Studies in Second Language Acquisition, **44** (2022), 536–561
doi:10.1017/S0272263121000358

Research Article

MASSED TASK REPETITION IS A DOUBLE-EDGED SWORD FOR FLUENCY DEVELOPMENT

AN EFL CLASSROOM STUDY

Yuichi Suzuki*

Kanagawa University

Keiko Hanzawa

Tokyo University of Science

Abstract

To examine the effects of task repetition with different schedules, English-as-a-foreign-language classroom learners performed the same oral narrative task six times under three different schedules. They narrated the same six-frame cartoon story (a) six times consecutively in one class (massed practice), (b) three times at the beginning and at the end of a class (short-spaced practice), and (c) three times as a part of two classes 1 week apart (long-spaced practice). The results yielded by an immediate posttest using a novel cartoon showed that massed practice reduced breakdown fluency (mid-clause and clause-final pauses) the most. However, the participants in the massed-practice group showed degraded speed (slower articulation rate) and repair fluency (more verbatim repetition). The effects of repetition schedule seem limited on a 1-week delayed posttest involving a novel cartoon. Yet, when participants narrated the same practiced cartoon 1 week later, massed practice also resulted in more verbatim repetition.

The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at https://www.iris-database.org/iris/app/home/detail?id=york%3a939256

This study was supported by Yokohama Academic Foundation for the first author and Grant-in-Aid for Scientific Research (KAKENHI) from Japan Society for the Promotion of Science (19K13274) to the second author. We would like to express our gratitude to Gavin Bui for sharing his data set for reanalysis. We are also grateful to Atsushi Miura, Chisaki Inaba, Miyu Koyama, Misaki Kuratsubo, Naoya Nishiura, and Kohta Shimada for their dedicated assistance in data coding.

* 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

INTRODUCTION

Task repetition plays a critical role in developing second language (L2) knowledge and skills (Bygate, 2018). In task-based language teaching and learning, the facilitative role of repeating a speaking task was found for fluency development. A body of L2 research on task repetition (e.g., Ahmadian & Tavakoli, 2011; Bygate, 1996, 2001; de Jong & Perfetti, 2011; Lambert et al., 2017; Lynch & Maclean, 2000; Thai & Boers, 2016) has shown that task repetition leads to changes in utterance fluency, reflected as speed (e.g., articulation rate), breakdown (e.g., pauses), and repair fluency (e.g., repetitions).

Task repetition research intersects with the idea of L2 practice—repeated engagement of L2 use in a systematic and deliberate way with a goal of developing automatized knowledge and skills (DeKeyser, 2007; Lyster & Sato, 2013; Suzuki et al., 2019b). One way to enhance L2 learning through repeated practice is by manipulating temporal spacing between practices (e.g., massed vs. spaced practice). Massed practice refers to repeated practice without any temporal intervals between sessions and trials, whereas spaced or distributed practice involves repeating tasks with temporal intervals. The advantage of spaced practice over massed practice is called spacing effect, and the effect of different spacing durations (e.g., short vs. long intervals) is called lag effect. Both spacing and lag effects are collectively termed distributed practice effects (Cepeda et al., 2006). The distributed practice effects—phenomena originally examined extensively in cognitive psychology—are worthy of further exploration in L2 learning research for both theoretical and practical reasons. Theoretically, researchers can assess the extent to which the findings obtained in the field of cognitive psychology can be translated to multifaceted aspects of L2 learning. Practically, establishing distributed practice effects for certain aspects of L2 learning can help maximize the outcome of repeated practice without changing the total practice time.

Authors of recent L2 studies inspired by cognitive psychology research have investigated how systematically manipulating the timing of repeated practice (e.g., massed vs. spaced schedules) can enhance proceduralization of some aspects of L2 knowledge, such as lexical, pronunciation, and grammar (e.g., Kasprowicz et al., 2019; Li & DeKeyser, 2019; Nakata & Elgort, 2021; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017a). The current study extends the line of investigation into repeated engagement of the same speaking task under different schedules (e.g., Bui et al., 2019; Suzuki, 2021b). In this study, English-as-a-foreign-language (EFL) learners engaged in the same monologue task six times under different schedules (massed, short-spaced, and long-spaced). The goal of this short-term classroom intervention study was elucidating the effects of different task repetition schedules on the development of L2 utterance fluency.

LITERATURE REVIEW

DISTRIBUTED PRACTICE EFFECTS IN COGNITIVE PSYCHOLOGY AND L2 RESEARCH

Cognitive psychology research has yielded a substantial body of knowledge regarding distributed practice effects (e.g., Cepeda et al., 2006; Toppino & Gerbier, 2014). Spacing effect (the advantage of spaced over massed learning) is a robust finding for diverse forms of learning in a variety of subjects (e.g., mathematics, verbal memory, and spelling). In contrast, extant findings on lag effects are equivocal in cognitive psychology research

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

because the optimal lag is influenced by a variety of factors (Cepeda et al., 2008; Rohrer, 2015). Several theories have been proposed to account for the distributed practice effects (see Toppino & Gerbier, 2014 for a review). According to the study-phase retrieval theory (e.g., Toppino & Bloom, 2002), for instance, learning improves the most when a learner makes most effort to retrieve a previously learned item in a repeated practice session. In other words, the repeated performance should *not* be too short or massed (i.e., little effort exercised to retrieve previously learned materials) or too long (i.e., failure to retrieve previously learned materials). This view is tied to Bjork's (1994) desirable difficulty framework, which stipulates that, when practice challenges leaners to bring out maximal effort, it can promote robust learning and retention (Suzuki et al., 2019a, 2020). Therefore, creating a desirably difficult situation through optimal spacing is important to make repeated practice most effective.

Research on distributed practice effects has also flourished in the L2 field over the last decade. Many researchers have demonstrated that spacing effects seem also applicable to L2 vocabulary and grammar learning. In deliberate L2 vocabulary learning in paired associate format, for instance, by inserting temporal spacing between retrieval trials, retention of vocabulary knowledge can increase from 160% to 250% relative to massed practice (e.g., Nakata, 2015; Nakata & Suzuki, 2019). In an empirical study on L2 grammar learning (Miles, 2014), spaced practice also seems more effective than massed practice at least for the acquisition of receptive grammatical knowledge.

A growing number of authors have conducted empirical studies to investigate the optimal lags for different aspects of L2 learning: pronunciation (Li & DeKeyser, 2019), vocabulary (Nakata, 2015; Rogers & Cheung, 2020a, 2020b; Serrano & Huang, 2018), and grammar (Bird, 2010; Kasprowicz et al., 2019; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017a). In contrast to the spacing effects, the findings pertaining to lag effects are inconsistent. For instance, within the domain of L2 grammar learning, some studies show the superiority of longer-spaced practice (Bird, 2010; Rogers, 2015), while other researchers claim that shorter-spaced practice may be as effective as longer-spaced practice (Kasprowicz et al., 2019) or even more effective than longer-spaced practice (Suzuki, 2017; Suzuki & DeKeyser, 2017a).

With the aforementioned desirable difficulty framework in mind, two L2-related factors are identified as potential moderators of distributed practice effects in L2 learning (cf., Suzuki et al., 2019a). First, complexity involved in information processing required for a given task has been found to be an important factor influencing distributed practice effects (Donovan & Radosevich, 1999). For instance, when learning complexity is relatively high (e.g., describing a picture orally using vocabulary and grammar rules involves more complex psycholinguistic processes than remembering L2 words in paired associate formats), the benefits of (longer-)spaced practice may be attenuated due to the failure to retrieve previously learned information. Second, according to skill acquisition (DeKeyser, 2020) and retention theories (Kim et al., 2013), acquisition of declarative and procedural knowledge involves different learning processes. Declarative knowledge can be acquired even with one encounter, whereas procedural knowledge acquisition necessitates repeated use of target knowledge and skills. Accordingly, because proceduralization occurs gradually over multiple learning sessions, it may be more effective to immediately repeat the activity before the skill and knowledge decays. In other words, more concentrated repetition may be a more efficient way for learners to fine-tune their

procedural knowledge in comparison with longer-spaced repetition. Based on these two key factors that moderate L2 learning difficulty, it is reasonable to assume that the optimal distribution of practice varies depending on L2 tasks and skills to be acquired (Rogers, 2017). To further our understanding of distributed practice effects, investigations into speaking skill acquisition are important because L2 speaking involves demanding mental processes, such as conceptualization, formulation, and articulation (Levelt, 1989) and relies on procedural knowledge for real-time speech processing (Kahng, 2014; Kormos, 2006; Suzuki, 2021b). The effects of speaking task repetition on L2 fluency have been examined in a wide range of task-based language learning studies, which will be reviewed next.

