Language Teaching (2021), **54**, 424–433
doi:10.1017/S0261444820000439

CAMBRIDGE UNIVERSITY PRESS

REPLICATION RESEARCH

Input spacing in second language classroom settings: Replications of Bird (2010) and Serrano (2011)

John Rogers

English Language Education, The Education University of Hong Kong, Hong Kong
Email: rjrogers@eduhk.hk

Abstract

This paper proposes the replication of Bird's (2010) and Serrano's (2011) studies, which examined distribution of practice effects in second language acquisition (SLA). These studies, which took place in authentic classroom contexts, produced conflicting results regarding the degree to which the learning of a second language (L2) benefited from distributed instruction. In the first part of the paper, I discuss the distribution of practice research in the learning and teaching of L2s. I then describe Bird's (2010) and Serrano's (2011) work, and highlight the strengths and limitations of the approaches of these studies. Finally, a number of approaches to approximate replications are suggested for each study in order to assess the reliability, internal validity, and generalizability of the original findings

1. Introduction

It is well acknowledged that there is limited scope for sufficient practice within a typical L2 classroom. This is true with regard to the number of practice opportunities, as well as to the total amount of time available (DeKeyser, 2017; Lightbown & Spada, 2019). One area that has been suggested as influencing the effectiveness of practice is how learning is distributed over a given period of time. *Distributed learning*, or *input spacing*, refers to the degree that learning/practice takes place over multiple training sessions. *Massed learning* refers to practice that is compressed or concentrated into a single session (such as staying up all night cramming before a final exam). Practice that is distributed, or spaced, takes place in two or more training episodes (such as studying a little every day in the days leading up to an exam). *The spacing effect* refers to the apparent superiority of distributed learning over massed learning with regard to learning and retention (Rohrer, 2015; Rogers, 2017). The term *lag effect* is used to describe the effects of spacing gaps of varying lengths, such as the impact of having a short break between training sessions versus having a longer gap, such as several days. The term *distributed practice effect* has been suggested as a blanket term to encompass both lag effects and spacing effects (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Toppino & Gerbier, 2014; Rogers, 2017).

An understanding of distribution of practice effects has important theoretical and practical pedagogical implications with regard to how L2s are learned and how they might best be taught (Lightbown & Spada, 2019; Suzuki, Nakata, & DeKeyser, 2019, 2020; Rogers & Leow, 2020). With regard to curricula, distribution of practice can inform course design on multiple levels, such as how frequently class meetings should occur and how long they should last. Classes may take place every day over a relatively short period, such as an intensive language course over the summer months, or may be spread over a longer period, such as over the course of an entire school year. Similarly, distribution of practice effects can influence the content or teaching schedule within a curriculum. For example, teachers may revisit/review previously taught material in the days prior to a test, or this might be structured so that content is revisited and reviewed regularly and systematically over the course of the entire semester (e.g., Nakata, Tada, Mclean, & Kim, 2020).

© The Author(s), 2020. Published by Cambridge University Press

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

Empirical research in the field of SLA has investigated the effects of spacing at the curricular level (e.g., Spada & Lightbown, 1989; Lightbown & Spada, 1994; Lapkin, Hart, & Harley, 1998; Freed, Segalowitz, & Dewey, 2004; White & Turner, 2005; Serrano & Muñoz, 2007; Collins & White, 2011; Serrano, 2011; for reviews, see Serrano, 2012; Muñoz, 2014). These studies, which have largely been observational in nature, have been at least partially motivated by examining the effectiveness of existing L2 learning programmes, and have generally been more concerned with global aspects of L2 development, such as overall language proficiency. The results of these studies have been remarkably consistent in providing support for more intensive learning conditions/periods of instruction.

