number of participants in this study, further research should be conducted to elucidate the effectiveness of blocked and interleaved practice in L2 speaking training, instead of drawing definitive conclusions from the findings yielded by this single experiment.

Potential Developmental Signature of L2 Fluency: Clause-Final Pause Frequency and Duration

In the present study, blocked practice led to higher clause-final pause frequency than did interleaved practice both during training and in the pretest–posttest comparison. This is somewhat surprising because blocked practice, which was found to be effective in promoting proceduralization, led to seemingly degraded performance as measured by clause-final frequency. To the best of the author's knowledge, a steady increase in pause frequency both during the fluency training phase and between the pretest and posttest has never been documented in prior fluency research. Exploring these potentially interesting patterns may be worthwhile for future research.

This unexpected finding could be interpreted as indicating the superiority of interleaved practice for this specific aspect of fluency. However, the participants assigned to the blocked-practice condition arguably improved fluency in the sense that they were able to pause at the appropriate clausal boundary to plan their speech more effectively. Furthermore, the blocked-practice group tended to pause at the clausal boundary more frequently but for a shorter duration ($d = -0.37$) compared to the interleaved-practice group. This is indicated by a significant negative correlation between clause-final pause frequency and duration in the pretest–posttest comparison for the blocked-practice condition, $r = -.51$, $p = .02$, only, which was absent in both blocked-practice, $r = -.23$, $p = .29$, and control, $r = -.16$, $p = .54$, group. Because a clause-final pause tends to reflect content planning and conceptualization (e.g., Kahng, 2017; Kormos, 2006; Saito et al., 2018), shorter and more frequent clause-final pauses may indicate potential developmental changes in subcomponents of speech processing such as microplanning where syntactic encoding of propositional form of the message is presumably carried out (Levelt, 1989).

No Transfer from Fluency Training to a New Task: Mean Length of Run and Phonation/Time Ratio

The results for the two fluency measures (mean length of run and phonation/time ratio) that yielded the largest effect sizes for group difference during training indicated no meaningful difference between pretest and posttest performance. One potential reason for the lack of transfer of the training effects to improvements between pretest and posttest may be in part related to the nature of these fluency subcomponents. Mean length of run and phonation/time ratio can be construed as a composite measure of speed and breakdown fluency (Bosker et al., 2013). Given the combined nature of multiple aspects of fluency improvement, it might have necessitated a greater number of practice opportunities to transfer their effect to a new task. Another possibility is that mean length of run may depend upon specific lexical and grammatical constructions that relate to the particular task (e.g., formulaic sequences, see for example Tavakoli & Uchihara, 2020). Because the picture prompts in the training were different from those used in the pretest and posttest and because describing them would have required different words and phrases, the limited overlap of linguistic features between the practice and test tasks might have lessened the potential for transferability of this subcomponent of fluency.

In contrast, the improvements evidenced in other fluency components (e.g., articulation rate) might have contributed to superior posttest performance in part due to enhanced lexical retrieval. Because the learners who participated in this study had limited speaking ability, they frequently struggled with lexical retrieval. The lexical retrieval process might be one of the areas that is highly conducive to task-repetition training for leaners at this proficiency level. Furthermore, articulation rate may reflect not only the efficiency of lexical and grammatical encoding but also the execution of articulatory motor gestures (Suzuki & Kormos, 2019). The fact that at least some articulatory motor skills were shared across the training and posttest tasks might have contributed to their transferability. Enhanced articulation processes through blocked practice may have increased the likelihood of transfer from training gains to a posttest performance (to a greater extent than through interleaved practice). These explanations are neither comprehensive nor account for all the measures, but they raise interesting questions regarding knowledge and skill transfer. Hence, it would be beneficial to determine what aspects of fluency are more likely to transfer from one task to another.

Limitations and Directions for Future Research

Several directions for future research can be proposed based on the present study. First and foremost, given the small sample size, it is necessary to replicate the current study using a larger number of participants. Generalizability of the current findings should also be attested with a different sample of L2 learners in a different context (e.g., learners at a higher proficiency level in classroom setting) using different types of speaking tasks (e.g., opinion/instruction tasks, see Lambert et al., 2017; or dialogue tasks, see van Os et al., 2020).