RESEARCH ON TASK REPETITION FOR L2 FLUENCY DEVELOPMENT

Task repetition enhances L2 fluency development (e.g., Ahmadian & Tavakoli, 2011; Bygate, 1996, 2001; de Jong & Perfetti, 2011; Lambert et al., 2017; Lynch & Maclean, 2000; Thai & Boers, 2016). The benefits of repeating the exact same task for enhancing fluency can be attributed to two phenomena. First, when L2 learners repeat the same task (e.g., oral narrative), they can free up their attentional resources for conceptualization (e.g., generating the content of speech) and use these additional resources for formulating accurate and fluent speech involving linguistic encoding and articulation at a subsequent performance (Bygate, 1996; Fukuta, 2016). Second, L2 learners who perform the same task again have presumably activated linguistic expressions they had produced in the first performance. This idea may be consistent with the view that the priming mechanisms (Bock & Griffin, 2000) support learning through task repetition for formulating and producing the same and/or similar L2 utterances more efficiently. The words and syntactic patterns that are primed and repeated across consecutive performances may facilitate L2 fluency development (de Jong & Perfetti, 2011; de Jong & Tillman, 2018).

One of the most detailed analyses of speed, breakdown, and repair fluency changes during task repetition was documented by Lambert et al. (2017). In their study, Japanese university EFL learners engaged in a paired speaking task (instruction task, narration task, or opinion task) six times. The performance changes across six repetitions were analyzed in terms of speech rate (the number of pruned syllables), mid-clause and clause-final filled pauses, and the number of repetitions and self-repairs. These analyses revealed a significant steady improvement in all aspects of fluency examined in their study until the fourth or fifth performance. Although Lambert et al. (2017) convincingly showed the benefits of task repetition through pair speaking tasks, the extent to which the effects of task repetition are durable ("retention") and carry over to fluency development that is measured by a posttest involving a different prompt ("transfer") remain unknown.

SYSTEMATIC TASK REPETITION SCHEDULES FOR FACILITATING L2 FLUENCY DEVELOPMENT

In task repetition research, an emerging line of investigations focuses on factors pertinent to systematic task repetition that can assist L2 fluency development, such as increasing the time pressure on task-repetition performance (e.g., Arevart & Nation, 1991), manipulating task type repetition (e.g., Bygate, 2001; Kim & Tracy-Ventura, 2013; Lambert et al., 2021), and variations in tasks to be repeated (de Jong & Perfetti, 2011; Suzuki, 2021b). Although

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

temporal spacing between task repetitions can affect L2 fluency development, little attention has been paid to this factor so far. Across previous empirical studies, the intervals between same-task repetitions vary considerably, such as massed (immediate) repetition (e.g., Lambert et al., 2017; Lynch & Maclean, 2000), across a few days (e.g., Ahmadian & Tavakoli, 2011; Gass et al., 1999; Kim & Tracy-Ventura, 2013), weeks (e.g., Fukuta, 2016), or months (e.g., Bygate, 2001). While the findings yielded by these studies indicate that task repetition facilitates fluency development, there is a paucity of research focusing specifically on the effects of temporal lags between task repetitions on L2 speech fluency development.

To the best of our knowledge, Bui et al. (2019) conducted the first and only study in which the task repetition interval was systematically manipulated in research design to investigate the distributed practice effects in speaking tasks. In their study, EFL learners in Hong Kong engaged in oral picture description task twice under five different schedules (0-day [massed], 1-day, 3-day, 7-day, and 14-day intervals). Their findings indicated that different amounts of spacing influence complexity−accuracy−fluency (CAF) changes from the first (Time 1) to the second (Time 2) speech. Most relevant to the current study on fluency development, the massed group showed the largest gain from Time 1 to Time 2 in their speed fluency (words per minute), whereas 7-day interval group showed the largest reduction in breakdown (filled pauses) and repair fluency (repetitions). According to Bui and collaborators, immediate repetition allowed leaners to recycle the linguistic expressions (e.g., lexical items) for the subsequent performance, as those expressions were primed and accessed more readily. In contrast, the observed reduction in repair fluency could be in part due to the less verbatim repetition between the two performances. Learners in the 7-day interval group might be more likely to use newly formulated messages in the second performance, while maintaining the understanding of the task information process (e.g., content, planning). While Bui et al. (2019) revealed that different task-repetition schedules influence fluency changes from Time 1 to Time 2, their findings also brought to light an important question—to what extent repetition schedule influences the "retention" and "transfer" of the fluency training effect.

THE CURRENT STUDY

To investigate the effects of different task repetition schedules on fluency development, EFL learners narrated the same story involving a six-frame cartoon six times under three different schedules. This was a classroom-based research employing a quasiexperimental research design with four intact English classes at a Japanese university. Based on the number of task repetition in Lambert et al. (2017), the current learners narrated the same story six times with different temporal distribution. The four classes were respectively assigned to a massed practice group (repeating the oral narrative task six times consecutively), a short-spaced practice group (repeating the oral narrative task three times each at the beginning and at the end of class), a long-spaced practice group (repeating the oral narrative task three times in the first and the second week), and a control group. The three experimental conditions are illustrated as follows:

Massed: Task XXXXXX
Short-spaced: Task XXX---45 min---Task XXX
Long-spaced: Task XXX----------7 days----------Task XXX

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

A pretest, immediate posttest, and 1-week delayed posttest, involving different stories from the one used for the practice tasks, were administered to measure the transfer of fluency improvement to different content of the same task type. In addition, the narrative task used in the training phase was also administered after 1 week to measure the retention of the task repetition practice effect. The following three research questions were addressed:

1. What are the effects of three different task-repetition schedules (massed, short-spaced, and long-spaced) on fluency during the training phase?
2. To what extent are the task repetition effects of three different schedules durable after 1 week?
3. To what extent does the task repetition through three different schedules transfer to fluency gains measured by performance on new oral narrative tasks?

METHODS

PARTICIPANTS

The study sample consisted of 79 first-year students at a private university in Japan who had been studying English as a foreign language for at least 6 years before entering university. They were recruited from four intact English classes, which were assigned to the massed ($n = 20$), the short-spaced ($n = 23$), the long-spaced ($n = 21$), or the control group ($n = 15$). Based on their mean score ($M = 421.84$, $SD = 98.88$) on a standardized English test (Test of English for International Communication [TOEIC]), their English proficiency was estimated to fall between A2 (elementary) and B1 (intermediate) level in the Common European Framework of Reference for Language (CEFR) benchmark (Tannenbaum & Wylie, 2008). According to one-way ANOVA results, there were no statistically significant differences in the TOEIC scores among the four groups, $F(3, 75) = 1.57$, $p = .20$, $\eta^2 = .059$.

MATERIALS

Training Material

Two picture prompts (Chase and Surprise) were used for the fluency training in the present study. The prompts were adopted from Heaton (1996) and have been used in many studies on L2 oral production (e.g., de Jong & Tillman, 2018; de Jong & Vercellotti, 2016; Suzuki, 2021b; Tavakoli & Foster, 2008). These two prompts were randomly assigned to each participant according to the seat arrangement in each class.[1] Each of the two prompts consists of a six-frame picture story with a similar narrative structure and little causal reasoning (i.e., the main character is chased by another character and experiences a surprising event in the end). The materials are presented in Appendix A in Online Supplementary File and are available in the IRIS digital repository of data collection instruments (Marsden et al., 2016).

In the training session(s), the participants engaged in the narrative task six times under different practice schedules (see Figure 1). The participants were instructed to narrate the unusual event depicted in the six-frame cartoon to a friend who has not seen the event before. In the instructions, they were told that "Yesterday, you saw an unusual event

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

described in the six-frame cartoon on the following slide. You are going to explain the story to a friend who hasn't seen the event before." Before the first narration, participants listened to a model speech twice to familiarize them with the narrative content and facilitate their narration. Participants were not allowed to take notes while they were listening to the story.

