Applied Psycholinguistics (2022), **43**, 513–550
doi:10.1017/S0142716421000631

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

Lag effects in grammar learning: A desirable difficulties perspective

Jonathan Serfaty* and Raquel Serrano

Department of Modern Languages and English Studies, University of Barcelona, Barcelona, Spain
*Corresponding author. Email: jonny.serfaty@gmail.com

(Received 16 June 2021; revised 1 October 2021; accepted 12 December 2021; first published online 03 February 2022)

Abstract

This paper examined lag effects in the learning of second language (L2) grammar. Moreover, following the Desirable Difficulty Framework for L2 practice, the present study investigated whether lag effects could be explained by other sources of difficulty. Using digital flashcards, 117 English language learners (aged 10–18) learned two grammatical structures over two different sessions at a 1-day or 7-day intersession interval (ISI). Learners' performance was analyzed at two retention intervals (RIs) of 7 and 28 days, respectively. Linguistic difficulty was compared by examining two different structures, while learner-related difficulty was analyzed by comparing learners who differed in terms of age, proficiency, and time required to complete the training. Results showed no main effect of ISI, a main effect of RI, and a small but significant ISI × RI interaction. Linguistic difficulty and age did not interact with ISI or RI. However, longer lags led to significantly higher scores for faster learners and learners of higher proficiency, while shorter lags promoted significantly higher scores for slower learners and learners of lower proficiency. The findings provide some support for the Desirable Difficulty Framework in its potential to explain L2 lag effects.

Keywords: Lag effects; desirable difficulty; retrieval effort; digital flashcards; grammar

The effect of input spacing on learning has attracted the attention of cognitive psychology researchers for over a century, but it is only in the past decade that this line of research has become prominent in the field of second language acquisition. Many publications have shown that time distribution has an impact on second language (L2) learning outcomes, but it is still not clear what the optimal distribution of L2 grammar practice should be.

Research on input spacing has mainly focused on two phenomena. Firstly, the spacing effect, which refers to the idea that time delays between repetitions of stimuli build memory better than massing them, given the same amount of exposure (Cepeda et al., 2006). The effect has been demonstrated in L2 learning, mostly using vocabulary (e.g., Bahrick & Hall, 2005; Bloom & Shuell, 1981; Koval, 2019; Nakata,

2015; Pavlik & Anderson, 2005) but also in the learning of grammar (Miles, 2014). The second phenomenon concerns lag effects, which refers to the differential outcomes of shorter versus longer intersession lags. These effects have been shown in vocabulary learning on the scale of delays within a single session, over several days, and even weeks, though it has not been found as consistently as the spacing effect (Toppino & Gerbier, 2014). The few existing studies in lag effects for second language (L2) grammar learning have produced evidence in favor of longer lags (Bird, 2010; Rogers, 2015), shorter lags (Suzuki, 2017; Suzuki & DeKeyser, 2017a) or little difference between conditions (Kasprowicz et al., 2019). These studies have used different types of treatments, participants, and target knowledge, which makes it difficult to generalize their findings or offer specific pedagogical recommendations.

A possible explanation for these conflicting findings can be found in the Desirable Difficulty Framework, hereafter DDF (Bjork, 1994, 1999, 2018; Schmidt & Bjork, 1992). The basic tenet of the framework is that adding complexity can decrease performance levels during training, but leads to better retention of the attained knowledge. One possible way to add difficulty is to expand the time delay, or lag, between learning episodes. More recently, Suzuki et al. (2019) have applied this framework to L2 practice. Drawing on the multicomponential nature of L2 difficulty proposed by Housen and Simoens (2016), the framework identifies three sources of difficulty for L2 practice that may influence outcomes, namely linguistic difficulty, learner-related difficulty, and the practice condition. According to the proposed framework, the optimal difficulty of training should depend on all three sources. Therefore, the differential results of lag effects for grammar learning reported in the past might be explained by the effects of other sources of difficulty.

The present paper aims to assess whether the DDF for L2 practice proposed by Suzuki and colleagues can account for differential lag effects in L2 grammar learning by explicitly testing lags under different levels of linguistic and learner-related difficulty. Although the DDF can be used in order to explain and compare the results of previous studies retrospectively, to the best of the authors' knowledge, the present study constitutes the first direct attempt to use this framework to account for lag effects in L2 practice. It is hoped that the findings contribute to the theoretical discussion of lag effects in SLA and the feasibility of the DDF as an avenue for determining best practice in L2 learning.