Second, because the posttest was administered only 1 day after the treatment, it is crucial to investigate the durability of training transfer effects in future research using delayed posttests (e.g., de Jong & Perfetti, 2011). Also, although the current findings suggest that short-term blocked practice may be beneficial for fluency development, studying the efficacy of longer-term training (e.g., one semester or longer) is important because extensive practice is required for efficient, automatized speech processing (Saito et al., 2018).

Third, as one of the reviewers pointed out, if the participants completed the tasks at different times of the day and in different settings, this could potentially have introduced extraneous variables that are not considered in the present analysis. Similarly, the prompts employed in the assessments might have imposed different cognitive loads on the participants, even though the picture prompts used for pretest and posttest were counterbalanced. To address this concern, independent-samples t tests were conducted on the nine fluency measures to compare the data from the two picture prompts, as used at pretest. No significant differences were detected ($p > .10$ for all measures), suggesting the two prompts elicited similar data. Nonetheless, in future investigations, it would be beneficial to conduct an independent pilot study on these prompts to confirm their similarity.

Fourth, a more detailed performance analysis during training may highlight other differences between blocked and interleaved practice. For instance, repeated use of the same lexical and N-grams across training (and indeed testing) performances might be examined (de Jong & Tillman, 2018; Tavakoli & Q4 Uchihara, 2019; Saito, 2020) because blocked practice may encourage the use of the same linguistic constructions and thus facilitate L2 proceduralization. According to usage-based perspectives in L2 acquisition (e.g., Ellis & Wulff, 2015), linguistic constructions can range from phonological, lexical, to grammatical structures, with varying levels of abstraction (e.g., *give it to George*, *give* O *to* O, V+ O + *to* O). It would be worthy of examining the overlap of linguistic constructions across the training performances.

Last, authors of future research can simply examine the spacing effects of task repetition on transfer (e.g., AAA [massing] vs. A__A__A [spacing]). To the best of the author's knowledge, only one empirical study has been carried out to systematically investigate the effects of task-based practice distribution for L2 fluency development (Bui, Ahmadian, & Hunter, 2019). That study's findings indicated that the amount of spacing between repeated tasks can influence some aspects of L2 fluency (i.e., speed and breakdown fluency). This suggested that the temporal distribution of repetition practice may also be an important variable in L2 fluency training that is worthy of investigation. Because the practice schedules in the current study (blocked and interleaved) can be interpreted, at least in part, as massed versus spaced configurations, further exploratory analyses were conducted (see Appendix S8 in the online Supporting Information for more details). These reanalyzed results, however, suggest that the pretest-posttest changes were better accounted for by the blocked and interleaved distinction than the massed and spaced distinction. In sum, future research needs to be carried out to systematically explore the effects of temporal spacing as well as sequence of tasks. This line of research can potentially elucidate the optimal conditions for L2 proceduralization and fluency development that are transferable to different tasks.

Pedagogical Implications

Researchers have just started to improve their understanding of optimal task-repetition conditions for proceduralization and skill transfer from training to a new task. This understanding is directly relevant to L2 teachers' and learners' pedagogical decisions about the most beneficial task sequencing. All other aspects (e.g., the amount of practice) being equal, at least two variables play a key role in proceduralization and successful skill transfer: (a) the amount of variation in task types and (b) the sequence of tasks (Kerr & Booth, 1978; Willey & Liu, 2018). Four possible task-repetition practice sequences are proposed below:

1. AAAAAAAAA (no task variability, blocked)
2. AAABBBCCC (moderate task variability, blocked)
3. ABCABCABC (moderate task variability, interleaved)
4. ABCDEFGHI (high task variability, interleaved)

According to findings of the current study and the findings reported by de Jong and Perfetti (2011), Sequence 2 has been found to be more effective than Sequences 3 and 4. Because no research has been conducted to compare the first sequence with the remaining three, the relative effectiveness of a practice sequence without any variability is unknown, but it may be difficult to sustain learners' motivation for repeating the same task many times without some extra affective and motivational support (see Lambert et al., 2017, for further discussion). One pedagogical implication of the present study could be a tentative suggestion that moderate variability in tasks, such as that given in Sequence 2, should be provided for effective fluency development through task repetition.