After listening to the model speech, the students were given 90 seconds for planning their narration. During the planning phase, participants were provided with the picture prompt and 13 useful words, along with their Japanese translations. After the preparation time, the participants narrated the story for 120 seconds, aided only by the picture prompt. They had to start their narration by saying "Yesterday, I saw an unusual event." This combination of preparation (90 seconds) and performance (120 seconds) was repeated in all fluency training sessions (i.e., six times). Constant time limit was imposed throughout because increasing time pressure can sometimes reduce the amount of repetition in subsequent performances, which may be considered as a less ideal condition for proceduralization (e.g., de Jong & Perfetti, 2011; de Jong & Tillman, 2018).

Three New Prompts for Pretest, Immediate, and Delayed Posttests

Three six-frame picture prompts (Bicycle, Bus, and Race) that were unfamiliar to the participants were used in the pretest and posttests (see Appendix B in Online Supplementary File). The three prompts were adopted from Heaton (1996) and were similar to the two training picture prompts in terms of the story structure. They also had a tight sequential structure and required little causal reasoning. In the pretest and posttests, the participants were provided with 4–6 guiding questions and a list of 12 useful English words along with their Japanese translations. All test prompts are available in the IRIS digital repository of data collection instruments (Marsden et al., 2016). The pretest and posttests followed the same procedure as the training session, with the exception of the time allocated for the preparation. Because no model speech was presented for the pretest or posttests, additional 60 seconds were allocated for the preparation phase (extending it to 180 seconds, compared to 120 seconds allowed during training). The order of the three test prompts (Bicycle, Bus, and Race) was counterbalanced across participants to minimize the task effects.

PROCEDURE

The study took place in four intact English classes where the second author of this article was the instructor. Figure 1 illustrates the pretest–training–posttest design schedules for each group. In Week 1, all participants took a pretest, as a part of which they individually read the instructions on the computer screen and recorded their narrations at a pace controlled by the instructor.

In Week 2, the training phase started for the three experimental groups. In the massed condition, the participants narrated the same story six times consecutively. In the short-spaced group, participants engaged in the same narrative task three times at the beginning and at the end of the class. It took about a total of 45 minutes for performing the narrative six times. For the remaining 45 minutes, participants engaged in regular class activities (i.e., reading a passage with comprehension questions and a dictation task of the passage), which was not relevant to the training task. In the long-spaced condition, the fluency

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

training session was conducted across 2 weeks (Weeks 2 and 3). The same narrative task was performed three times each week. Immediate posttest was administered after the sixth narrative performance to measure the transfer of fluency training effects on a new narrative task.

One week after the last training session (Week 3 or Week 4), a delayed posttest with a new narrative task was administered for examining the long-lasting transfer of fluency training.[2] Additionally, the same prompt used in the training, denoted as "Narrative (Retention)" in Figure 1, was presented to measure retention of fluency gains in the training task. Participants in the control group only took pretest, immediate posttest, and delayed posttest.

DATA CODING

In total, there were 237 speech samples for the pretest, immediate posttest, and delayed posttest (79 students [4 groups] × 3 times), and 448 samples for the fluency training session(s) and 1-week retention (64 students [3 groups] × 7 times). The first part of narration ("Yesterday, I saw an unusual event") and the final sentence that is not relevant to the story narration (e.g., "Thank you for listening") were removed before fluency coding. The speech samples were annotated using a free sound analysis software PRAAT (Boersma & Weenink, 2016). With the aid of the script developed by de Jong and Wempe (2009), which automatically detects unfilled (silent) pauses of at least 200 millisecond duration, three trained coders manually identified filled and unfilled pauses. They also transcribed the speech samples into Analysis of Speech (AS) units (Foster et al., 2000). Their work was subsequently assessed by the other two coders to ensure the accuracy of all transcriptions. Intercoder reliability was also checked using 20% of the data coded by multiple coders (16 out of 79 participants' data) and was confirmed acceptable for all fluency measures (Cronbach alpha > .85). Consistent with the operationalizations and measurements used in prior studies on this topic (Bui et al., 2019; Lambert et al., 2017),

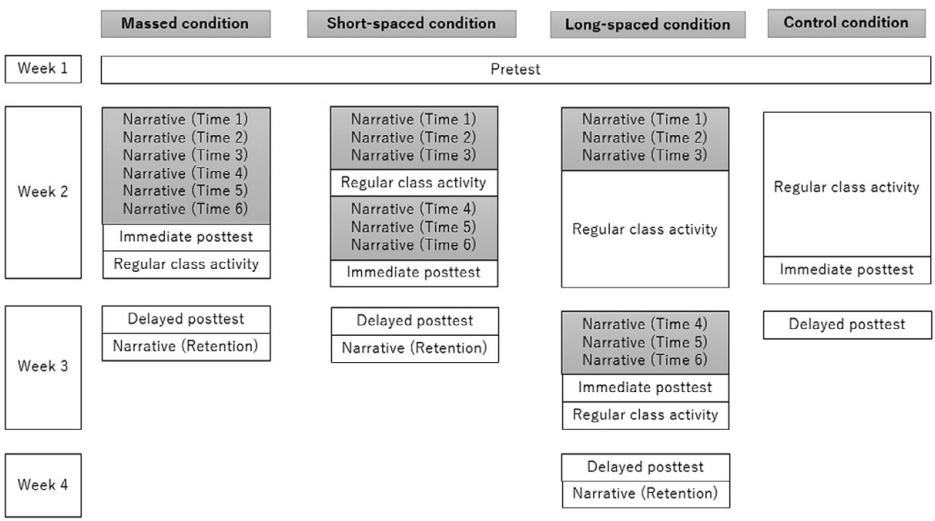

FIGURE 1. Research design.

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

the following seven fluency measures were coded to cover speed, breakdown, and repair fluency (Skehan, 2003):

- Speed fluency
  - i. Articulation rate (the number of syllables per minute of speech, excluding pauses)
- Breakdown fluency
  - ii. Mid-clause pause duration (mean duration of mid-clause filled and unfilled pauses)
  - iii. Clause-final pause duration (mean duration of clause-final filled and unfilled pauses)
  - iv. Mid-clause pause frequency (number of mid-clause filled and unfilled pauses per minute)
  - v. Clause-final pause frequency (number of clause-final filled and unfilled pauses per minute)
- Repair fluency
  - vi. Repetition frequency (number of repetitions per minute)
  - vii. Repair frequency (number of self-repairs per minute)

First, speed fluency is represented by articulation rate because it is relatively independent from the breakdown and repair fluency measures. Second, in breakdown fluency, pause location (mid- vs. final clauses) may be indicative of different cognitive processes (Kahng, 2018). While final-clause pause is often associated with conceptualizing and planning the content of speech, mid-clause pauses are more likely to signal linguistic retrieval and sentence construction in the formulation (Lambert et al., 2017). Therefore, mid-clause and clause-final pauses were coded separately based on AS unit. Third, repair fluency is characterized by how often speakers punctuated with repetitions and repairs. Compared to speed and breakdown fluency, repair fluency may be influenced by factors that are not directly relevant to L2 speaking ability such as L1 speaking style (Zuniga & Simard, 2019). Yet, as they may serve as a proxy for the efficiency of speech monitoring (Hanzawa, 2021) and may be susceptible to intervention (Lambert et al., 2017), repetition and repair counts were included in the analyses.

STATISTICAL ANALYSIS

Training Performance

To compare the changes in training performance across the three experimental groups, a series of two-way mixed ANCOVAs were conducted. Each of the seven fluency measures was used as a dependent variable and Condition and Time were the between-subject variables. Condition was coded as three levels (massed, short-spaced, and long-spaced). To narrow down the scope of analysis, three critical time points (Time 4, Time 6, and Retention [1 week after Time 6]) were included as Time. Time 4 best represents the critical difference in temporal spacing among the three groups. At Time 4, there is no spacing, 45-minute lag, and 1-week lag since Time 3 for the massed, short-spaced, and long-spaced conditions, respectively. Performance at Time 6 and Retention is also of interest, as it allows examining the extent to which the three schedules impact fluency changes at the end of training and 1-week retention, respectively. The interaction between Condition and Time was also included in the model to identify any group differences at different time points. Each fluency measure at Time 1 was included as a covariate to control for potential baseline differences among the three groups.[3]

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

Because the four dependent variables (mid-clause pause duration, clause-final pause duration, clause-final pause frequency, and repetition) were not normally distributed, log-transformation was performed to correct the distributions. No outliers were identified ($z > 3.29$; Tabachnick & Fidell, 2013). When a main effect or interaction was significant in the two-way mixed ANCOVAs, follow-up univariate ANCOVAs were conducted for each performance (i.e., Time 4, Time 6, and Retention) with Condition (massed, short-spaced, and long-spaced) as the between-subject variable for those fluency measures that were significant. The fluency score at Time 1 served as the covariate to control for any differences in the initial training performance. Multiple pairwise comparisons were performed with Bonferroni correction.