Furthermore, a number of more experimental studies have examined the learning of different aspects of an L2, such as vocabulary learning (e.g., Nakata, 2015; Rogers & Cheung, 2018, 2020; Serrano & Huang, 2018; Nakata & Suzuki, 2019; Nakata & Elgort, 2020), as well as the learning of L2 grammar (Bird, 2010; Miles, 2014; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017; Kasprowicz, Marsden, & Sephton, 2019). These studies have tended to be more focused in that they have been concerned with isolated aspects of L2 development, such as the learning of a particular set of vocabulary items/grammatical structures. These studies have also been motivated more explicitly by empirical research in the field of cognitive psychology, which has examined the relationship between the time between training sessions (the intersession interval, or ISI) and the time between training and testing (the retention interval, or RI). Work in the field of cognitive psychology has suggested a relationship between these variables (Glenberg, 1976; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), and mathematical models of the ISI/RI relationship have suggested that there is an optimal ISI/RI ratio, in which the optimal ISI is approximately 10–40% of the RI, although this varies to some extent depending on the nature of the task (Cepeda et al., 2008).

SLA studies have increasingly begun to examine whether the optimal ISI/RI ratio suggested by Cepeda et al. (2008) necessarily applies to SLA. In contrast to the relatively consistent findings of the curricular-level research mentioned above, the results of these studies have been mixed, and have painted an unclear picture regarding the degree to which L2 learning benefits from distribution of practice, with some studies finding support for the claim that distributed practice results in more robust learning (e.g., Bird, 2010), and other studies failing to find similar effects (e.g., Suzuki, 2017). Given the conflicting findings in the extant research, and the methodological limitations of this body of research, it is not currently possible to state with confidence that the benefits of distributed practice have been demonstrated robustly in empirical SLA research (Rogers, 2017).

An understanding of the effects of practice on learning is important for the field as it can help to inform both SLA theory and pedagogical practice (Suzuki et al., 2019, 2020; Rogers & Leow, 2020). This importance is increasingly being recognized within the field, and recent years have seen a growing number of studies in this area, with some isolated attempts to replicate studies that have failed to produce results in the predicted direction (Rogers & Cheung, 2020; Suzuki, 2017). Although it is undeniable that such attempts at replication are steps in the right direction, extant research has generally suffered from a number of methodological limitations. Accordingly, additional replication efforts are warranted in order to establish the reliability, internal validity, and generalizability of previous findings.

In this article, I propose replications of two studies examining distribution of practice effects in SLA: an experimental study by Bird (2010), and a curricular-level investigation by Serrano (2011). There are a number of theoretical, pedagogic, and methodological justifications for replications of these studies. First, on a theoretical level, replications of both of these studies would provide additional evidence regarding the degree to which the optimal ISI/RI ratios proposed within cognitive psychology (Cepeda et al., 2008) apply to SLA. Furthermore, distribution of practice effects is linked closely to the benefits of repetition, review, and practice. Such benefits have found support in skill-based theories of SLA (e.g., DeKeyser, 2017), and in recent frameworks emphasizing desirable difficulties in learning (Lightbown, 2008; Suzuki et al., 2019, 2020; Rogers & Leow, 2020). Replications might also assist in addressing the methodological limitations of Bird's (2010) and Serrano's (2011) studies, such as the lack of delayed testing or confounds related to repeated retrieval opportunities at testing. Finally,

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

Bird's (2010) and Serrano's (2011) studies are among the few studies to have investigated distribution of practice effects in authentic classroom contexts (see also Miles, 2014; Kasprowicz et al., 2019). Replications of these studies would meet wider calls for ecologically valid classroom-based research within the field of SLA (Spada, 2005, 2015; Kasprowicz & Marsden, 2018; Lightbown & Spada, 2019; Rogers & Révész, 2020).

2. Study 1: Bird (2010)

There are a number of factors that makes Bird's influential (2010) study a prime candidate for replication. First, Bird's (2010) study was the first SLA study to investigate the degree to which the optimum ISI/RI ratios proposed by cognitive psychology (Cepeda et al., 2008) apply to the field of SLA. Bird's (2010) study is widely cited as an example of the benefits of distributed learning over massed learning within the field of SLA. It is also frequently cited further afield across the educational sciences as evidence of the degree to which the benefits of distributed practice can be generalized across different content areas (see, e.g., review articles by Rohrer, 2015, Carpenter, 2017, among others). Finally, Bird's (2010) study provided positive evidence of the benefits of distributed practice for the learning and long-term retention of L2 grammar. Although similar studies that failed to reveal such benefits have been replicated (e.g., Suzuki & DeKeyser, 2017; Suzuki, 2017; see also Rogers & Cheung, 2018, 2020), to the best of my knowledge, no study that has shown the positive benefits of input spacing for the learning of L2 grammar has been replicated/reproduced to date.