Conclusions

The chief objective of this study was to determine and to contrast the effects of blocked and interleaved practice on L2 fluency development and proceduralization. Contrary to previous findings indicating an advantage for interleaved practice in a variety of domains including L2 grammar learning (Nakata & Suzuki, 2019b; Pan et al., 2019; Suzuki & Sunada, 2019; Suzuki et al., 2020), the results obtained in the current study suggest that blocked practice can facilitate L2 fluency development and proceduralization (articulation rate and mid-clause pause duration) more effectively than interleaved practice. Furthermore, the participants in the blocked-practice condition exhibited interesting patterns in clause-final pausing behaviors (i.e., increased frequency yet shorter duration over time), which may signal developmental changes in the underlying speech processing. Learning gains were found in some aspects of fluency (mean length of run and phonation/time ratio) during training, but they did not reliably result in pretest–posttest improvements.

From a broader perspective, the current study demonstrates that insights from research on optimal and systematic L2 practice (see Suzuki et al., 2019b for more detail) can inform efficient task-repetition practice in task-based language teaching and learning. Synergy between interdisciplinary lines of inquiry (into L2 practice informed by cognitive psychology and task-based language teaching) can generate new research questions and yield important findings for informing L2 teaching and learning.

Final revised version accepted 6 June 2020

Notes

1 Practically, the distinction between massed and spaced practice overlaps with that between blocked and interleaved practice to a large extent. When learners undertake three tasks in the interleaved schedule, each task will be repeated after completing the two other tasks, thus resulting in a spaced schedule (e.g., A[BC]A[BC]A[BC]A). When the repeated practice of one task type is not interspersed with other tasks (blocked schedule), this format essentially corresponds to a massed schedule for that particular task (e.g., AAAABBBBCCCC).

2 The current study focused primarily on blocked and interleaved practice (rather than massed and spaced practice). Because blocked and interleaved practice are closely related to massed and spaced practice, the differences are addressed in more detail in the Discussion section and in Appendix S8 in the online Supporting Information.

3 The Common European Framework of Reference for Languages levels were estimated based on their English certificates such as Test of English for International Communication (TOEIC®) scores (https://www.ets.org/toeic).

4 The participants in the control group were not part of the random group assignment because only the blocked-practice and interleaved-practice group members volunteered to engage in the training outside their regular classes.

5 One of the reviewers pointed out that the interleaving condition did not incorporate randomization (e.g., ABC-CBA-CAB, rather than ABC-ABC-ABC) even though randomization has been typically used in previous studies on interleaved practice. Unpredictability stemming from randomization may be an important variable in the effectiveness of interleaved training schedules (as argued by Pan, Lovelett, Phun, & Rickard, 2019).

6 The treatment-specific effect sizes (Cohen's *d*s) concerned the pretest–posttest changes (rather than training data) for which blocked and interleaved practice were compared, because the posttest in the current study was administered 1 day after the treatment.

7 According to the univariate ANOVAs, there was no significant main effect of condition for the pretest fluency measures ($p > .10$).

8 The blocked-practice group significantly reduced repetitions than did the interleaved-practice group only on Day 2–2.

Open Research Badges

This article has earned an Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at https://www.iris-database.org. All proprietary materials have been precisely identified in the manuscript.

References

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: Does it transfer to a new task? *The Language Learning Journal*, *39*, 269–280. https://doi.org/10.1080/09571736.2010.545239

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, *15*, 35–59 https://doi.org/10.1177/1362168810383329

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

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. https://doi.org/10.1177/1745691617690642

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

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*, 159–175. https://doi.org/10.1177/0265532212455394

Brady, F. (2004). Contextual interference: A meta-analytic study. *Perceptual and Motor Skills*, *99*, 116–126. https://doi.org/10.2466/pms.99.1.116-126

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*, 1029–1052. https://doi.org/10.1037/bul0000209

Bui, G., Ahmadian, M. J., & Hunter, A. M. (2019). Spacing effects on repeated L2 task performance. *System*, *81*, 1–13. https://doi.org/10.1016/j.system.2018.12.006

Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 136–146). Oxford, UK: Heinemann.

Q5

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In P. S. M. Bygate & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23–48). Harlow, UK: Pearson Longman.

Bygate, M. (2018). *Learning language through task repetition* (Vol. 11). Amsterdam, The Netherlands: John Benjamins.

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*, *43*, 1699–1719. https://doi.org/10.1037/xlm0000406

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. https://doi.org/10.1037/0033-2909.132.3.354

de Croock, M. B. M., & Van Merriënboer, J. J. G. (2007). Paradoxical effects of information presentation formats and contextual interference on transfer of a complex cognitive skill. *Computers in Human Behavior*, *23*, 1740–1761. https://doi.org/10.1016/j.chb.2005.10.003

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*, 533–568. https://doi.org/10.1111/j.1467-9922.2010.00620.x

de Jong, N., & Tillman, P. C. (2018). Grammatical structures and oral fluency in immediate task repetition: Trigrams across repeated performances. In M. Bygate (Ed.), *Task-based language teaching* (Vol. 11, pp. 43–73). Amsterdam, The Netherlands: John Benjamins.