Literature review

Lag effects in cognitive psychology

The cognitive psychology literature has examined the effects of intersession interval (ISI), defined as the delay between study sessions, and retention interval (RI), the time from the final study session to the posttest, on learning and retention. Throughout this study, ISIs and RIs will be measured in days (e.g., ISI-1 is an intersession interval of one day) unless specified otherwise.

Cepeda et al. (2006)'s meta-analysis found that longer ISIs were better for longer RIs, though most studies were on the scale of hours. Expanding this idea to a longer scale, Cepeda et al. (2009) used six ISIs from 5 minutes to 14 days, tested at an RI of 10 days (RI-10), for the retention of Swahili–English word pairs. Scores were significantly higher for ISI-1 (10% of RI) than for ISI-0, with a 34% difference in scores. No other pairwise comparison reached statistical significance, with gradually

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

decreasing scores as ISI increased. That is, the lag effect was nonmonotonic, and a longer lag after a certain optimal point was actually somewhat detrimental to retention. Cepeda et al. (2009) reported a second experiment in which participants learned the names of obscure objects with ISIs from 5 min to 6 months, assessed after 6 months. Here, the 1-month ISI (17%) fared best. This pattern has been found in studies up to an RI of 350 days (Cepeda et al., 2008), namely that the optimal ISI is approximately 10–20% of the RI (Rohrer & Pashler, 2007).

Thus, findings from cognitive psychology have suggested that the optimal ISI is largely dependent on its ratio with the RI. However, when applying this to L2 grammar practice, the situation becomes less clear. Bird (2010) and Rogers (2015) produced evidence supporting a longer lag for better retention, whereas Suzuki and DeKeyser (2017a) and Suzuki (2017) found advantages for a shorter lag, regardless of RI. Finally, Kasprowicz et al. (2019) found no clear advantage to either lag. This body of research suggests that lag effects may differ according to various criteria.

Lag effects have previously been associated with the DDF (Bjork, 1994, 1999, 2018; Schmidt & Bjork, 1992) based on study-phase retrieval theories. Pyc and Rawson (2009) demonstrated that retrieval of previous presentations becomes more difficult with longer lags and that when successful retrievals are more effortful than easier retrievals, knowledge is more durable. Thus, an optimal lag would induce the highest retrieval effort while still facilitating successful retrieval. Too short a lag would induce suboptimal effort, and too long a lag would lead to unsuccessful retrieval.

However, retrieval effort may also depend on other factors. The DDF for L2 Practice (Suzuki et al., 2019) cites three main sources of difficulty: linguistic difficulty, learner-related difficulty, and the practice condition. The following section will discuss previous findings for lag effects on L2 grammar practice by first considering practice conditions and then exploring how lag effects might depend on linguistic and learner-related sources of difficulty.

Lag effects according to Suzuki et al. *(2019)'s DDF for optimal L2 practice*

Practice condition

Practice, defined here as activities engaged in for the intentional development of L2 knowledge and skills (DeKeyser, 2007), may be performed under more or less difficult conditions, regardless of what is being learned or who is learning it. This could include blocked or interleaved presentations, recognition or recall training, deductive or inductive rule learning, explicit or implicit feedback, among many others. In the case of lag effects, a longer lag would create a more difficult practice condition by requiring more effort in retrieving previously attained knowledge (Pyc & Rawson, 2009).

Bird (2010) was the first to compare lag effects for L2 grammar learning. During four sessions of ISI-3.3 or ISI-14, 38 Malaysian English language learners (ELLs) studied two pairs of grammatical structures, counterbalanced with ISI within participants. Both treatment and assessment were grammaticality judgement tests (GJTs). Both ISIs led to significant gains at RI-7, but at RI-60, the longer ISI-14 led to significantly better retention than ISI-3.3. Notably, ISI-14 with RI-60 was the only combination that approximated Rohrer and Pashler's (2007) optimal ratio at 23%.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Further support for longer lags in grammar learning was found by Rogers (2015), who examined the effects of implicit learning of complex grammatical structures among 37 ELLs in Qatar. During five sessions of either ISI-2.5 or ISI-7, subjects saw sentences that used the target structure and answered yes/no comprehension questions about their meaning. GJTs were administered immediately and at RI-42, which was within the optimal ratio for the longer-lag group (17% vs 5%). As with Bird (2010), groups made similar initial gains but at RI-42 only the longer lag group maintained their gains.