Bird (2010) investigated the effects of input spacing on the learning of L2 grammar. In this study, 38 learners of English (university students, Malay first language (L1); roughly upper-intermediate level of proficiency: IELTS 5.8 to 6) were trained to differentiate between two different verb distinctions, namely past simple versus present perfect, and present perfect versus past perfect. The materials for this study were form-focused worksheets that presented the target grammatical items in an isolated, decontextualized manner. Two sets of materials were created for this study. The first set consisted of 160 sentences, half of which were simple past sentences (for example, 'I saw that movie with my brother last week'), and half were present perfect sentences (such as 'Grant has just finished his breakfast'). The second set consisted of 160 sentences, which were divided into 80 present perfect sentences and 80 past perfect sentences (such as 'I hadn't seen that movie before she recommended it'). Each of the 320 sentences was unique. These sentences were then divided to create eight worksheets for each set, each of which contained twenty sentences. In each of the worksheets, 15 items were modified randomly to be erroneous sentences in which the main verb phrase was incorrect (for example, 'I have seen* that movie with my brother last week'). Ten of the worksheets were used as part of the training phase. The remaining six worksheets were used for the testing phase of the study. During the training phase, participants were reminded that the error was only with regard to the verb phrase used, thus focusing the participants' attention on this aspect.

The study followed a pretest, posttest, delayed posttest design. Spacing was manipulated within-participants, with one group of participants studying the past simple/present perfect materials under short spacing conditions, and the present perfect/past perfect materials under long spacing conditions; the second group had this order reversed. The subjects participated in five training sessions in total. The short-spaced conditions followed a three-day ISI, and the long-spaced conditions a 14-day ISI. The practice sessions lasted for approximately one hour in total, and included intentional, form-focused language practice. The students were presented with practice sheets that mirrored the format of the testing phase. After 30 minutes, the participants were given the correct answers, as well as brief explanations of why the answers were correct. The explanations were scripted to ensure comparability across the different groups.

Two posttests were administered following a seven-day and a 60-day RI. This resulted in ISI/RI ratios of 42% and 5% for the three-day ISI, and 200% and 23% for the 14-day ISI on the seven-day and 60-day delayed posttests, respectively. As noted, the test materials were identical to those used in the training stage of the study. Each worksheet consisted of 20 statements, of which 15 were

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

incorrect and five were correct. The participants were required to judge the grammaticality of each statement and, in the case of ungrammatical sentences, to correct them. One mark was awarded for the correct identification of an item as being correct/incorrect. A second mark was allocated for providing the correction. Following the seven-day RI posttest, the instructor asked the groups if they had noticed the different practice schedules or studied the material outside of the class. None of the participants reported studying for more than a few minutes outside of the class.

Data were analysed using conventional inferential statistics, namely a mixed between-within ANOVA, followed by planned comparisons using *t*-tests at the different time points. The results of the study indicated that both spacing conditions resulted in significant improvements from the pretest to the seven-day delayed posttest, with no significant differences between the two lag conditions. In other words, no advantages were seen for either spacing condition following a seven-day delay. However, at 60-day delayed testing, the three-day ISI showed a significant decrease in retention, whereas the 14-day day condition did not, resulting in significantly higher performance for 14-day ISI conditions over three-day ISI conditions. Thus, the results of the study suggest that the spacing between training sessions influences the durability of learning that occurs across different linguistic materials. These results are also in line with cognitive psychology studies investigating the impact of learning schedules (e.g., Cepeda et al., 2008).

2.1 Approaches to replication

A replication of Bird's (2010) study would be of undoubted interest to SLA researchers, researchers in cognitive psychology and in the educational sciences, as well as to teachers and learners interested in how instruction and study time might be optimized. Although an exact replication of the original study would be valuable, I propose several avenues for an approximate replication. The original study was a relatively small-scale study that investigated the learning of L2 grammar by university students in a classroom setting. One obvious avenue for replication would be to adhere to the original experimental procedures and materials as closely as possible, but to involve a different population of learners from that in the original study (university students, L1 Malay). Apart from examining participants with a different L1 from those in the original study, future research might examine populations of different ages, such as younger learners, older learners, or other non-traditional populations. Sampling across diverse populations of learners is important because the vast majority of distributed practice research across the cognitive sciences has taken place with adult populations (Cepeda et al., 2006). Furthermore, the SLA research that has investigated young learners has painted an unclear picture regarding the degree to which the benefits of spacing extend to young learners (e.g., Rogers & Cheung, 2018, 2020; Serrano & Huang, 2018; Kasprowicz et al., 2019). Such an effort would be in line with recent interest in the generalizability of results in the field of SLA (e.g., Porte & McManus, 2018), as well as in the psychological sciences (e.g., Heinrich et al., 2010), and would provide evidence of the generalizability of Bird's (2010) findings.

