Studies in Second Language Acquisition, **43** (2021), 1138–1156
doi:10.1017/S0272263120000236

Replication Study

DOES IT MATTER WHEN YOU REVIEW?

INPUT SPACING, ECOLOGICAL VALIDITY, AND THE LEARNING OF L2 VOCABULARY

John Rogers *

The Education University of Hong Kong

Anisa Cheung

The Education University of Hong Kong

Abstract

This study is a conceptual replication of Rogers and Cheung's (2018) investigation into distribution of practice effects on the learning of L2 vocabulary in child EFL classrooms in Hong Kong. Following a pretest, treatment, delayed posttest design, 66 primary school students (Cantonese L1) studied 20 vocabulary items over three training episodes under spaced-short (1-day interval) or spaced-long (8-day interval) learning conditions. The spacing of the vocabulary items was manipulated within-participants, and learning was assessed using crossword puzzles following a 4-week delay. While Rogers and Cheung (2018) resulted in minimal overall learning with a slight advantage for the spaced-short group, this study found large learning gains across the experimental conditions with no significant differences between the two learning schedules. Taken together, these results provide evidence that the results from previous research examining input spacing with adult populations in laboratory contexts might not generalize to authentic child learning contexts.

INTRODUCTION

How learning and instruction might be optimized remains an important goal for second language acquisition (SLA) research, with clear implications for classroom practice (Rogers & Leow, 2020; Suzuki et al., 2019, 2020). One area that has received considerable attention within the field of cognitive psychology is how the distribution of practice might influence the quantity and quality of learning that takes place. *Distribution of practice*, also referred to as *input spacing*, refers to whether and how learning is spaced

* Correspondence concerning this article should be addressed to John Rogers, Department of English Language Education, The Education University of Hong Kong, Room 22B, 1/F, Block B4, 10 Lo Ping Road, Tai Po, N.T., Hong Kong. E-mail: rjrogers@eduhk.hk.

over multiple learning episodes. *Massed practice* refers to experimental conditions in which learning is concentrated into a single, uninterrupted training session, whereas *distributed* or *spaced practice* refers to learning that is spread over two or more training episodes. The term *the spacing effect* refers to the phenomenon of distributed practice being superior to massed practice for learning, a finding that is particularly evident in delayed testing (Rogers, 2017a; Rohrer, 2015). The term *lag effect* refers to the impact of spacing gaps of varying lengths, such as a one-day gap between training sessions versus a one-week gap. The term *distributed practice effect* is used as a blanket term for both spacing and lag effects (Cepeda et al., 2006; Rogers, 2017a).

A growing amount of SLA research has examined distribution of practice effects on the learning of second language (L2) grammar (e.g., Bird, 2010; Kasprowicz et al., 2019; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017). These studies have been motivated in part to examine the degree to which empirical research in cognitive psychology generalizes to SLA, in particular research that has examined the relationship between the timing between practice sessions (intersession interval, or ISI) and the amount of time between the final practice session and testing (retention interval, or RI). For example, an influential study in the field of cognitive psychology by Cepeda et al. (2008) comprehensively mapped the optimum ISI/RI ratios with regard to the learning of trivia facts. The results of this study indicated that the optimum spacing between training sessions is dependent on when the knowledge will later be used (i.e., when it will be tested), and that the optimal ISI is approximately 10 to 30% of RI (Rohrer & Pashler, 2007).

With regard to the learning of L2 vocabulary, the focus of the present investigation, a number of studies have examined the effects of input spacing with adult populations (e.g., Cepeda et al., 2009; Küpper-Tetzel & Erdfelder, 2012; Nakata, 2015; Nakata & Suzuki, 2019; Pavlik & Anderson, 2005). While these studies have generally found advantages for more distributed conditions, studies examining distributed practice effects with nonadult populations, that is, young learners in authentic contexts, have returned conflicting results: either no difference between the spacing conditions (e.g., Küpper-Tetzel et al., 2014) or greater learning effects for more intensive conditions (e.g., Rogers & Cheung, 2018).

Given that these results contradict the findings of a wealth of laboratory-based studies from the field of cognitive psychology, it is prudent to replicate these studies to establish the external validity of these findings (Porte & McManus, 2018; Rogers & Révész, 2020). This study represents a conceptual replication of a recent study (Rogers & Cheung, 2018) that set out to examine distributed practice effects with young learners under ecologically valid learning conditions. The aim of the present replication is to examine the degree that the results of Rogers and Cheung (2018) generalize to a different teaching and learning context, with the broader goal of establishing whether input spacing is a viable method within an authentic teaching and learning environment. In the following sections, we first review the motivation of Rogers and Cheung's (2018) original study. We then provide a rationale for the present conceptual replication, before describing the present study.

MOTIVATION FOR THE ORIGINAL STUDY

The motivation for Rogers and Cheung's (2018) original study was rooted in questions of the generalizability of previous input spacing research in authentic SLA teaching and

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

learning contexts. Although distribution of practice effects are frequently cited as one of the most widely researched and robust findings in all the educational sciences, it has been argued that only a handful of these studies are "educationally relevant" (Rohrer, 2015). As such, Rogers and Cheung (2018) contend that there is a far less stable foundation to base any claims of the benefits of distributing practice over longer periods than is typically claimed.