In the statistical analyses, the alpha level for statistical significance was set at .05. The effect size magnitudes in ANCOVAs were 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: $\eta_p^2 = .1379$). The effect size of group difference—Hedges's *g*—was computed using the adjusted posttest scores. The estimate from Hedges's *g* is very similar to Cohen's *d* but is more accurate for a sample size below 20. Its magnitude can be interpreted in the same way as Cohen's *d*. Treatment-specific (i.e., spacing effects) benchmark was established for interpreting Hedges's *g*. In a meta-analysis of 63 studies with 112 effect sizes (Donovan & Radosevich, 1999), the overall effect size of spacing effect (comparison between massed vs. spaced practice) was $d = .46$. More relevant to the current study, Bui et al. (2019) compared the fluency change between massed task repetition and 1-week interval repetition on the same two fluency measures as the current study's measurements (i.e., repetition and mid-clause pause frequency). The mean effect size of difference was 0.50 ($g = 0.59$ and 0.43 for repetition and mid-clause pause frequency, respectively). Based on the treatment-specific effect sizes from Donovan and Radosevich's meta-analysis and Bui et al.'s study, the effect size above 0.50 was considered meaningful in the current study. This magnitude of effect size is considered small according to a L2 field-general benchmark (Plonsky & Oswald, 2014): small (0.40), medium (0.70), and large (1.00).

Pretest–Posttest Changes

To compare the pretest–posttest changes among the four groups, a series of two-way mixed ANCOVAs were conducted. Each of the seven fluency measures was used as a dependent variable. Time (immediate and delayed posttests) was the within-subject variable, and Condition (massed, short-spaced, long-spaced, and control) was the between-subject variable. The interaction between Condition and Time was also included in the model to identify any group differences at different time points. Fluency measure on the pretest was used as a covariate to control for potential differences among the groups.

Because the four dependent variables (mid-clause pause duration, clause-final pause duration, repetition, and repair) were not normally distributed, a log-transformation was performed to correct the distributions. Because the data for mid-clause pause duration were not adequately corrected after a log, a square root, or an inverse transformation, a rank ANCOVA was conducted using the ranks of the pretest and posttest scores. Data pertaining to one participant was identified as an outlier ($z > 3.29$; Tabachnick & Fidell, 2013) for mid-clause pause duration and clause-final duration, respectively, and these

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

records were thus excluded from the following analyses. When a main effect or interaction was significant in the two-way mixed ANCOVAs, follow-up univariate ANCOVAs were conducted for each of the two posttests with Condition (massed, short-space, long-spaced, and control) as the between-subject variable for those fluency measures that were significant. The pretest score served as the covariate for estimating the posttest scores while controlling for the potential differences in the pretest performance. Multiple pairwise comparisons were performed with Bonferroni correction. The effect sizes were interpreted against the same benchmarks established for the training data.

RESULTS

PERFORMANCE CHANGES AND 1-WEEK RETENTION

Figure 2 illustrates the mean scores for each fluency measure during the training sessions (Time 1−Time 6) and on the retention performance 1 week later. Overall, the participants in all three conditions exhibited gradual changes on all fluency measures.

When inspecting group differences, as expected, a noticeable difference emerged at Time 4. Among the three groups, the long-spaced practice led to the largest changes from Time 3 to Time 4. In particular, at Time 4, participants assigned to the long-spaced condition tended to pause longer both within and between clauses. At the same time, they tended to pause less frequently within the clause. At the end of the training phase (Time 6), however, the long-spaced group achieved the same level of fluency as the massed and short-spaced groups. In terms of 1-week retention performance, there were no notable differences among the three groups, except for the increase in repetition in the massed compared with the long-spaced group.

Table 1 summarizes the mixed ANCOVA results for the training data.[4] Significant main (Condition) and/or interaction (Time × Condition) effects were found for four fluency measures (articulation rate, mid-clause pause duration, clause-final pause duration, and clause-final pause frequency). A marginally significant main effect ($p = .05$) was found for repetition. The sizes of partial eta squared for the main or interaction effects are presented in Table 1, which shows medium to large effect sizes (.08−.14). The interaction effects suggest that the effects of condition varied depending on the timing of performance, which was further analyzed in follow-up univariate ANCOVAs.

The follow-up univariate ANCOVAs (see Appendix D in Online Supplementary File for full results) showed a significant main effect of Condition on three fluency measures (mid-clause duration, clause-final pause duration, and clause-final pause frequency) at Time 4. While no significant main effect of Condition was detected at Time 6, it exhibited significant main effect on repetition for the 1-week retention.

To locate the group differences on some of the fluency measurements at Time 4 and 1-week retention, post-hoc multiple comparisons were conducted with Bonferroni correction. Figure 3 illustrates the differences among the groups on fluency measures that showed significant differences (see Table 2 for the summary of all effect sizes). Significant group differences were noted for three breakdown fluency measures at Time 4. First, the massed group and short-spaced group decreased the mid-clause pause duration more than the long-spaced group did with meaningful effect sizes ($g = -0.73$ [−1.37, −0.10] and $g = -0.99$ [−1.37, −0.10], respectively). Second, the massed and short-spaced groups decreased the

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

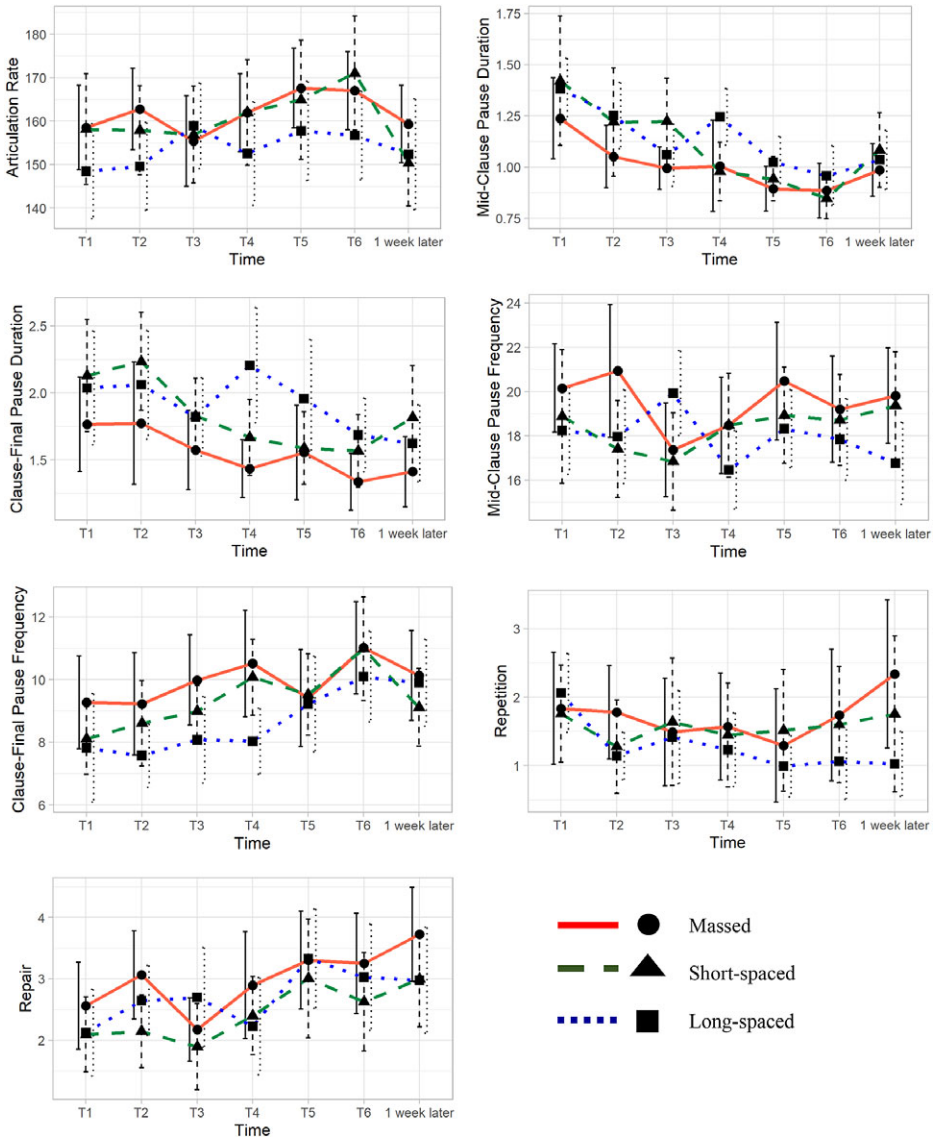

FIGURE 2. Performance change during training and at 1-week retention test.
Note: The numerical values for means, SDs, and 95% CIs are presented in Appendix C in the Online Supplementary File.