A second avenue for replication would be to maintain fidelity with the original study as far as possible, but to have a different linguistic focus. With regard to L2 grammar, it would be interesting to examine two or more aspects of grammar at different levels of complexity. Some meta-analyses in the field of cognitive psychology have suggested the complexity of the target stimuli moderates the effects of spacing (Donovan & Radosevich, 1999), with more simple stimuli benefiting more from the effects of spacing in relation to more complex material. Moreover, recent theoretical models have suggested that linguistic complexity may interact with spacing interventions to influence learning (Suzuki et al., 2019; see also Rogers & Leow, 2020; Suzuki, Nakata, & DeKeyser, 2020). Accordingly, it would be worth examining two aspects of grammar at different levels of complexity as a means of testing such models.

Another avenue for replication would be with regard to the learning tasks utilized as part of the training stage of the study. As noted previously, Bird's (2010) study involved form-focused tasks: specifically, worksheets that required participants to make grammaticality judgements of

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

decontextualized sentences and to correct the incorrect utterances. A replication study could explore multiple possibilities. The first is that a more contextualized and meaning-focus task might be used in the training phase of the study. Such an approach would be more in line with current approaches to the learning and teaching of language, such as task-based approaches which emphasize the positive benefits of tasks in that they provide more authentic practice opportunities and are more likely to facilitate incidental acquisition than traditional form-focused exercises (e.g., Ellis, Skehan, Li, Shintani, & Lambert, 2020). However, given that form-focused instruction is still utilized in many (if not most) learning contexts, it would also be worth exploring alternate learning tasks, such as ecologically valid tasks that are commonly used within the context in which the research is set (Rogers & Cheung, 2018, 2020).

One limitation of Bird's (2010) study that a replication might address would be to examine a wider range of ISIs and RIs. This limitation is not unique to Bird's (2010) study, but extends across the field of SLA, in which studies of input spacing have examined a relatively limited range of ISIs and RIs. Typically, these studies have examined only two different ISIs, usually one relatively short ISI and one slightly longer one. With regard to RIs, these studies have typically included only two posttests, with the initial posttest occurring either immediately following training (e.g., Miles, 2014; Rogers, 2015) or shortly thereafter (e.g., Bird, 2010), and the delayed posttest occurring following a delay of at least a few weeks. Typically, the ISI/RI ratios are manipulated, at least partly, in such a way that some of the ISI/RIs ratios fall within the optimum range suggested by Cepeda et al. (2008), and some outside of this range. The issue is that the studies to date have each investigated a limited number of ISI/RI ratios, and this piecemeal approach has not provided consistent evidence in support of or against the ratios proposed by Cepeda et al. (2008). It would benefit the field if studies were to investigate a wider range of ISIs (for example, three or more), as well as multiple RIs. It also seems important to vary the ranges sufficiently to ensure that the different spacing schedules result in differences in learning. For example, while it is likely that comparing a true massed condition (that is, a zero-day ISI) to a four-day ISI would lead to differences in learning, the degree to which the comparison of a three-day ISI to a seven-day ISI would lead to the same results is unclear.

With regard to investigating a range of ISIs and RIs, an important methodological limitation of SLA research concerns research design. As retrieval mechanisms have been implicated as an explanatory mechanism for spacing effects, it is important to recognize that testing sessions represent an opportunity for retrieval to take place, and that retrieval attempts have been shown to influence learning, regardless of their success (see, e.g., Roediger et al., 2011; Karpicke, 2017). Accordingly, research designs in which the subjects take part in multiple posttests can potentially confound the effects of spacing. As an example, Suzuki (2017) discussed this potential confound and, at the reviewers' request, recalculated the ISI/RI ratios in the study, thus muddying the interpretation of the results. Studies in cognitive psychology (e.g., Cepeda et al., 2008) have also manipulated the testing stage between-participants; in other words, each experimental group only took part in testing at a single point in time. A replication of Bird's (2010) study should adopt a similar practice, and manipulate the testing stage to ensure that each experimental group only takes part in a single posttest.