A second criticism levied by Rogers and Cheung (2018) concerns the ecological validity of previous, "educationally relevant" research. Specifically, Rogers and Cheung argue that previous research, including classroom-based studies, has strived for high degrees of internal validity in their experimental designs, at the cost of external and ecological validity. To elaborate, *internal validity* relates to experimental control and refers to the level of certainty that the results of the experiment can be attributed to the experimental treatment, that is, that the observed changes in the dependent variable are a result of the independent variable. Any factor that allows for an alternative interpretation of the findings represents a threat to internal validity (Shadish et al., 2002). *External validity* refers to the degree that results hold true outside of the particular study, that is, the generalizability of the findings. External validity is best established through replication (Porte & McManus, 2018; Shadish et al., 2002). It is widely acknowledged that there is a constant tension between internal and external validity in experimental research (e.g., Hulstijn, 1997; Rogers & Révész, 2020). Related to external validity is the construct of *ecological validity*, which is related to the "ecology" of a particular context. A study can claim to have ecological validity if the experiment is similar to the context to which it aims to generalize. Ecological validity in this sense refers not only to the location in which the experiment takes place but also to the interrelations of all aspects of the setting (e.g., Van Lier, 2010).

Rogers and Cheung argue that many studies that meet Rohrer's criteria of educational relevance (e.g., Küpper-Tetzel et al., 2014) have overemphasized internal validity in their experimental designs at the cost of decreased external and ecological validity. It goes without saying that a high level of experimental control is not an issue in itself and that tightly controlled laboratory studies have greatly contributed to our understanding of SLA (see, e.g., Hulstijn, 1997). The issue is that much previous research that has claimed ecological validity has adopted a narrow interpretation in that they operationalize this construct to refer to only the location in which the experimental study takes place. Although these studies have taken place in a classroom, they have imposed artificial experimental conditions that would not be present normally in a classroom environment. At best, this has meant that the experimental conditions have not been validated with regard to the learning environment in which the study takes place (Rogers & Révész, 2020). At worst, the studies have imposed artificial experimental conditions that do not reflect an authentic learning environment, such as not allowing participants to take notes (e.g., Küpper-Tetzel et al., 2014) or forbidding participants from engaging in specific cognitive strategies during instruction (e.g., Pavlik & Anderson, 2005).

To address these issues, Rogers and Cheung set out to examine whether the effects of input spacing extend to an authentic teaching and learning environment. Using a within-participants experimental design, Rogers and Cheung examined the learning of L2 vocabulary, specifically 20 English adjectives related to describing people, across four classrooms in a primary school in Hong Kong. What was innovative about Rogers and

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

Cheung's (2018) study was that they asked the teachers to teach the target L2 lexical items as they normally would, with the condition that each individual teacher was consistent in their approach across the training episodes. Vocabulary was learned under two different spacing conditions: a spaced-short condition with a 1-day ISI (i.e., gap between training sessions) and a spaced-long condition with an 8-day ISI. Learning was assessed after a 28-day delay (RI) using a multiple-choice test, which asked the learners to circle the picture that most closely matched the meaning of the target item. Classroom observations were carried out over the course of the experiment by the research team to examine how the teachers taught the material over the training conditions.

The results of the lesson observations indicated that the four teachers largely adopted similar approaches in first presenting the material to the students, then emphasizing the pronunciation/spelling of the target items, followed by choral drilling and another form of form-focused practice, for example, completing crossword puzzles. The results of the delayed posttest showed minimal learning gains across the four classrooms: roughly 10% improvement across all target items. This minimal amount of learning might be explained in terms of the lack of transfer appropriateness across the training conditions, which emphasized the pronunciation of items, and the testing condition, which reflected the degree to which participants could recognize the orthography of target items. Despite the minimal amount of learning, participants learned the spaced-short items at significantly higher rates than the spaced-long items, going against predictions of theoretical models of lag effects.

JUSTIFICATION FOR REPLICATION

There are a number of theoretical, pedagogic, and methodological reasons that justify a replication of Rogers and Cheung's (2018) study. On a theoretical level, it is important to investigate lag effects in L2 learning because lag effects are closely linked with the benefits of repetition, review, and practice. Such benefits find support in skill-based theories of SLA (e.g., DeKeyser, 2015) where deliberate practice can aid in the transition from declarative/explicit knowledge to procedural knowledge. Lag effects have also been identified as an area of research with useful pedagogical implications because they may help bring about "desirable difficulties" in learning. Desirable difficulties occur as a result of conditions that "trigger the encoding and retrieval processes that support learning, comprehension, and remembering" (Bjork & Bjork, 2011, p. 58). In other words, long-term memory is strengthened as a result of making retrieval effortful, for example through spacing practice over longer periods (see Lightbown, 2008; Rogers & Leow, 2020; Suzuki et al., 2019, 2020 for discussions of desirable difficulties in L2 learning) and so may lead to more efficient and effective L2 practice (Suzuki et al., 2019).

Further, as noted, a limited number of ecologically valid studies have examined the effects of the distribution of practice across all domains of learning. Additional research is needed to justify any claims as to its pedagogical applications. In addition, there are even fewer studies examining the effects of input spacing on the learning of L2 vocabulary in authentic teaching and learning contexts with nonadult populations of learners. By carrying out such a replication, the present study would meet wider calls for more ecologically valid research within the field of SLA with nontraditional populations

(Kasprowicz & Marsden, 2018; Lightbown & Spada, 2019; Rogers & Révész, 2020; Spada, 2005, 2015).

This replication study also sets out to address some of the limitations of Rogers and Cheung's (2018) study. First, while one of the strengths of Rogers and Cheung's study is the use of an ecologically valid research design, the study lacked the experimental control that can be found in other laboratory and cognitively oriented classroom-based research. As such, a replication is needed to help avoid "leaps of logic" in extrapolating research findings from one context to another (Hatch, 1979; Spada, 2015) and to ensure that the results are real and do not reflect an artefact of the environment in which the study took place. A replication would also help in building a body of evidence toward the generalizability of distribution of practice effects in SLA in that "extrapolating from one 'controlled' experimental study … may not be as great as extrapolating from several 'less controlled' classroom studies (i.e., with intact classes) that report similar findings in distinctive settings" (Spada, 2005, p. 334). In other words, a replication would provide evidence as to the applicability of input spacing in authentic learning environments.