clause-final pause duration significantly more than the long-spaced group did with meaningful effect sizes ($g = -1.09$ [−1.75, −0.44] and $g = -1.04$ [−1.69, −0.38], respectively). Third, the long-spaced group paused significantly less at the clause boundary than the short-spaced group with a meaningful effect size ($g = 0.78$ [0.14, 1.41]). At the 1-week retention test, the massed group increased the number of repetitions significantly more than the long-spaced group with a meaningful effect size ($g = 0.94$ [0.29, 1.58]).

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

TABLE 1. Summary of mixed ANCOVAs on the training data

	Condition			Condition × time		
	F	*p*	ηp^2	*F*	*p*	ηp^2
Articulation rate	0.16	.85	.01	3.38	.01*	.10
Mid-clause pause dur.	1.98	.15	.06	2.64	.04*	.08
Clause-final pause dur.	2.97	.06$^+$	.09	4.66	.002*	.14
Mid-clause pause freq.	1.23	.30	.04	0.64	.63	.02
Clause-final pause freq.	0.31	.73	.01	4.08	.004*	.12
Repetitions	3.06	.05$^+$	.09	1.22	.31	.04
Repairs	0.68	.51	.02	0.31	.87	.01

$^+p < .10$; $^*p < .05$.

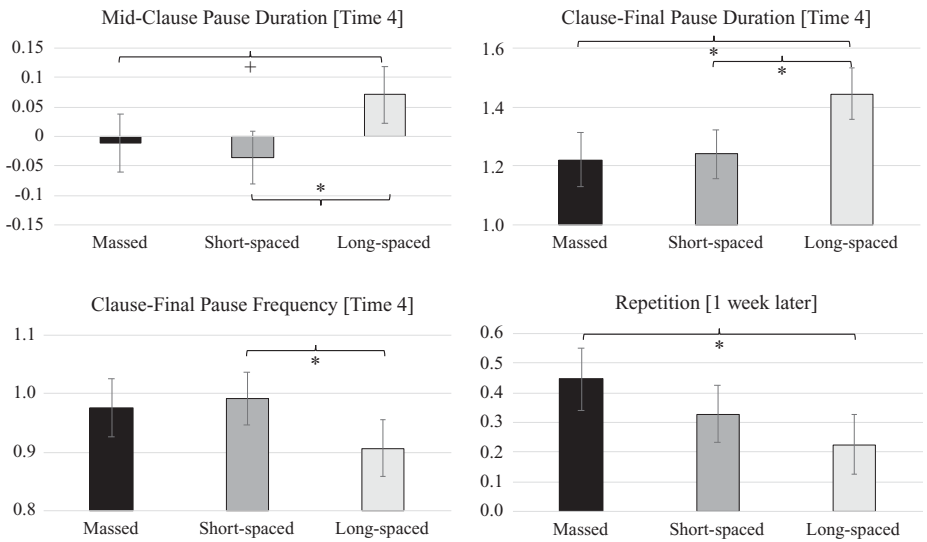

FIGURE 3. Mean scores (adjusted for the first practice performance) for fluency measures with significant effects at Time 4 and 1-week retention. The error bars represent 95% confidence intervals. Note that data transformation resulted in negative values in mid-clause pause duration.
Note: $^+p < .10$; $^*p < .05$.

PRETEST–POSTTEST CHANGES

Descriptive statistics for pretest, immediate, and delayed posttests are presented in Appendix F in Online Supplementary File. A series of two-way mixed ANCOVAs were conducted on all seven fluency measures.[5] As shown in Table 3, significant main and/or interaction effects with large effect sizes (.11 to .22) were found for three fluency measures (articulation rate, mid-clause pause frequency, and repetition). Additionally, a marginally significant interaction ($p = .05$) was noted for clause-final pause frequency with a medium−large effect size (.10). These four fluency measures were thus subjected to follow-up univariate ANCOVAs.

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

TABLE 2. Effect sizes for the multiple comparisons for the fluency measures at Time 4, Time 6, and 1-week retention

		Time 4		Time 6		1-week retention	
		Short-spaced	Long-spaced	Short-spaced	Long-spaced	Short-spaced	Long-spaced
Articulation rate	Massed	−0.02 [−0.62, 0.58]	0.12 [−0.49, 0.73]	−0.24 [−0.84, 0.36]	0.18 [−0.43, 0.80]	0.44 [−0.16, 1.05]	−0.02 [−0.63, 0.59]
	Short-spaced		0.14 [−0.47, 0.73]		0.42 [−0.18, 1.02]		−0.47 [−1.07, 0.13]
Mid-clause pause duration	Massed	0.22 [−0.38, 0.83]	−0.73^{+} [−1.37, −0.10]	0.27 [−0.33, 0.88]	−0.11 [−0.72, 0.50]	0.03 [−0.57, 0.63]	0.18 [−0.43, 0.80]
	Short-spaced		−0.99* [−1.37, −0.10]		−0.39 [−0.98, 0.21]		0.16 [−0.43, 0.75]
Clause-final pause duration	Massed	−0.09 [−0.69, 0.51]	−1.09* [−1.75, −0.44]	−0.17 [−0.77, 0.43]	−0.52 [−1.14, 0.11]	−0.41 [−1.01, 0.20]	−0.18 [−0.79, 0.43]
	Short-spaced		−1.04* [−1.69, −0.38]		−0.35 [−0.94, 0.25]		0.23 [−0.36, 0.83]
Clause-final pause frequency	Massed	−0.13 [−0.73, 0.47]	0.61 [−0.01, 1.24]	−0.12 [−0.72, 0.48]	0.04 [−0.57, 0.66]	0.22 [−0.38, 0.82]	−0.2 [−0.82, 0.41]
	Short-spaced		0.78* [0.14, 1.41]		0.17 [−0.42, 0.76]		−0.43 [−1.03, 0.17]
Repetitions	Massed	0.15 [−0.45, 0.75]	0.4 [−0.22, 1.02]	0.05 [−0.55, 0.65]	0.54 [−0.09, 1.16]	0.05 [−0.11, 1.11]	0.94* [0.29, 1.58]
	Short-spaced		0.25 [−0.36, 0.87]		0.48 [−0.12, 1.08]		0.44 [−0.16, 1.04]

Note: Negative effect sizes indicate that the condition (massed and short-spaced) on the left side showed smaller values than the conditions at the top (short-spaced and long-spaced). See Appendix E in Online Supplementary File for more detailed information.
*$p = .05$; $^{+}p < .10$.

TABLE 3. Results of mixed ANCOVAs on the posttests

	Condition			Condition × time		
	F	*p*	ηp^2	*F*	*p*	ηp^2
Articulation rate	3.33	.02*	0.12	7.11	<.001*	.22
Mid-clause pause dur.	2.13	.10	0.08	0.46	.71	.02
Clause-final pause dur.	0.16	.93	.01	1.42	.24	.06
Mid-clause pause freq.	2.50	.07^{+}	.09	2.91	.04*	.11
Clause-final pause freq.	0.79	.50	.03	2.76	.05^{+}	.10
Repetitions	3.00	.04*	.11	1.80	.15	.07
Repairs	0.20	.90	.01	0.96	.42	.04

Note: For the effect of time, see Appendix G in Online Supplementary File.
$^{+}p < .10$; $^{*}p < .05$.