3. Study 2: Serrano (2011)

Serrano (2011) is an influential, oft-cited study investigating the effects of regular versus intensive modes of study on L2 development. This study can be viewed as the culmination of a body of research stretching back to the 1980s that has investigated distribution of learning at the curricular level (e.g., Spada & Lightbown, 1989; Lightbown & Spada, 1994; White & Turner, 2005; Serrano & Muñoz, 2007; Collins & White, 2011). As noted in the introduction, these curricular-level studies have generally shown positive benefits for more intensive, massed conditions. Although it may appear that sufficient positive evidence of the benefits of intensive instruction is available, there are a number of reasons that additional research is needed. First, the extant curricular-level studies cited have been conducted in a limited range of educational contexts, specifically either the Spanish (e.g., Serrano & Muñoz, 2007;

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

Serrano, 2011) or Canadian English as a foreign language (EFL) context (e.g., Lightbown & Spada, 1994; Collins & White, 2011). Accordingly, whether their findings of these studies can be generalized or are artifacts of their particular teaching and learning contexts is unclear. A further, more pressing reason that a replication of these studies is warranted is due to a methodological limitation. As noted in the cognitive psychology literature on input spacing, the effects of distributed practice are most likely to be seen in delayed testing (Cepeda et al., 2006). Studies that do not include delayed posttests are often viewed as methodologically flawed in that they are argued to have 'stacked the deck' in favour of massed conditions, and have created a biased experimental condition. Recent reviews (e.g., Rohrer, 2015) from the field of cognitive psychology have excluded studies such as Serrano's (2011) based on the lack of a delayed posttest. Thus, approximate replications of curricular-level research might address some of these perceived shortcomings.

Serrano (2011) compared the learning gains made during a short, intensive course versus those made in a longer, more extensive language course. The courses in question were language courses offered by the University of Barcelona. The regular classes in this study met twice a week for two-hour sessions (four hours a week in total) over the course of the entire academic year (a total of 110 hours). The students in the intensive classes also received 110 hours of instruction, but this instruction was condensed into a period of four and a half weeks during the summer term. These intensive classes met five days a week for five hours per session (25 hours per week in total). The participants in this study were university-age students, who were enrolled in these courses to earn elective credits. Two levels of language proficiency were analysed in this study, namely intermediate and advanced. In total, 14 different intact classes took part in the study, each of which was taught by a different instructor.

Unlike more experimental distribution of practice research, such as Bird (2010), other curricular-level investigations (e.g., Collins & White, 2011) have been more observational in nature in that they did not prescribe a particular set of training materials or interventions, but allowed the instruction to be carried out as usual throughout the study instead. Such an approach increases the ecological validity of the study, but can also result in a decrease in internal validity due to potential confounding variables, such as differences in methodological approaches by different teachers in the study. With regard to instruction, Serrano (2011) noted that the courses followed a somewhat traditional curriculum, which included communicative activities in conjunction with grammatically oriented tasks. Each class that took part in the study was taught by a different instructor, but Serrano (2011) included a wide number of different sections as a means of offsetting any instructor effects. The students were not assigned randomly to either the regular or to the intensive condition; that is, they choose to register for their respective courses, though the author noted that the population could be considered homogenous with regard to ability and motivation.

Serrano (2011) adopted a pretest/posttest design. Data were collected at each time point using three different types of tasks, which were a proficiency test, a written task, and an oral narration task. The proficiency test consisted of a battery of tasks, namely a listening task, a reading task, a sentence conversion task, and a cloze task. The writing task consisted of a composition on the topics of 'My best friend' (at pretesting) and 'Someone I admire' (at posttesting). These compositions were then analysed based on a number of measures of writing quality, such as accuracy, fluency, lexical complexity, and syntactic complexity. The oral narration task required students to tell a narrative based on a series of pictures about two children going on a picnic. These samples were then analysed in terms of fluency, accuracy, and complexity.