De Jong, N., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*, 387–404. https://doi.org/10.1177/1362168815606161

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*, 113–132. https://doi.org/10.1515/iral-2016-9993

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. https://doi.org/10.1017/S0142716412000069

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*, 385–390. https://doi.org/10.3758/BRM.41.2.385

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). Task repetition for language learning: A perspective from skill acquisition theory. In M. Bygate (Ed.), *Learning language through task repetition* (pp. 27–42). Amsterdam, The Netherlands: John Benjamins.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*, 533–557. https://doi.org/10.1017/s0272263109990015

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work: Journal of Prevention, Assessment & Rehabilitation*, 159–165.

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

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

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375. https://doi.org/10.1093/applin/21.3.354

Freed, B. F., Segalowitz, N. S., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, *26*, 275–301. https://doi.org/10.1017/S0272263104262064

Fukuta, J. (2016). Effects of task repetition on learners' attention orientation in L2 oral production. *Language Teaching Research*, *20*, 321–340. https://doi.org/10.1177/1362168815570142

Goto, K., Maki, H., & Kasai, C. (2010). The minimal English test: A new method to measure English as a second language proficiency. *Evaluation & Research in Education*, *23*, 91–104. https://doi.org/10.1080/09500791003734670

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning*. San Diego, CA: Academic Press.

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

Heaton, J. B. (1996). *Composition through pictures*. Essex, UK: Longman.

Horvath, J. C., Lodge, J. M., & Hattie, J. (2016). *From the laboratory to the classroom: Translating science of learning for teachers*. New York, NY: Routledge.

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

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*, 809–854. https://doi.org/10.1111/lang.12084

Kahng, J. (2017). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*, 569–591. https://doi.org/10.1017/S0142716417000534

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. https://doi.org/10.1002/acp.1801

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

Kerr, R., & Booth, B. (1978). Specific and varied practice of motor skill. *Perceptual and Motor Skills*, *46*, 395–401. https://doi.org/10.1177/003151257804600201

Kim, Y., & Tracy-Ventura, N. (2013). The role of task repetition in L2 performance development: What needs to be repeated during task-based interaction? *System*, *41*, 829–840. https://doi.org/10.1016/j.system.2013.08.005

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

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*, 167–196. https://doi.org/10.1017/S0272263116000085

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

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

Lightbown, P. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding second language process* (pp. 27–44). Clevedon, UK: Multilingual Matters.

Lynch, T., & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research*, *4*, 221–250. https://doi.org/10.1177/136216880000400303

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York, NY: Routledge.

Mclaughlin, B. (1987). *Theories of second-language learning*. London, UK: Routledge.

Merbah, S., & Meulemans, T. (2011). Learning a motor skill: Effects of blocked versus random practice: A review. *Psychologica Belgica*, *51*, 15–48. https://doi.org/10.5334/pb-51-1-15

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

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. https://doi.org/10.1017/S0272263114000825

Nakata, T., & Suzuki, Y. (2019a). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, *41*, 287–311. https://doi.org/10.1017/S0272263118000219

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

Oba, T. (2018). *The effects of form-focused practice and corrective feedback on proceduralizing simple and complex L2 grammar during EFL task-based interaction*. (Unpublished doctoral dissertation). McGill University, Montreal, QC, Canada.

Pan, S. C., Lovelett, J., Phun, V., & Rickard, T. C. (2019). The synergistic benefits of systematic and random interleaving for second language grammar learning. *Journal of Applied Research in Memory and Cognition*, *8*, 450–462. https://doi.org/10.1016/j.jarmac.2019.07.004

Pan, S. C., Tajran, 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*, *111*, 1172–1188. https://doi.org/10.1037/edu0000336

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

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*, 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

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

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*, 355–367. https://doi.org/10.1007/s10648-012-9201-3

Saito, K. (2020). Multi- or single- word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*, 548–588. https://doi.org/10.1111/lang.12387

Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics*, *39*, 593–617. https://doi.org/10.1017/S0142716417000571

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

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. https://doi.org/10.1002/tesq.445