One limitation of Rogers and Cheung's (2018) study that the present study set out to address is the low levels of learning demonstrated by the learners. Theoretical accounts of the benefits of distributed practice often include some aspect of the benefits of retrieval (Toppino & Gerbier, 2014). In the case of Rogers and Cheung (2018), it could be argued that the low levels of learning mask any benefits of spaced practice. As such, we have set out to better link the materials used during training and testing, with a view to using materials that are both transfer-appropriate and valid with regard to the teaching and learning context in which this experimental study is set.

PRESENT STUDY

Like Rogers and Cheung (2018), the present study adopted a within-participants experimental design (Rogers & Révész, 2020) to examine the impact of different spacing schedules on the learning of L2 vocabulary in a classroom setting. A summary comparing the methodological features of Rogers and Cheung (2018) and the present study is presented in Table 1.

PARTICIPANTS

An *a priori* power analysis was carried out using G*power 3.1 (Faul et al., 2009). Anticipating a medium-sized effect ($f = .25$) within a mixed within-between 2×2 ANOVA experimental design, the analysis here revealed that a minimum of 54 participants were required to achieve a power of .95 with an alpha level of .05.

The participants who were recruited for this study were 87 children (L1 Cantonese, aged 8–9), studying in Primary Grade 4 in a Hong Kong Primary school in a low socioeconomic setting in Hong Kong. These participants were drawn from three intact classrooms. Each class was taught by a different instructor (three instructors in total), each with multiple years of experience teaching in the local context. The children who took part in this study had been studying English for nearly four years, having begun learning English in kindergarten as part of the Hong Kong Primary Curriculum. The children in this study were aware that they were taking part in an experimental study, and informed

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

TABLE 1. Comparison of methodological features of Rogers and Cheung (2018) and the present study

	Rogers and Cheung (2018)	Present study
Participants	52 primary students (Cantonese L1), 8–9 years old	66 primary students (Cantonese L1), 8–9 years old
Target items	20 English adjectives, not part of regular curriculum	20 English words, part of English curriculum
Materials	Word list and corresponding clip-art images	Word list and corresponding clip-art images; PowerPoint presentations; crossword puzzles
Training sessions	2 training conditions (within-group): 1-day ISI and 8-day ISI	2 training conditions (within-group): 1-day ISI and 8-day ISI
Training procedure	Teachers asked to teach the target items freely but remain consistent in their approach across training episodes	Teachers asked to use the provided materials in teaching the target items, and remain consistent in how they use these materials across the training episodes
Testing session	4-week delayed	4-week delayed
Testing measures	Multiple-choice recognition test (form recognition)	Crossword puzzle production test (form recall)
Qualitative measures	Classroom observations	Postexperimental interviews

consent was collected from their parents, teachers, and the school administration prior to the commencement of the study. Data were excluded from participants whose guardians did not complete and return the consent forms and from participants who missed one or more of the testing and/or training sessions, resulting in a final participant pool of 66 participants.

MATERIALS

All materials for this study were chosen and/or designed based on considerations of their appropriateness within the local teaching and learning context, with input from the stakeholders involved in the project, that is, teachers and other school representatives.

Twenty English words were selected for this project from the content vocabulary words in the students' course book. These 20 words comprised a mix of word classes: prepositions, nouns (e.g., body parts/food), and action verbs. To help control for item frequency effects, these 20 words were divided into two separate word lists of 10 words each. This division was carried out with the aim of maintaining a balance, as far as possible, with considerations of lexical rarity (Cobb, 2016), word class, and, in the case of verbs, regular and irregular forms (see Table 2).

TRAINING AND TESTING MATERIALS

Two PowerPoint presentations were created for the teachers to use when presenting the target items: one for List A and one for List B. Each slide was animated to present information incrementally. A completed slide can be seen in Figure 1.

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

TABLE 2. Lexical rarity and word groups of target items included in the study

	List A	List B
Prepositions	between	behind
Body parts	eyes, **paw**	**nose**, **feet**
Action verbs	walked, sat, **roared**	**poured**, looked, **stole**
Food	*chili*, **soup**	bread, *pies*
Other nouns	tears, **brush**	paint, **smoke**

Note: K1 words in normal font, K2 words in bold, off-list words in italics.

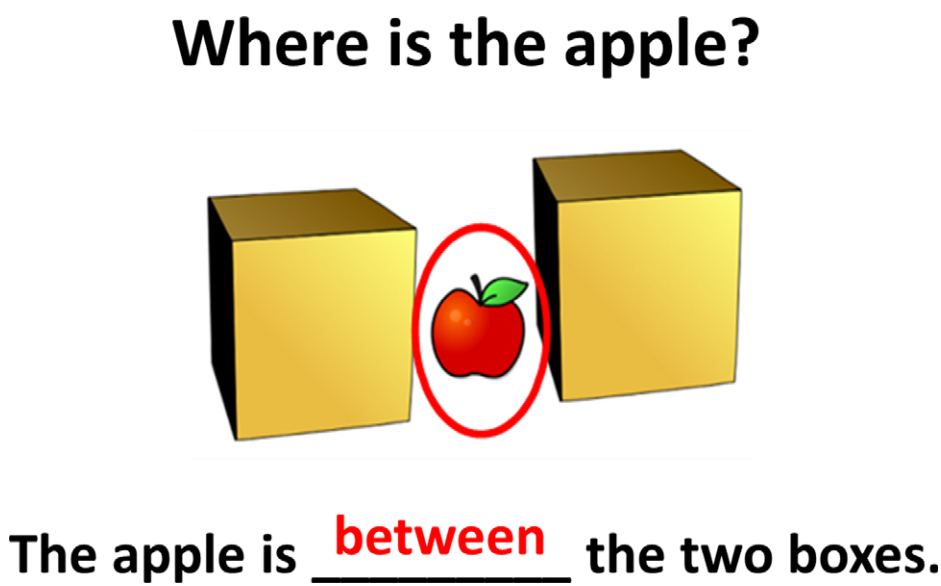

FIGURE 1. An example slide from PowerPoint presentations used in the training phase of the experiment.