The pretest was administered close to the beginning of the course (after approximately 20 hours of classroom instruction), and the posttest was administered towards the end of the course (approximately 80 hours after the pretest). Data were analysed by comparing learning gains using a series of *t*-tests to compare the average gains of students in the regular classes versus those in the intensive classes across the individual outcome measures. The results of this study indicated that, at the intermediate level, students in the intensive courses showed significantly higher gains than did those in the regular courses based on a number of measures of L2 development. These gains were most evident in different aspects of the proficiency test, particularly in the listening, sentence conversion, and reading tasks that measured vocabulary knowledge, which showed medium-sized effects. No significant

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

differences were seen in the cloze exercises or in reading questions measuring comprehension. With regard to the writing task, a number of measures of writing quality were analysed, but significant differences were only seen for one aspect, which was lexical complexity. No other differences were seen in terms of gains in written performance, and no differences were noted across any of the measures of spoken performance. For the advanced learners, no significant differences were seen between the regular and intensive groups across any of the measures of general proficiency, or in writing or speaking performance, with small-sized effects throughout. These results indicate that learners with lower levels of ability may benefit more from intensive instruction, whereas advanced learners do not seem to benefit to the same degree.

3.1 Approaches to replication

As noted, the generalizability of Serrano's (2011) study, as well as that of other curricular-level studies, is questionable due to the limited number of contexts in which this body of research has been conducted. Accordingly, a straightforward approach to replication would be to carry out an approximate replication in a different teaching and learning context. Such a replication would assist in avoiding leaps of logic when generalizing research findings from one context to another (Spada, 2015). As Serrano's (2011) study investigated university-aged students studying EFL in Spain, a closely approximate replication might follow Serrano's (2011) approach by comparing learning gains in a university course over the course of a semester, or an entire academic year, versus the learning gains in more intensive programmes, such as courses during the summer months, while keeping the total number of instructional hours constant. Although examining a range of course lengths would be beneficial for our understanding of practice effects in SLA, the length and intensity of the different training conditions will undoubtedly depend on the particular context in which a study is set.

Another possibility for replication that is worth exploring would be to examine different age groups. As discussed with regard to Bird's (2010) study, age has been identified as a factor that may influence the effects of spaced instruction. As the curricular-level studies to date have drawn samples from primary (e.g., Collins & White, 2011) or tertiary settings (e.g., Serrano, 2011), one future direction would be for further approximate replications to examine the effects of spacing across a number of age ranges, such as young learners, and learners in middle school and high school settings. Exploring the effects of spacing on a range of different ages within a single study would allow for more direct evidence towards the degree that the benefits of spacing is moderated by the age of the population sample in the study.

An interesting aspect of Serrano's (2011) study was that it examined the effects of spacing across two different levels of language proficiency, namely intermediate and advanced. The results showed that intermediate students benefited from more intensive language courses based on a number of measures, whereas spacing seemed to have no effect on learning gains for advanced learners. This result is interesting in that it suggests that intensive instruction might be more beneficial at lower levels of learning. This finding can be related to skill-based theoretical accounts of language learning whereby intensive exposure can lead to proceduralization/automatization processes, which are more likely to occur at the intermediate level (see DeKeyser, 2007; Serrano, 2011, for discussions). This finding can also be linked to the broader literature on the benefits of input spacing, in which level of expertise has long been discussed as a potential moderating factor in the benefits of spacing (see Ruch, 1928, for an early discussion; for a more recent discussion, see Carpenter, 2017). Future replication attempts might also investigate the effects of spacing across multiple proficiency levels. Such data would be beneficial in light of the theoretical discussions outlined in Section 2.1 above, and on practical grounds with regard to how curricula might best be structured to promote learning; for example, whether more intensive instruction at an early stage in the learning process followed by more extensive instruction at later stages would be beneficial.