Shea, C. H., Kohl, R., & Indermill, C. (1990). Contextual interference: Contributions of practice. *Acta Psychologica*, *73*, 145–157. Retrieved from http://doi.org/10.1016/0001-6918(90)90076-R

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*, 1–14. https://doi.org/10.1017/S026144480200188X

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics*, *30*, 510–532. https://doi.org/10.1093/applin/amp047

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*, 97–111. https://doi.org/10.1515/iral-2016-9992

Suzuki, S., & Kormos, J. (2019). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*, 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, Y. (2011). Why can they write coherently? Focusing on syntactic processing speed. [どうして『つながりのある文章』が書けるのか–文法処理速度に焦点を当てて–]. *STEP Bulletin*, *23*, 30–52.

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

Suzuki, Y. (in press). The cognitive approach. In T. Gregersen & S. Mercer (Eds.), *The Routledge handbook of psychology of language learning*. New York, NY: Routledge.

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

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

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

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*, *42*, 169–197. https://doi.org/10.1017/S0272263119000470

Suzuki, Y., Yokosawa, S., & Aline, D. (2020). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research*, https://doi.org/10.1177/1362168820913985 Q6

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson Education.

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*, 447–471. https://doi.org/10.1002/tesq.244

Tavakoli, P., & Hunter, A.-M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research*, *22*, 330–349. https://doi.org/10.1177/1362168817708462

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*, 506–547. https://doi.org/10.1111/lang.12384

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly*, *50*, 369–393. https://doi.org/10.1002/tesq.232

van Os, M., de Jong, N. H., & Bosker, H. R. (2020). Fluency in dialogue: Turn-taking behavior shapes perceived fluency in native and nonnative speech. *Language Learning*, https://doi.org/10.1111/lang.12416

Willey, C. R., & Liu, Z. (2018). Long-term motor learning: Effects of varied and specific practice. *Vision Research*, *152*, 10–16. https://doi.org/10.1016/j.visres.2017.03.012

Zuniga, M., & Simard, D. (2019). Factors influencing L2 self-repair behavior: The role of L2 proficiency, attentional control and L1 self-repair behavior. *Journal of psycholinguistic research*, *48*, 43–59. https://doi.org/10.1007/s10936-018-9587-2

Supporting Information

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

Appendix S1. Counterbalancing of Practice Tasks.
Appendix S2. The Number of Participants Identified as an Outlier for Each Fluency Measure.
Appendix S3. Results of Mixed ANCOVAs on Training Data.
Appendix S4. Descriptive Statistics for Training Data.

Appendix S5. Effect Sizes Between Blocked and Interleaved Conditions (Training Data).
Appendix S6. Results of Univariate ANCOVAs.
Appendix S7. Descriptive Statistics for Pretest and Posttest.
Appendix S8. Alternative Accounts: Exploratory Analysis from Perspective of Spacing Effects.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. Article accepted in *Language Learning* on 6 June 2020.

To Develop Fluency, What Is the Best Schedule for Recycling the Same Speaking Task?

What This Research Was About and Why It Is Important

To speak a second language (L2) fluently, repeating the same speaking task can be effective. However, we do not know whether it is better to repeat exactly the same task in a “block”, or mix up (“interleave”) different tasks to spread out the repetition. This study compared the effectiveness of blocked and interleaved task repetition schedules. Japanese university students provided an oral narration of cartoons, under either a blocked (Day 1: Task A-A-A, Day 2: Task B-B-B, Day 3: Task C-C-C) or interleaved (Day 1: Task A-B-C, Day 2: Task A-B-C, Task Day 3: A-B-C) schedule. The blocked repetition was more effective than interleaved repetition for fluency development.

What the Researcher Did

- Participants were 68 Japanese university students, with at least 6 years’ experience of studying English.
- Participants were assigned to either a blocked or an interleaved practice group. They engaged in speaking training outside of their regular classes, performing oral narrative tasks in which they described six-frame cartoon stories, three times a day for 3 consecutive days.
- A control group participated only in the pretest–posttest sessions.
- In the blocked practice, participants narrated Cartoon A three times on day 1, Cartoon B three times on day 2, and Cartoon C three times on day 3.
- In the interleaved practice, participants narrated different cartoons (Cartoon A, B, and C) once every day.

- Fluency development was measured by a pretest and a posttest (six-frame cartoon stories that were different to the training), administered 1 day after the final (i.e., third) speaking training session.