In addition, crossword puzzles were used across the training and testing phases of this study. These puzzles were created using a free online crossword puzzle generator.[1] Four crossword puzzles were created for each of the two lists of words (eight unique crossword puzzles in total; four puzzles for List A and four puzzles for List B). Half the crossword puzzles were used for the testing phase of the experiment and half were used for the training phase of the experiment. The clues were identical across all puzzles in that they were taken from the course materials. While the clues did not vary across the test versions, the order of the items and design of the crossword puzzles were unique for each version of the puzzle. An example of one puzzle is available as a supplementary file.

The materials for the testing phase were four crossword puzzles designed with the same considerations in mind as those used during the training phase. Two of these puzzles (one for Set A and one for Set B) were administered as the pretest; two of the puzzles (one for Set A and one for Set B) were administered as the posttest. It was decided to use separate

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

TABLE 3. Experimental design

	Monday	Tuesday	Wednesday	Thursday	Friday
Week 1			Prebriefing		
Week 2			Pretest		
Week 3					
Week 4			Training Session 1 (items 1–10)		
Week 5			Training Session 2 (items 11–20)	Training Session 3 (items 1–20)	
Week 6					
Week 7					
Week 8					
Week 9				28-day RI Posttest	

puzzles rather than combining the two sets into a larger set of 20 so that the testing phase more closely matched the training conditions. This was done on a theoretical level so that the testing conditions would more closely match the training conditions, which would in theory lead to better transfer of learning (e.g., Lightbown, 2008). On a practical level, the teachers in the study felt that a 20-item crossword puzzle would be too difficult to manage and that administering two shorter puzzles would be more appropriate for the students.

Only one posttest was administered to help control for testing effects as a potential confounding variable (Rogers & Cheung, 2018; Suzuki, 2017). The internal reliability of the two test versions ($\alpha = .86$ and .95) was acceptable, given Plonsky and Derrick's (2016) guidelines. The entire experiment, from prebriefing to postbriefing sessions, took 9 weeks in total. It comprised a prebriefing session, three training sessions, and two testing sessions (a pretest and a 28-day delayed posttest). The temporal distribution of the learning of the target vocabulary items was manipulated within-subjects and was identical to the spacing conditions adopted in Rogers and Cheung's (2018) study. This was done by first dividing the 20 target items into two lists of 10 items apiece. One of these lists was first studied in Training Session 1 (henceforth spaced-long items). The other list was first studied in Training Session 2 (henceforth spaced-short items). The order that the lists were studied was counterbalanced across the different classes that took part in the study. Both lists were reviewed in Training Session 3. This resulted in an 8-day ISI for the items studied in Session 1 and reviewed in Session 3, and a 1-day ISI for items studied in Session 2 and reviewed in Session 3.

The procedural timeline for the experiment can be seen in Table 3. All target items included in this study had not been previously taught by the instructors in the course. A prebriefing session was held with the teachers, led by a member of the research team. As part of this session, the consent forms were distributed and the experimental procedures and protocols were discussed. Pretests were then administered in class in the following week. Two weeks later, all three classes took part in Training Session 1, in which 10 spaced-long items were introduced and taught using the prescribed materials. The spaced-long items were counterbalanced among the three classes, with Class A and Class C studying one set, and Class B studying the other. One week later, the 10 spaced-short items were introduced and taught using the prescribed materials in Training Session 2. The following day, Training Session 3 took place, in which all 20 target items were

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

reviewed. The procedure for each training session was agreed upon through a collaborative discussion between the school principal, teachers who agreed to take part in the study, and the research team. Each training session consisted of an initial PowerPoint presentation that the teachers used to introduce the target items. This was followed by a crossword puzzle to practice the target items, and concluded with feedback from the teachers. Training Sessions 1 and 2 took approximately 15 minutes to complete. Training Session 3 took approximately 25 minutes to complete because of the higher number of target vocabulary items.

The surprise posttest was administered 28 days after Training Session 3 (i.e., the review session), resulting in an ISI/RI ratio of 3.6% and 28.6% for the spaced-short and spaced-long items, respectively. The timing of the posttest was identical to that of Rogers and Cheung's (2018) original study, which is based on providing the optimum ISI/RI ratio for the spaced-long condition as per Cepeda et al.'s (2008) model. Following the posttest, interviews were carried out with the teachers to discuss the results and the degree (if any) to which they deviated from the agreed-upon experimental procedure.