The follow-up univariate ANCOVAs (see Appendix H in Online Supplementary File for full results) showed that Condition exerted significant main effects on all four fluency measures on the immediate posttest. Because no meaningful difference was found on the delayed posttest, only the immediate posttest results will be reported (see Table 4 for all effect sizes). Figure 4 illustrates the differences among the four groups (massed, short-spaced, long-spaced, and control) on the four fluency measures that showed significant differences, which will be discussed in detail in the text that follows.

Articulation Rate

Multiple pairwise comparisons showed that the articulation rate was the slowest for the massed group, followed by the short-spaced, long-spaced, and control groups. Significant differences were observed between the massed group and the other three groups all with meaningful effect sizes ($-1.73 < g < -0.88$), and between the short-spaced group and the control group with a meaningful effect size ($g = -0.94$ [-1.63, -0.26]).

Mid-Clause Pause Frequency

Participants assigned to the massed group decreased their mid-clause pause frequency significantly more than did those in the control group with a meaningful effect size ($g = -0.98$ [-1.69, -0.27]). None of the other comparisons was significant. Nevertheless, massed practice resulted in fewer mid-clause pauses than short-spaced and long-spaced practice with meaningful effect sizes, albeit with the 95% confidence intervals overlapping zero ($g = -0.62$ [-1.23, 0.00] and -0.63 [-1.25, 0.00], respectively).

Clause–Final Pause Frequency

Similarly, the massed group paused significantly less at the clause boundary than the control group with a meaningful effect size ($g = -1.02$ [-1.73, -0.31]). Although none of the other comparisons was significant, massed practice resulted in fewer clause-final

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

TABLE 4. Effect sizes for the multiple comparisons for the fluency measures on the immediate and delayed posttests

		Immediate posttest			Delayed posttest		
		Short-spaced	Long-spaced	Control	Short-spaced	Long-spaced	Control
Articulation rate	Massed	−0.88* [−1.51, −0.25]	−1.38* [−2.06, −0.69]	−1.73* [−2.06, −0.95]	0.06 [−0.54, 0.66]	−0.23 [−0.84, 0.39]	0.21 [−0.46, 0.88]
	Short-spaced		−0.49 [0.11, −0.5]	−0.94* [−1.63, −0.26]		−0.29 [−0.88, 0.31]	0.16 [−0.49. 0.81]
	Long-spaced			−0.43 [−1.10, −0.24]			0.42 [−0.25, 1.09]
Mid-clause pause frequency	Massed	−0.62 [−1.23, 0.00]	−0.63 [−1.25, 0.00]	−0.98* [−1.69, −0.27]	0.07 [−0.53, 0.67]	−0.46 [−1.07, 0.17]	−0.4 [−1.08, 0.27]
	Short-spaced		−0.01 [−0.06, 0.58]	−0.42 [−1.08, 0.24]		−0.54 [−1.13, 0.07]	−0.5 [−1.16, 0.16]
	Long-spaced			−0.39 [−1.06, 0.28]			0.03 [−0.63, 0.69]
Clause-final pause frequency	Massed	−0.64 [−1.25, −0.02]	−0.36 [−0.98, 0.26]	−1.02* [−1.73, −0.31]	0.04 [−0.56, 0.64]	0.21 [−0.41, 0.82]	0.19 [−0.48, 0.86]
	Short-spaced		0.28 [−0.32, 0.87]	−0.45 [−1.11, 0.21]		0.17 [−0.42, 0.76]	0.17 [−0.48, 0.82]
	Long-spaced			−0.68 [−1.37, 0.00]			0 [−0.67, 0.66]
Repetitions	Massed	1.13* [0.48, 1.77]	0.61 [−0.02, 1.05]	0.38 [−0.30, 1.50]	0.27 [−0.33, 0.87]	−0.08 [−0.70, 0.53]	−0.07 [−0.74, 0.60]
	Short-spaced		−0.52 [−1.12, 0.08]	−0.73 [−1.40, −0.06]		−0.35 [−0.94, 0.25]	−0.34 [−0.99, 0.32]
	Long-spaced			−0.2 [−0.87, 0.46]			0.01 [−0.65, 0.67]

Note: Negative effect sizes indicate that the condition (massed and short-spaced) on the left side showed smaller values than the conditions at the top (short-spaced and long-spaced). See Appendix I in Online Supplementary File for more detailed information.
*$p = .05$.

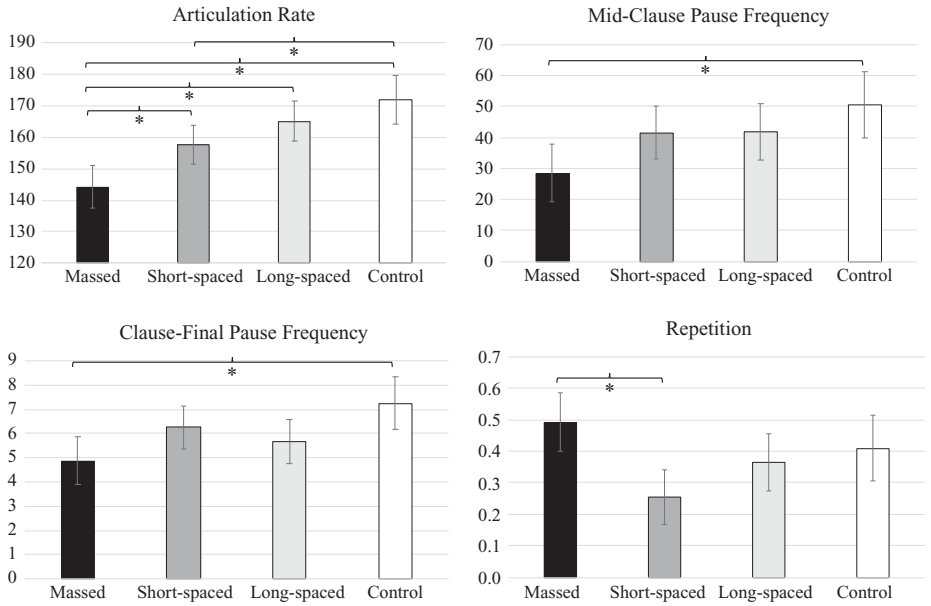

FIGURE 4. Mean immediate posttest scores (adjusted for the pretest scores) with significant effects for massed, short-spaced, long-spaced, and control conditions. The error bars represent 95% confidence intervals.
*p <.05.

pauses than short-spaced practice with meaningful effect size, albeit with the 95% confidence interval overlapping zero ($g = -0.64$ [−1.25, −0.02]).

Repetition

The massed group exhibited significantly more verbatim repetitions than the short-spaced group with a meaningful effect size ($g = 1.13$ [0.48, 1.77]). None of the other comparisons was significant.

DISCUSSION

EFFECTS OF TASK REPETITION SCHEDULES DURING FLUENCY TRAINING AND ON RETENTION

The current findings indicate that different task repetition schedules influence fluency changes during the training phase (Research Question 1) as well as 1-week retention (Research Question 2) with the effect sizes well above the treatment-specific benchmark ($g > .50$). The analyses further demonstrated that, at Time 4, fluency performance of the long-spaced group was most affected by repetition schedule. While participants assigned to the long-spaced group made fewer clause-final pauses than did those in the short-spaced group, they paused for longer durations at both within and between clause boundaries than the other two groups. This suggests that the longer mid-clause and

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

clause-final pauses in the long-spaced condition at Time 4 may indicate the learners' more effortful retrieval of information. However, the short-spaced practice led to comparable performance with massed practice during the training phase. This may imply that the brief interval (about 45 minutes) between the two blocks of three repetitions in a 90-minute class was short enough to maintain or improve their fluent performance at Time 4.

However, at the final (sixth) task performance, there was no significant main effect of practice schedule across any of the seven fluency measures. This suggests that, when the same task is repeated six times, the differences in task distribution (i.e., six times consecutively, three times at the beginning and at the end of 90-minute class, and three times per class with a 1-week gap) exerted little influence on fluency performance at the end of the training phase.