Another important variable that future replications might manipulate is with regard to the RI, or the timing of the posttest. Serrano (2011) included a single immediate posttest. However, the lack of delayed testing is problematic in that the effects of distributed practice are argued to be

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

more apparent following a delay (Cepeda et al., 2006). Accordingly, future research should strive to incorporate testing at multiple time points to obtain a more robust picture of the learning and retention as a result of the different spacing conditions. However, researchers should utilize multiple posttests with caution, due to the fact that testing provides retrieval opportunities, and taking part in an immediate posttest might influence performance on later, delayed posttests. Some recent SLA research (Rogers & Cheung, 2018, 2020) has attempted to avoid this confound by including only a single delayed posttest. However, there are a number of disadvantages to such an approach. Having only one posttest forces researchers to decide on a single RI and, as in Rogers and Cheung's (2018, 2020) studies, this can result in 'stacking the deck': in other words, creating an advantage for a particular spacing condition. Similar to the suggestions outlined for a replication of Bird's (2010) work, it is suggested that future research should adopt multiple posttests at different RIs. Most importantly, testing should be manipulated between participant groups, with each experimental group only taking part in a single posttest at a particular time point. The obvious caveat is that this approach would also necessitate an increase in the overall sample size/number of classes taking part in the study, which may not be practical. However, such an approach is recommended when feasible.

One of the strengths of Serrano's (2011) study upon which future replications should build is the use of multiple measures of learning. Serrano (2011) included a variety of tasks designed to measure general L2 proficiency, as well as measures of spoken and written performance. Given the multifaceted nature of L2 development, future replications should follow suit and utilize a battery of measures, which might include both receptive and productive measures, measures that tap into different types of knowledge (such as explicit versus implicit knowledge), or various measures related to the complexity, accuracy, and fluency (CAF) of linguistic performance (see, e.g., Michel, 2017, for an overview of CAF measures in SLA). Such an approach would not only allow for a more complete picture of the development that has taken place but also might shed light on whether spacing has differential effects on different aspects of L2 development.

A replication of Serrano's (2011) study in a different educational context would also inherently include a number of other differences from the original study, such as differences in the language curriculum, methodological approaches, training materials, students' learning habits, and so on. In this sense, it is vital that any replication also documents the nature of classroom instruction qualitatively. A number of previous curricular-level studies have identified differences in methodological approaches and classroom activities between intensive and extensive language courses, even those that occur in the same context (Collins et al., 1999; Collins & White, 2011). Such documentation is important because it might shed light on the internal validity of the study, specifically the degree to which any observed differences are the result of the spacing of instruction rather than being due to differences in pedagogical approaches. Such data might also provide avenues for follow-up studies examining interactions between spacing and different instructional conditions. In addition to classroom data, it may also be fruitful to survey and interview students regarding their language habits outside of the classroom. This is important given the growing recognition of how students' formal and informal language exposure outside of class influences proficiency development (e.g., Lee & Dressman, 2018; Ortega, 2019). Collecting such data would provide a fuller picture of the nature of learning that has taken place, and would provide greater insights into how spacing of instruction influences learning in authentic contexts.

4. Conclusion

Which findings generalize and which do not? This is arguably a pressing question across the education sciences. It is also a question that replication is best suited to address (Porte & McManus, 2018). This paper has suggested a number of ways in which studies by Bird (2010) and by Serrano (2011) might be replicated and extended. The major question addressed by these studies is the degree to which longer gaps between training influence the learning and retention of different aspects of an L2. Given the

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

growing interest in distribution of practice effects in SLA (see, e.g., Suzuki et al., 2019), and the conflicting findings to date, it appears that replications of these studies are warranted. Such replications would allow for greater clarity regarding the reliability and validity of these studies, and the generalizability of their findings; furthermore, they could increase our understanding of the degree to which the distribution of training influences L2 learning, and its applicability to authentic teaching and learning contexts.

References

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

Carpenter, S. K. (2017). Spacing effects on learning and memory. In J. T. Wixted & J. H. Byrne (Eds.), *Cognitive psychology of memory: Vol. 2. Learning and memory: A comprehensive reference* (2nd ed., pp. 465–485). Oxford, UK: Academic Press.

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*(3), 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.

Collins, L., & White, J. (2011). An intensive look at intensity and language learning. *TESOL Quarterly*, *45*(1), 106–133.

Collins, L., Halter, R. H., Lightbown, P. M., & Spada, N. (1999). Time and the distribution of time in L2 instruction. *TESOL Quarterly, 33*, 655–680.

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge, UK: Cambridge University Press.

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). New York, NY: Routledge.