SCORING AND STATISTICAL ANALYSES

Each crossword puzzle consisted of 10 items. Two separate dichotomous scoring methods, strict and lenient, were used to score the data, and separate analyses were run for each. Under strict scoring, each item was scored as 1 if participants provided the correct spelling of the target item, and 0 if the correct spelling was not provided, that is, if the participant made any spelling mistake, no matter how minor, then the item was counted as incorrect. Under lenient scoring, 1 point was awarded if the participants provided a correct spelling or an incorrect spelling that did not impede the completion of the crossword puzzle, and 0 points were awarded for cases in which students could not provide the missing information. Approximately 25% of the scripts were double marked independently by members of the research team. A reliability analysis showed a high level of agreement between the markers: $r\,(70) = .992$, $p < .001$ (see Plonsky & Derrick, 2016 for a discussion of reliability standards in SLA research). Visual inspection of the data revealed a skewed distribution of data. A Shapiro–Wilk test of normality confirmed a nonnormal distribution for both the pretest ($W = .946$, $p = .006$) and posttest scores ($W = .799$, $p < .001$) for strict scoring, and a normal distribution for the pretest using lenient scoring ($W = .968$, $p = .088$) and nonnormal distribution for the posttest ($W = .729$, $p < .001$).

Two different statistical approaches were undertaken to analyze the data. First, given the nonnormal distribution for three out of the four sets of data, nonparametric statistical procedures, specifically Mann–Whitney U tests and Wilcoxon signed rank tests, were carried out using SPSS V25. We elected to use these nonparametric tests as they were identical to the tests run in Rogers and Cheung's (2018) study, thus allowing for greater comparability. The alpha level for these tests was set at .05. Effect sizes were calculated in Pearson's correlation coefficient r. For interpreting effect sizes, we adopt Plonsky and Oswald's (2014) field-specific guidelines of $r = .25$ as small-, .4 as medium-, and .6 as large-sized effects, respectively.

In addition to the aforementioned nonparametric tests, the data were further analyzed using more current methods, specifically a series of logit mixed-effects models (Linck &

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

Cunnings, 2015). These analyses were carried out using the lme4 package in R (Bates et al., 2015). Within these models, distribution (spaced-short vs. spaced-long), time (pretest vs. posttest), and class (Class A vs. Class B vs. Class C) were included as fixed effects with crossed random effects for subjects and items. The models were developed incrementally using a maximum-likelihood technique (Cunnings, 2012; Linck & Cunnings, 2015; see also Rogers, 2017b; Suzuki & Sunada, 2019 for similar approaches). First, an initial null model was created using only random intercepts for participants and items. Following this, the fixed effects were added incrementally and each model was compared against the null model using the *ANOVA* function in the lme4 package. This was followed by random slopes. If the model that included the fixed effect was significant against the null model, this result was interpreted as indicating that the fixed effect in question had a significant relationship with the dependent variable and should be included in any subsequent analyses. If the result was nonsignificant, this was interpreted as no significant relationship, and that this fixed effect could be excluded from any models to follow. The best-fitting model was then analyzed to determine which fixed effects, if any, reached statistical significance.

RESULTS

Descriptive statistics were generated with regard to pretest and posttest performance for the three classes and further broken down for spaced-short and spaced-long items. These results are presented in Table 4 for strict scoring and Table 5 for lenient scoring. First, we examined total learning across all participants and item types using a Wilcoxon signed-ranks test. This analysis revealed a significant difference in median ranks from pretest to posttest across all participants with a large-sized overall effect using strict scoring ($Z = -6.863$, $p < .001$, $r = .60$), and similar results under lenient scoring ($Z = -6.768$, $p < .001$, $r = .59$).

Similar analyses were run to examine spaced-short and spaced-long items independently. These analyzes showed a significant result for both spaced-short ($Z = -6.646$, $p < .001$, $r = .56$) and spaced-long items ($Z = -6.566$, $p < .001$, $r = .57$) for strict scoring, and for spaced-short ($Z = -6.257$, $p < .001$, $r = .54$) and spaced-long items ($Z = -6.320$, $p < .001$, $r = .55$) under lenient scoring, all of which indicated medium-sized effects. Finally, to compare scores across spaced-short versus spaced-long items, gain scores were first calculated from pretest to posttest, then these scores were compared using Mann–Whitney U tests. This test indicated a nonsignificant result between spaced-short and spaced-long items with a small-sized effect under strict ($U = 2{,}149.5$, $p = .90$, $r = .01$) and lenient scoring ($U = 1{,}979$, $p = .36$, $r = .08$).

As noted, in addition to the nonparametric tests mentioned previously, a series of logit mixed-effects analyses were carried out on the data. To interpret the data provided by these models, if a fixed effect reaches significance within the model, then this indicates that the effect in question is significant in accounting for variance in the overall dataset. As an example, if a model indicates that the fixed effect of distribution is significant, this points toward a significant difference in accuracy between the variables included within the fixed effect of distribution, that is, spaced-short versus spaced-long items. In addition to significant main effects, the mixed-effects model can also indicate a significant interaction between fixed effects. For example, if the model results in a significant

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

TABLE 4. Descriptive statistics of performance on pretest, posttest, and gain scores (%) under strict scoring