What the Researcher Found

- During the training, the blocked practice group performed the narration more fluently (e.g., faster articulation, fewer pauses within clauses), particularly on the third performance each day, compared to the interleaved practice groups.
- On the posttest, blocked practice resulted in greater fluency development (faster articulation and shorter pauses within clauses) than interleaved practice. Blocked practice also resulted in pausing more frequently *before* a clause, which might have served fluency *during* the clause.

Things to Consider

- In the current study, the same tasks (Cartoon A, B, C) were repeated exactly three times in both blocked and interleaved practice schedules. However, changing the order and distribution of task repetition can be done in different ways, and these can impact the efficacy of speaking training. When deciding how to distribute tasks, two factors can be considered: (a) task variability and (b) task sequence. For instance, there are four possible schedule options:

    1. Task AAA-AAA-AAA (no task variability, blocked)
    2. Task AAA-BBB-CCC (moderate task variability, blocked)
    3. Task ABC-ABC-ABC (moderate task variability, interleaved)
    4. Task ABC-DEF-GHI (high task variability, interleaved)

The current study showed that the second option was better than the third for fluency development.

Materials and data: Materials are publicly available at https://iris-database.org.

How to cite this summary: Suzuki, Y. (2020). To develop fluency, what is the best schedule for recycling the same speaking task? *OASIS Summary* of Suzuki (2021) in *Language Learning*. https://oasis-database.org/

Graphical Abstract

The contents of this page will be used as part of the graphical abstract of html only. It will not be published as part of main article.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. Article accepted in *Language Learning* on 6 June 2020.

To develop fluency, what is the best schedule for recycling the same speaking task?

What this research was about and why it is important

To speak a second language (L2) fluently, repeating the same speaking task can be effective. However, we do not know whether it is better to repeat exactly the same task in a 'block', or mix up ('interleave') different tasks to spread out the repetition. This study compared the effectiveness of blocked and interleaved task repetition schedules. Japanese university students provided an oral narration of cartoons, under either a blocked (Day 1: Task A-A-A, Day 2: Task B-B-B, Day 3: Task C-C-C) or interleaved (Day 1: Task A-B-C, Day 2: Task A-B-C, Task Day 3: A-B-C) schedule. The blocked repetition was more effective than interleaved repetition for fluency development.

What the researcher did

- Participants were 68 Japanese university students, with at least 6 years' experience of studying English.
- Participants were assigned to either a blocked or an interleaved practice group. They engaged in speaking training outside of their regular classes, performing oral narrative tasks in which they described six-frame cartoon stories, three times a day for three consecutive days.
- A control group participated only in the pretest-posttest sessions.
- In the blocked practice, participants narrated Cartoon A three times on day 1, Cartoon B three times on day 2, and Cartoon C three times on day 3.
- In the interleaved practice, participants narrated different cartoons (Cartoon A, B, and C) once every day.
- Fluency development was measured by a pretest and a posttest (six-frame cartoon stories that were different to the training), administered one day after the final (i.e., third) speaking training session.

What the researcher found

- During the training, the blocked practice group performed the narration more fluently (e.g., faster articulation, fewer pauses within clauses), particularly on the third performance each day, compared to the interleaved practice groups.
- On the posttest, blocked practice resulted in greater fluency development (faster articulation and shorter pauses within clauses) than interleaved practice. Blocked practice also resulted in pausing more frequently *before* a clause, which might have served fluency *during* the clause.

- **Things to consider**
  - In the current study, the same tasks (Cartoon A, B, C) were repeated exactly three times in both blocked and interleaved practice schedules. However, changing the order and distribution of task repetition can be done in different ways, and these can impact the efficacy of speaking training. When deciding how to distribute tasks, two factors can be considered: (a) task variability and (b) task sequence. For instance, there are four possible schedule options:
    1. Task AAA-AAA-AAA (no task variability, blocked)
    2. Task AAA-BBB-CCC (moderate task variability, blocked)
    3. Task ABC-ABC-ABC (moderate task variability, interleaved)
    4. Task ABC-DEF-GHI (high task variability, interleaved)

    The current study showed that the second option was better than the third for fluency development.

Materials and data: Materials are publicly available at https://iris-database.org.

How to cite this summary: Suzuki, Y. (2020). To develop fluency, what is the best schedule for recycling the same speaking task? *OASIS Summary* of Suzuki (2021) in *Language Learning*. https://oasis-database.org/