Although the three groups' performances were comparable at the end of the training phase (Time 6), there was one significant main effect of practice schedule on the number of self-repetitions. Specifically, the long-spaced group repeated the same words less frequently than the massed group. The effect size was 0.94, which is almost twice as large as the treatment-specific benchmark and corresponds to large size in the field-general benchmark. This pattern may be consistent with some of the findings reported by Bui et al. (2019), which revealed that 1-week interval repetition reduced the number of verbatim repetitions on the second performance. A novel aspect of the current findings is that long-spaced practice may be beneficial in reducing the verbatim repetition 1 week after the treatment. The 1-week interval might have deactivated the linguistic expressions on the repeated task performance at Time 4. After such a long interval, the participants had to deliberately encode linguistic information that had been deactivated 1 week ago. Possibly, effortful encoding positively contributed to the reduction of verbatim repetition (cf., desirable difficulty; Bjork, 1994; Suzuki et al., 2019a, 2020). In contrast, massed practice may exert adverse effects on repetition because learners who repeated the same task six times without spacing might be more likely to reuse a certain set of linguistic items that were activated and reinforced during each repetition through priming mechanisms (Bock & Griffin, 2000). These highly activated processes in the massed practice group might have adversely affected the verbatim repetition behaviors.

EFFECTS OF TASK REPETITION SCHEDULES FOR SHORT-TERM TRANSFER: DOUBLE-EDGED SWORD OF MASSED TASK REPETITION

The third research question of the current study concerned the extent to which the fluency gains through task repetition transfer to fluency changes in narrative performance based on new prompts. The results showed that the repetition schedules exerted large influence on the fluency performance on the immediate posttest requiring an unfamiliar narration with a new cartoon. Post-hoc multiple comparison highlighted a significant difference between massed practice and the other groups ($0.88 < |g| < 1.73$) on four fluency indices. Intriguingly, massed practice led to a trade-off effect between breakdown fluency and speed-repair fluency on the immediate posttest. In what follows, both advantages and disadvantages of massed practice are discussed.

Although there were no significant differences among the three experimental groups, massed practice enhanced short-term breakdown fluency development (i.e., fewer mid-clause pauses and clause-final pauses) with meaningful effect sizes. In particular, because

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

mid-clause pause is arguably related to proceduralization of L2 knowledge (Kahng, 2014; Kormos, 2006; Suzuki, 2021b), the current findings suggest that massed practice plays a potential facilitative role for certain aspects of L2 speaking skill development. Specifically, repeating the same task consecutively *en masse* might have allowed primed or preactivated linguistic items (e.g., words, chunks) to be retrieved more efficiently with fewer pauses, resulting in the superiority of massed practice.

However, the drawbacks of massed practice were also found for speed fluency (slower articulation rate) and repair fluency (more repetition). Inspection of Figure 2 suggests that the massed practice group reached a plateau in their articulation rate improvement at Time 5. This suggests that repeating the same task six times consecutively was likely to be too repetitive and exerted some adverse effects on learners, although massed repetition was effective in reducing breakdown fluency. This observation may be corroborated by a brief questionnaire administered to all three experimental groups after the immediate posttest. This questionnaire was included purely for exploratory purposes, but the participants in the massed practice group felt more bored and less focused during the task performance than the other two groups (see Appendix J in Online Supplementary File for the results). In other words, spacing (of both 45-minute and 1-week duration) seems effective in mitigating boredom and fatigue.

Furthermore, massed practice increased verbatim repetition significantly more than did short-spaced practice, suggesting that inserting a 45-minute period dedicated to other nonspeaking activities between speaking task repetitions is effective in reducing verbatim repetition in a new task. Recall that on the practiced (familiar/old) task performance 1 week after the training, verbatim repetition increased in the massed practice condition more than in the long-spaced practice condition. Taken together, longer spacing such as 1 week might be necessary to significantly reduce verbatim repetition in the performance of the same task, while short-spaced practice might have been optimal to diminish the carryover of verbatim repetition "habits" to a new task. It is speculated that different cognitive mechanisms may underlie these two phenomena (increased verbatim repetition in a familiar task vs. a novel task) and could be worthy of further investigations. In sum, massed task repetition is a double-edged sword for fluency training, the effectiveness of which needs to be considered more carefully (see the "Pedagogical Implications" section).

DIVERGENT PATTERNS ON DELAYED POSTTEST PERFORMANCE COMPARED TO PREVIOUS LITERATURE ON DISTRIBUTED PRACTICE EFFECTS

While both benefits and drawbacks of massed task repetition were revealed on the immediate posttest, virtually no significant differences among task repetition schedules were found with respect to any fluency measures at the 1-week delayed posttest involving new prompt (all effect sizes were below the treatment-specific benchmark, $g < 0.50$), except for the mid-clause pause frequency between short-spaced and long-spaced groups. Unobservable spacing effects or lag effects may seem trivial, but are indeed an interesting phenomenon in light of the body of literature on distributed practice effects in general. Meta-analyses of studies based on verbal recall tasks clearly indicate the advantage of spaced practice over massed practice on delayed posttests (Cepeda et al., 2006). Furthermore, several L2 studies showed the distributed practice effects on delayed posttests that

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

are administered at least 1 week after the treatment (e.g., for pronunciation learning, Li & DeKeyser, 2019; for vocabulary learning, Nakata, 2015; Nakata & Suzuki, 2019; for grammar learning, Bird, 2010; Rogers, 2015; Suzuki, 2017).

To the best of our knowledge, consistent with the present study's findings, authors of two L2 studies have documented no significant differences between short- and long-spaced practice on L2 grammar acquisition (Kasprowicz et al., 2019; Suzuki & DeKeyser, 2017a). In both studies, L2 grammar was learned through auditory processing with the support of written input in various listening and/or speaking activities. Because spoken input is more difficult to monitor than written input, it may increase the learning difficulty (Suzuki et al., 2019a). Oral narration task in the current study was presumably also complex and difficult for the participants, which might have reduced the benefit of spaced practice (Donovan & Radosevich, 1999). Another interpretation may pertain to the procedural nature of target skills (i.e., achieving fluent speech). Because the acquisition of procedural knowledge takes more practice opportunities than declarative knowledge (DeKeyser, 2020; Kim et al., 2013), only one block of six repetitions may not be sufficient to promote procedural knowledge, which makes it difficult to gain distributed practice effects. Yet, as the distributed practice effects were evident even with six task repetitions on the immediate posttest in the current study, a longer intervention study (e.g., over one semester) may be needed to demonstrate durability of its effects.

The diminished distributed practice effects may also be due to the timing of the delayed posttest (i.e., retention interval). According to cognitive psychologists (Cepeda et al., 2008; Rohrer & Pashler, 2007), the optimal spacing depends on the retention interval (i.e., the timing of delayed posttest). The optimal ratio of spacing and retention interval should be 10–30%. In the present study, however, the ratio of spacing and retention for the long-spaced group in the current study was 100% (7-day interval was adopted for learning in the long-spaced group as well as for the posttest). Consequently, it is likely that the delayed posttest was conducted too early to reveal any benefits of longer spacing. In other words, if the training effect is durable, the long-spaced group could have performed better on a delayed posttest administered 23–70 days after the intervention (corresponding to the 10–30% ratio for the 7-day learning interval).

Lastly, the lack of significant effect of spacing on the delayed posttest does not necessarily mean that distributed practice does not promote fluency. In addition to the aforementioned potential factors (e.g., complexity and types of knowledge), other moderating factors also need to be considered, such as experimental context, like laboratory versus classroom (Rogers & Cheung, 2020a), as well as frequency of study sessions (Suzuki, 2017). Instead of drawing any conclusions from the failure to observe long-term effects of distributed practice in speaking task repetition, further investigations are clearly warranted.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

Several limitations of the current study are addressed in the text that follows to provide directions for future research. First, the number of participants was small, particularly in the control group ($n = 15$). Although the coding of detailed fluency measures is laborious, the current findings need to be attested with a larger sample size for further generalizations that would facilitate a more precise estimation of distributed practice effects for fluency development.

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

Second, because random assignment of participants to each condition was not feasible in this classroom research, individual differences such as cognitive aptitudes should have been controlled at least statistically. As accumulating evidence suggests that distributed practice effects in L2 grammar learning are moderated by individual characteristics such as working memory and language analytic ability (Kasprowicz et al., 2019; Suzuki, 2019; Suzuki & DeKeyser, 2017b), it is worth exploring possible aptitude−treatment interaction between practice distribution and learners' cognitive aptitude profiles in future research (see Suzuki, 2021a, who demonstrated that some aspects of memory predict the effects of L2 fluency training).

Third, as the current study focused on fluency changes, other speech aspects such as complexity and accuracy (appropriateness) were not analyzed. Because task repetition schedules also seem to moderate complexity and accuracy (Bui et al., 2019), authors of future research in this domain need to expand the scope of outcome measures. When examining the effects of task repetition on accuracy and complexity, simply repeating the task as in the current study may not be sufficient to induce substantial performance changes (Ellis, 2009). Combining task repetition with some linguistic support (e.g., presenting models) may thus be useful.