		Pretest				Posttest				Gain Scores			
Overall		Mdn	M	SD	SE	Mdn	M	SD	SE	Mdn	M	SD	SE
$N = 66$	Total	22.50	27.50	20.10	2.47	90.00	76.52	29.04	3.57	47.50	49.02	26.65	3.28
	Spaced-short	20.00	24.39	23.08	2.84	90.00	73.64	32.52	4.00	50.00	49.24	32.41	3.99
	Spaced-long	20.00	30.61	27.95	3.44	100.00	79.39	30.73	3.78	50.00	48.79	32.79	3.28
Class A													
$n = 26$	Total	40.00	38.08	20.00	3.92	95.00	85.19	23.60	4.63	45.00	47.12	29.23	5.73
	Spaced-short	30.00	30.77	23.14	4.54	90.00	80.77	28.13	5.52	50.00	50.00	33.82	6.63
	Spaced-long	50.00	45.38	29.15	5.72	100.00	89.62	21.44	4.21	40.00	44.23	35.91	7.04
Class B													
$n = 16$	Total	20.00	20.31	19.53	4.88	65.00	56.87	31.24	7.81	30.00	36.56	22.64	5.66
	Spaced-short	25.00	31.25	27.29	6.84	65.00	58.75	35.19	8.80	35.00	27.50	23.80	5.95
	Spaced-long	0.00	9.38	19.14	4.74	50.00	55.00	36.51	9.13	40.00	45.62	34.44	8.61
Class C													
$n = 24$	Total	20.00	20.83	15.79	3.22	90.00	80.21	27.80	5.67	65.00	59.38	22.81	4.66
	Spaced-short	10.00	12.92	14.88	3.04	95.00	75.83	33.22	6.78	70.00	62.92	28.81	5.88
	Spaced-long	20.00	28.75	21.93	4.48	100.00	84.58	27.18	5.55	60.00	55.83	27.96	5.71

TABLE 5. Descriptive statistics of performance on pretest, posttest, and gain scores (%) under lenient scoring

		Pretest				Posttest				Gain Scores			
Overall		Mdn	M	SD	SE	Md	M	SD	SE	Mdn	M	SD	SE
$N = 66$	Total	30.00	34.70	22.08	2.72	95.00	79.70	29.21	3.60	47.50	45.00	27.15	3.34
	Spaced-short	30.00	29.85	25.02	3.08	100.00	77.46	33.00	4.03	50.00	47.27	33.45	4.12
	Spaced-long	40.00	39.55	31.79	3.91	100.00	82.27	30.37	3.74	40.00	42.73	33.45	4.12
Class A													
$n = 26$	Total	45.00	46.92	19.60	3.84	100.00	89.42	22.99	4.51	45.00	42.50	30.27	5.94
	Spaced-short	35.00	35.00	23.02	4.52	100.00	85.93	27.63	5.32	50.00	50.38	35.04	6.87
	Spaced-long	60.00	58.85	27.90	5.47	100.00	93.46	20.58	4.04	30.00	34.62	35.91	7.04
Class B													
$n = 16$	Total	25.00	25.31	21.41	5.35	70.00	61.88	33.41	8.53	35.00	36.56	25.61	6.40
	Spaced-short	45.00	40.63	31.30	7.83	75.00	64.38	37.77	9.44	30.00	23.75	27.05	6.76
	Spaced-long	0.00	10.00	18.97	4.74	60.00	59.38	37.50	9.38	45.00	49.38	34.15	8.54
Class C													
$n = 24$	Total	25.00	27.71	19.50	3.98	95.00	81.04	27.90	5.69	60.00	53.33	23.02	4.70
	Spaced-short	15.00	17.08	16.54	3.38	100.00	76.67	33.58	6.84	65.00	59.58	28.20	5.76
	Spaced-long	35.00	38.33	27.29	5.56	100.00	85.42	26.70	5.45	55.00	47.08	29.56	6.03

interaction between distribution and time, then this result points toward differences between pretest and posttest performance with regard to spaced-short versus spaced-long items.

The results of these models corroborated the results of the nonparametric tests mentioned in the preceding text in that a significant effect was found for the fixed effect of time, indicating that all participants' performance improved significantly from pretest to posttest. Most importantly, the fixed effect of distribution and its interactions with group and time were all nonsignificant, indicating that the distribution of items, spaced-short versus spaced-long, did not differ significantly with regard to pretest and posttest scores, and across the different groups, that is, classes that took part in the study. The results of the best-fitting models for both strict and lenient scoring can be found in Table 6.

QUALITATIVE RESULTS

The main objective of the postexperiment interviews was to collect information from the teachers as to whether and to what degree they deviated from the experimental procedure and to confirm that they maintained consistency across the different training sessions during the experiment.

Overall, all three teachers reported that they were consistent in their approaches across the training stages of the experiment and utilized the experimental materials as intended. Teacher A reported playing a miming game with students after presenting the PowerPoint and prior to asking students to complete the practice crossword puzzles during the training sessions. While going through the PowerPoint, the teacher also highlighted the past tense forms (e.g., *sat*, *roared*, *walked*) of the verbs in the study, by explicitly pointing these out to the students during the presentation. Teacher B reported that they did not deviate from the procedure and went through the PowerPoint with students, asking them to remember the words, prior to completing the practice crossword puzzles. Teacher C reported following the experimental procedure but also devised an additional activity where they showed the pictures one by one to the students and asked the students to spell the words to their partner. During the presentation and feedback, they reported highlighting the spelling of words that they perceived to be difficult for the students, such as *poured* and *roared.* All teachers reported checking the answers to the crossword puzzles with the students during the training and review sessions by eliciting the answers from the students and showing the students the correct answer on the overhead projector.

DISCUSSION

This study examined the optimum learning schedules of L2 vocabulary in child classrooms under ecologically justified learning conditions. Participants learned the target vocabulary items in two learning sessions following either a spaced-short (1-day ISI) or a spaced-long condition (8-day ISI). Learning was measured through a 4-week delayed posttest, which consisted of crossword puzzles that were transfer-appropriate to the training conditions of the study. The results here indicated medium to high learning gains across all target items, with no significant differences between the two spacing conditions.