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*(5), 795–805.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2020). *Task-based language teaching: Theory and practice*. Cambridge, UK: Cambridge University Press.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, *26*(2), 275–301.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning & Verbal Behavior*, *15*(1), 1–16.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83.

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), *Learning and memory: A comprehensive reference, Vol. 2: Cognitive psychology of memory* (pp. 487–514). Oxford, UK: Academic.

Kasprowicz, R., & 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*(6), 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 classroom. *Modern Language Journal*, *103*(3), 580–606.

Lapkin, S., Hart, D., & Harley, B. (1998). Case study of compact core French models: Attitudes and achievement. In S. Lapkin (Ed.), *French second language education in Canada: Empirical studies* (pp. 3–31). Toronto, Canada: University of Toronto Press.

Lee, J. S., & Dressman, M. (2018). When IDLE hands make an English workshop: Informal digital learning of English and language proficiency. *TESOL Quarterly*, *52*(2), 435–445.

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). Bristol, UK: Multilingual Matters.

Lightbown, P. M., & Spada, N. (1994). An innovative program for primary ESL students in Quebec. *TESOL Quarterly*, *28*(3), 563–579.

Lightbown, P. M., & Spada, N. (2019). Teaching and learning L2 in the classroom: It's about time. *Language Teaching*, 1–11. doi.org/10.1017/S0261444819000454

Michel, M. C. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen & S. Masatoshi (Eds.), *Routledge handbook of instructed second language acquisition* (pp. 50–68). New York, NY: Routledge.

Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System*, *42*, 412–428.

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, *35*(4), 463–482.

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*(4), 677–711.

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press

Nakata, T., & Elgort, I. (2020). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*. doi: 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*(2), 287–311.
Nakata, T., Tada, S., Mclean, S., & Kim, Y. A. (2020). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly*. Advance online publication, 1–23.
Ortega, L. (2019). SLA and the study of equitable multilingualism. *Modern Language Journal*, *103*(S1), 23–38.
Porte, G., & McManus, K. (2018). *Doing replication research in applied linguistics*. New York, NY: Routledge.
Roediger, H. L. III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford, UK: Elsevier.
Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*(4), 857–866.
Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, *38*(6), 906–911.
Rogers, J., & Cheung, A. (2018). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*, *24*(5), doi: 10.1017/S0272263120000236.
Rogers, J., & Cheung, A. (2020). Does it matter when you review? Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition*. doi.org/10.1017/S0272263120000236
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). New York, NY: Routledge.
Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review, 27*, 635–643.
Ruch, T. C. (1928). Factors influencing the relative economy of massed and distributed practice in learning. *Psychological Review*, *35*(1), 19–45.
Serrano, R. (2011). The time factor in EFL classroom practice. *Language Learning*, *61*(1), 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). Bristol, UK: 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*(4), 971–994.
Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System*, *35*, 305–321.
Spada, N. (2005). Conditions and challenges in developing school-based SLA research programs. *Modern Language Journal*, *89*(3), 328–338.
Spada, N. (2015). SLA research and L2 pedagogy: Misapplications and questions of relevance. *Language Teaching*, *48*(1), 69–81.
Spada, N., & Lightbown, P. M. (1989). Intensive ESL programmes in Quebec primary schools. *TESL Canada Journal*, *7*, 11–32.
Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning*, *67*(3), 512–545.
Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, *21*(2), 166–188.
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*(1), 313–319.
Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. *The Psychology of Learning and Motivation*, *60*, 113–189. Elsevier Academic Press.
White, J., & Turner, C. (2005). Comparing children's oral ability in two ESL programs. *Canadian Modern Language Review*, *61*(4), 491–517.

John Rogers is an assistant professor in the Department of English Language Education, The Education University of Hong Kong, where he teaches courses related to second language acquisition and pedagogy. His research interests are broadly situated within the field of SLA, with a specific focus on research methodology, as well as the role and efficacy of practice in the teaching and learning of second languages.

Cite this article: Rogers, J. (2021). Input spacing in second language classroom settings: Replications of Bird (2010) and Serrano (2011). *Language Teaching*, *54*(3), 424–433. https://doi.org/10.1017/S0261444820000439

https://doi.org/10.1017/S0261444820000439 Published online by Cambridge University Press