Last but not the least, the current operationalization of task repetition schedules could be further extended to examine a greater variety of options. In the current study, one task was repeated at least three times immediately in all experimental groups as follows:

Massed: XXXXXX
Short-spaced: XXX---45 min---XXX
Long-spaced: XXX----------7 days----------XXX

This operationalization of massed, short-spaced, and long-spaced task repetition could have attenuated the differences among the three interval conditions because repeating the same task three times consecutively might be already effective regardless of when the second block of task repetition practice takes place. To scrutinize distributed practice effects in fluency training, it may be worth adopting a simpler operationalization, such as:

Massed: XXX
Short-spaced: X---30 min---X---30 min---X
Long-spaced: X----------7 days----------X----------7 days----------X

In future research, the effects of unit size of blocked repetition (e.g., three times in the current study) as well as intervals between repetitions should also be explored.

PEDAGOGICAL IMPLICATIONS

Task repetition is an effective teaching technique for fostering L2 learners' fluency (Bygate, 2018; Lambert et al., 2017; Tavakoli & Hunter, 2018). Novel contribution from the current short-term classroom intervention is that massed task repetition is a double-edged sword. Massed practice reduced breakdown fluency the most but led to slower articulation rate and greater repetition on the immediate posttest, while potentially reducing motivation. Despite some potential values of massed practice, leaners may not be motivated to engage in massed practice in the current form. Therefore, it is

https://doi.org/10.1017/S0272263121000358 Published online by Cambridge University Press

advisable to avoid simply repeating the same monologue tasks six times. As shown in the current study, inserting other activities for periods as short as 45 minutes can reduce the drawbacks of massed repetition. This is an easy option without requiring any additional resources. However, changing the task format might render massed practice more effective. If speaking practice is performed in pairs, for instance, massed task repetition (even six times) could still be effective (cf., Lambert et al., 2017). As opposed to the current monologue practice, learners would be more engaged in repeated practice as they get to learn from and interact with their peers. Furthermore, providing feedback and models of narration adjusted to L2 learners' interlanguage level is also important for engaging them in repetition activities and potentially facilitates more accurate and appropriate language use (Lynch, 2018). As recommended by Johnson (1996), provision of corrective feedback immediately after task performance (and just before repeated performance) helps learners notice and correct their linguistic errors. Mindful engagement in repeated narration—conscious effort to use more correct linguistic forms—will likely impose "desirable" challenges on learners, motivating them to become better L2 users (Bjork, 1994; Suzuki et al., 2019a, 2020).

CONCLUSIONS

The objective of this short-term classroom intervention study was to further our understanding of distributed practice effects in the context of L2 speaking task repetition. By examining the differential effects of task repetition schedules on L2 fluency development, massed task repetition was found effective in reducing breakdown fluency (mid-clause and clause-final pauses) but led to lower speed fluency (articulation rate) and repair fluency (verbatim repetition) on the immediate posttest. No significant effect of repetition schedule was found on 1-week delayed posttest. These findings demonstrate the importance of studying distributed practice in the context of L2 learning, which can reveal new insights and contributes to a very large body of literature of cognitive psychology on distributed practice effects. Clearly, the question on distributed practice that needs to be addressed has changed from "whether or not" to "under what conditions" spacing creates the optimal learning conditions for different aspects of L2 acquisition. The time is ripe for SLA researchers to harness an interdisciplinary perspective and apply cognitive psychology findings, with the aim of *re*informing cognitive psychology by investigating L2 learning, which entails acquisition of one of the most complex cognitive skills.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S0272263121000358.

NOTES

[1]Two different prompts were employed for the training session for a different research purpose. A computerized psycholinguistic task was adopted to assess the processing of collocation; however, due to technical issues, the task was not administered and the use of different prompts lost their purpose. Ad-hoc comparisons using independent-samples *t*-tests showed that there were no significant differences in the seven

fluency measures (see the "Data Coding" section) between the two prompts in the first and the final (sixth) performance ($p > .05$).

[2]The posttest was conducted after one week for two reasons. First, as this was a weekly course, this was a minimal gap between tests. Second, if it was postponed further, participants could have improved their speaking ability outside of this intervention.

[3]ANCOVA was chosen rather than repeated measures ANOVA because ANCOVA is more appropriate for estimating the posttraining scores in each group after controlling for the pretest score (Dimitrov & Rumrill, 2003).

[4]According to the one-way ANOVAs, there was no significant main effect of condition for any fluency measures at Time 1 ($p > .10$).

[5]According to the one-way ANOVAs, there was no significant main effect of condition for any fluency pretest measures ($p > .10$), with the exception of clause-final pause duration ($p = .01$).

REFERENCES

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

Arevart, S., & Nation, P. (1991). Fluency improvement in a second language. *RELC Journal*, *22*, 84–94. https://doi.org/10.1177%2F003368829102200106

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

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, *129*, 177–192. https://doi.org/10.1037/0096-3445.129.2.177

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

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). Heinemann.

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). Pearson Longman.

Bygate, M. (2018). *Learning language through task repetition*. John Benjamins Publishing Company.

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning a temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

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.), *Language learning through task repetition* (pp. 43–73). 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., & 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. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press.

DeKeyser, R. M. (2020). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 83–104). Routledge.

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest–posttest designs and measurement of change. *Work: Journal of Prevention, Assessment & Rehabilitation*, *20*, 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, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*, 474–509. https://doi.org/10.1093/applin/amp042

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

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

Gass, S., Mackey, A., Alvarez Torres, M. J., & Fernández García, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, *49*, 549–581. https://doi.org/10.1111/0023-8333.00102

Hanzawa, K. (2021). Development of second language speech fluency in foreign language classrooms: A longitudinal study. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688211008693

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

Johnson, K. (1996). *Language teaching and skill learning*. Blackwell Publishers.

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. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*, 569–591. https://doi.org/10.1017/S0142716417000534

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

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

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*. Routledge.

Lambert, C., Aubrey, S., & Leeming, P. (2021). Task preparation and second language speech production. *TESOL Quarterly 55*, 331–365. https://doi.org/10.1002/tesq.598

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*. 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

Lynch, T. (2018). Perform, reflect, recycle: Enhancing task repetition in second language speaking classes. In M. Bygate (Ed.), *Learning language through task repetition* (pp. 193–222). John Benjamins Publishing Company.

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%2F136216880000400303

Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. García Mayo, J. Gutierrez-Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71–92). John Benjamins Publishing Company.

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). Routledge.

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

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., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, *37*, 233–260. https://doi.org/10.1177/0267658320927764

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

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

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, *38*, 906–911. https://doi.org/10.1093/applin/amw052

Rogers, J., & Cheung, A. (2020a). Does it matter when you review? Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition*, 1–19. https://doi.org/10.1017/S0272263120000236

Rogers, J., & Cheung, A. (2020b). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*, *24*, 616–641. https://doi.org/10.1177/1362168818805251

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review*, *27*, 635–643. https://doi.org/10.1007/s10648-015-9332-4

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

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

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*, 1–14. https://doi.org/10.1017/s026144480200188x

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. (2019). Individualization of practice distribution in second language grammar learning: A role of metalinguistic rule rehearsal ability and working memory capacity. *Journal of Second Language Studies*, *2*, 170–197. https://doi.org/10.1075/jsls.18023.suz

Suzuki, Y. (2021a). Individual differences in memory predict changes in breakdown and repair fluency but not speed fluency: A short-term fluency training intervention study. *Applied Psycholinguistics*. Advance online publication. https://doi.org/10.1017/S0142716421000187

Suzuki, Y. (2021b). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*, *71*, 285–325. https://doi.org/10.1111/lang.12433

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

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

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., Nakata, T., & DeKeyser, R. M. (2020). Empirical feasibility of the desirable difficulty framework: Toward more systematic research on L2 practice for broader pedagogical implications. *The Modern Language Journal*, *104*, 313–319. https://doi.org/10.1111/modl.12625

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

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto the common European framework of reference: An application of standard-setting methodology. *ETS Research Report Series*, *2008*, i–75. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *58*, 439–473. https://doi.org/10.1111/j.1467-9922.2008.00446.x

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

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

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

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

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