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

TABLE 6. Results of best-fitting models of logit mixed-effects models for strict and lenient scoring

	Strict Scoring				Lenient Scoring			
	Participant		Item		Participant		Item	
Random effects	Variance	*SD*	Variance	*SD*	Variance	*SD*	Variance	*SD*
Intercept	13.75	3.71	1.93	1.39	22.79	4.77	3.42	1.85
Time	10.21	3.20	–	–	20.08	4.48	–	–
Distribution	12.95	3.60	–	–	9.54	3.09	–	–
Time*Distribution	6.30	2.51	–	–	6.52	2.55	–	–
Group	–	–	.09	.30	–	–	.11	.33
Fixed effects	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Intercept	−5.29	1.58	−3.35	<.001	−8.97	2.64	−3.49	<.001
Time	4.67	1.32	3.53	<.001	8.85	2.48	3.57	<.001
Group	−.96	.70	−1.37	.17	.29	1.01	.29	.78
Distribution	2.36	1.71	1.38	.17	3.35	2.59	1.29	.20
Time*Group	.34	.58	.58	.56	−.97	.93	−1.04	.30
Time*Distribution	−1.74	1.27	−1.37	.17	−2.68	2.38	−1.12	.26
Group*Distribution	−.44	.78	−.56	.58	−.87	.96	−.90	.37
Time*Group*Distribution	.28	.56	.50	.62	.74	.83	.90	.37

Note: Time = pretest vs. posttest; distribution = spaced-short vs. spaced-long; group = Class A vs. Class B vs. Class C. * denotes interaction. Model Formula: accuracy ~ time*-group*distribution + (time*distribution|participant) + (grouplitem), glmerControl (optimizer = "bobyqa"), family = binomial.

These results are in line with the existing research that has examined lag effects in child L2 classrooms (Kasprowicz et al., 2019; Küpper-Tetzel et al., 2014; Rogers & Cheung, 2018; Serrano & Huang, 2018), where no advantages have been found for spaced-long conditions in comparison with spaced-short conditions. Taken as a whole, these results provide growing evidence that lag effects, in particular the optimum ISI/RI ratios reported in the cognitive psychology literature, might not translate directly to conditions with lower levels of experimental control.

Across the SLA literature, there is mixed empirical evidence as to the advantages of distributed practice. Some studies have provided support for the ISI/RI ratio framework proposed within the cognitive psychology literature (Cepeda et al., 2008) when examining the learning of isolated L2 grammatical structures (Bird, 2010; Rogers, 2015). Other studies, however, have returned conflicting results with advantages for more intensive conditions, or similar amounts of learning across both conditions (Suzuki, 2017; Suzuki & DeKeyser, 2017). Studies that have examined more intensive versus more extensive learning schedules on a programmatic level have also reported either advantages for more intensive learning conditions or similar results across conditions (e.g., Collins & White, 2011; Serrano, 2011; Serrano & Muñoz, 2007). Taken together with the results from the present study and the four child L2 classroom-based studies cited in a preceding paragraph, there does not appear to be a clear advantage at present for longer spacing conditions across the extant SLA literature.

There are a number of plausible explanations for this trend. The first, as has been suggested in the literature (e.g., Rogers, 2017a; Serrano, 2012) is that the optimum ISI/RI intervals of Cepeda et al. (2008) are not directly applicable to SLA, due to the fact that SLA is arguably more complex that the data on which Cepeda et al.'s (2008) model is based, that is, the learning of trivia facts (see Serrano, 2012 for a discussion). This interpretation is speculative as the majority of SLA studies to date, including the present study, have justified the timing of their delayed posttests in light of Cepeda et al.'s (2008) model, with mixed results. A systematic exploration of the learning of SLA content over a wide range of other ratios would provide evidence necessary for this interpretation. In this regard, future SLA research might begin by establishing the validity of ISI/RI ratios to the learning of SLA content under controlled laboratory settings, followed by research to explore the degree that this generalizes to authentic teaching and learning contexts. Another possibility is the methodological differences between SLA studies and those in cognitive psychology. One difference, as noted in the preceding text, is that posttests are typically manipulated experimentally between-participants in studies in cognitive psychology, whereas SLA studies tend to do so within-participants, thus creating a potential confound due to repeated retrieval opportunities. A final possibility is that the benefits of distributed practice have been overstated in the literature and previous findings are not robust in the face of the increased variability present in authentic classroom environments. This is evident in the current study, where the results show a high degree of variability within groups, as reflected by the standard deviations within each group. Regardless of the explanation, it is clear that further research, both laboratory- and classroom-based, is needed to substantiate any claims as to the benefits of longer spacing conditions for SLA and, in particular, instructed SLA.

A further point of discussion concerns the conceptualization of spacing within the broader SLA and cognitive psychology literature. As noted, studies interested in spacing operationalize retention with regard to the distance from the final training session and the test, in other words the RI. However, as pointed out by a reviewer, the distance from the initial training session to the testing phase is also valuable from a pedagogical perspective in that instructors and students may be concerned with the amount of vocabulary that can be retained for long periods following initial exposure, as well as following later review. Future SLA research may explore total retention time (i.e., time from initial exposure to testing) as a potential moderating variable on the effects of spacing.

It is also important to highlight some of the methodological limitations of the present study. Unlike the previous research on which the current experiment is based (Rogers & Cheung, 2018), we were unable to carry out lesson observations due to reasons of practicality/agreement with the teachers and other stakeholders in the host school. Therefore, we relied on postobservation interviews with the teachers to collect data regarding their fidelity to the experimental protocols. Although the teachers reported using the materials as intended, and we have no reason to believe otherwise, it is important to acknowledge that we do not have direct evidence of the teachers' classroom practice throughout this study, and their responses may have been influenced by, for example, social desirability bias. Two other limitations concern measurement. First, as the same clues were included across the crossword puzzles used throughout the experiment, it is possible that the results here might have been influenced by a testing effect. Second, the outcome measure used in the current study, crossword puzzles, only captures one aspect of vocabulary knowledge, specifically form recall (Schmitt, 2010). It would be advantageous for future research to utilize multiple measures of vocabulary knowledge, which would provide a more complete picture of lexical development (Webb, 2005).

A final point concerns ecological validity. As noted, some previous quasi-experimental distribution of practice research has taken a narrow view of ecological validity in operationalizing this construct as an experimental study that takes place in a classroom setting, regardless of the levels of experimental control and authenticity of the experimental manipulations. To truly make any claims about the degree to which research findings generalize to authentic learning environments, SLA and otherwise, interventions need to be empirically tested within authentic environments with higher degrees of ecological validity. Researchers might do so by validating and justifying their experimental interventions and instruments with regard to the context they aim to generalize to (Lightbown & Spada, 2019; Rogers & Révész, 2020). In doing so, a parallel decrease in effect sizes might be expected. However, if the effects of an intervention cannot be seen in a quasi-experimental study that is ecologically valid in its experimental manipulations, we question whether any effects of the intervention would be seen when implemented "at the chalkface," that is, in real teaching and learning environments.

CONCLUSION

By way of conclusion, we would like to comment briefly on the practical implications of this research. It is unlikely that anyone, whether researcher or teaching practitioner, would question that review and revision are beneficial for learning. On a theoretical level, such benefits find support in, for example, skill-based theories of language learning (e.g.,

https://doi.org/10.1017/S0272263120000236 Published online by Cambridge University Press

DeKeyser, 2017), where repeated practice is necessary for proceduralization and automatization to take place. The question posed by studies examining distribution of practice effects is not whether, but rather when, and how often, review should take place to optimize learning (Suzuki et al., 2019). With regard to how often, there is some unsurprising evidence that "more is better" (Bahrick et al., 1993). When it comes to when a review should take place, it perhaps goes without saying that a review period close to the assessment is likely to be most beneficial to students' performance on the assessment (Cepeda et al., 2008). With regard to the long-term retention of L2 vocabulary studied by young learners as part of normal classroom instruction, what evidence there is indicates that the timing of the review does not significantly influence learning or retention. In other words, the practical takeaway for language teachers from this growing body of research is that it does not appear to matter when L2 vocabulary is reviewed as part of classroom instruction, as long as it *is* reviewed.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263120000236.

NOTE

[1]https://worksheets.theteacherscorner.net/make-your-own/crossword/

REFERENCES

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*, 316–321.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

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

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236–246.

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.

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.

Cobb, T. (2016). VocabProfile [computer software]. http://www.lextutor.ca/vp/eng

Collins, L., & White, J. (2011). An intensive look at intensity and language learning. *TESOL Quarterly*, *45*, 106–133.

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

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

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen, & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15–32). Routledge.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Hatch, E. (1979). Apply with caution. *Studies in Second Language Acquisition*, *2*, 123–143.

Hulstijn, J. H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, *19*, 131–143.

Kasprowicz, R. E., & Marsden, E. (2018). Towards ecological validity in research into input-based practice: Form spotting can be as beneficial as form-meaning practice. *Applied Linguistics*, *39*, 886–911.

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

Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, *20*, 37–47.

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

Lightbown, P., & Spada, N. (2019). *In it together: Teachers, researchers, and classroom SLA*. Plenary presented at the annual meeting of the American Association of Applied Linguistics. Atlanta, GA.

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

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207.

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

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

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*, 559–586.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, *100*, 538–553.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Porte, G., & McManus, K. (2018). *Doing replication research in applied linguistics*. Routledge.

Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*, 857–866.

Rogers, J. (2017a). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, *38*, 906–911.

Rogers, J. (2017b). Awareness and learning under incidental learning conditions. *Language Awareness*, *26*, 113–133.

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

Rogers, J., & Leow, R. P. (2020). Toward greater empirical feasibility of the theoretical framework for systematic and deliberate L2 practice: Comments on Suzuki, Nakata, & DeKeyser (2019), *Modern Language Journal*, *104*, 309–312.

Rogers, J., & Révész, A. (2020). Experimental and quasi-experimental designs. In J. McKinley, & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 133–143). Routledge.

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review*, *27*, 635–643.

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, *16*, 183–186.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Press.

Serrano, R. (2011). The time factor in EFL classroom practice. *Language Learning*, *61*, 117–145.

Serrano, R. (2012). Is intensive learning effective? Reflecting on the results from cognitive psychology and the second language acquisition literature. In C. Muñoz (Ed.), *Intensive exposure experiences in second language learning* (pp. 3–22). Multilingual Matters.

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.

Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System*, *35*, 305–321.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Spada, N. (2005). Conditions and challenges in developing school-based SLA research programs. *The Modern Language Journal*, *89*, 328–338.

Spada, N. (2015). SLA research and L2 pedagogy: Misapplications and questions of relevance. *Language Teaching*, *48*, 69–81.

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

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

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*, *103*, 1–29.

Suzuki, Y., Nakata, T., & DeKeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *Modern Language Journal*, *103*, 713–720.

Suzuki, Y., Nakata, T., & DeKeyser, R. (2020). Empirical feasibility of the desirable difficulty framework: Toward more systematic research on L2 practice for broader pedagogical implications. *Modern Language Journal*, *104*, 313–319.

Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. *Psychology of Learning and Motivation*, *60*, 113–189.

Van Lier, L. (2010). The ecology of language learning: Practice to theory, theory to practice. *Procedia-Social and Behavioral Sciences*, *3*, 2–6.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*, 33–52.