Different results were obtained by Suzuki and DeKeyser (2017a) and Suzuki (2017) with grammar tasks that involved oral production. In the former, Suzuki and DeKeyser (2017a) taught the Japanese present continuous structure to undergraduate beginners in two 50-min sessions at either ISI-1 or ISI-7. The lessons included vocabulary learning, grammar explanations, comprehension practice, and oral production practice. Participants were given a rule application test and a sentence completion test. For accuracy, no statistical differences between ISI groups were found, though there was a marginally significant advantage to the shorter ISI-1 for reaction times at RI-28. This seemed to contradict earlier findings from Bird (2010) and Rogers (2015). Suzuki (2017) then conducted a conceptual replication of the study using an artificial language, with more stringent controls. This time it was the accuracy scores that gave a significant advantage to the shorter ISI for all tests.

Lastly, a grammar study was conducted by Kasprowicz et al. (2019) using multiple-choice computer games to teach French morphology in a primary school setting. Participants studied in either three sessions of 60 min at ISI-7 or six sessions of 30 min at ISI-3.5. In both conditions, high accuracy rates (>75%) were recorded during training and posttest scores were low, with only a marginal advantage to the ISI-3.5 group because they had started with lower pre-test scores.

In line with the DDF, the different results in terms of lag effects reported in the literature could be explained by other aspects of the practice condition, for example the types of tasks used during training and/or testing, which might have induced differing levels of difficulties. For example, Bird (2010) and Rogers (2015) used GJTs, which can only indicate a learner's ability to recognize specific L2 structures, rather than produce them. Studies involving both recall and recognition have consistently reported substantially higher scores for recognition (e.g., Bahrick & Phelps, 1987). Regarding the treatment for Rogers (2015), grammar learning was incidental, measured after exposure to forms in a task that was not language focused. Consequently, these studies likely induced relatively low levels of retrieval effort. In line with the predictions of the DDF, a longer lag was beneficial in these cases, as it added desirable difficulty to the practice condition. On the other hand, Suzuki and DeKeyser (2017a) and Suzuki (2017) included productive recall activities. In these studies, retrieval effort was high, with training that involved the retrieval and manipulation of newly learned linguistic forms both productively and receptively in timed oral tasks. As might be expected, the shorter lag was best, as the task was itself already difficult. Finally, as Suzuki et al. (2019) suggest, the lack of differences reported by Kasprowicz et al. (2019) can be interpreted as neither lag being sufficient to induce enough desirable difficulty to improve scores.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Linguistic difficulty

Linguistic difficulty refers to relative difficulties of target features such as saliency, allomorphy, and complexity (Housen & Simoens, 2016). In the case of vocabulary learning, Bahrick and Phelps (1987) found that items were better retained 8 years after learning with an ISI-30 schedule than with ISI-1 or massed learning. They also analyzed results according to per item difficulty. The number of presentations required to learn each word for each subject was recorded, and it was found that the easier items were better remembered 8 years later, regardless of ISI. These findings exhibited an advantage to a more difficult practice condition, but a disadvantage to higher linguistic difficulty.

Prior research into the interaction of lag effects and linguistic difficulty for L2 grammar learning has only compared difficulty on the scale of a single word. Suzuki (2017) compared words requiring one or two morphological changes and found no interaction with lag, though a facilitatory effect of the shorter lag during training was stronger for more complex target forms, involving more changes. This suggests that the shorter lag may aid in more difficult target knowledge, and that this effect may be amplified when form complexity is increased. In sum, there is a dearth of evidence regarding the interaction between lag effects and target forms, and Suzuki et al. (2019) called for more experiments examining lag effects using structures of differing degrees of linguistic difficulty. The present study aims to contribute to this line of research.

Learner-related difficulty

Learner-related difficulty comprises prior knowledge, affective factors, and cognitive abilities. This source is more difficult to measure, due to the subjective nature of learners' experiences. However, it is possible to infer difficulty from learner attributes. For example, Suzuki and DeKeyser (2017b) and Suzuki (2019) found that some aptitude measures (language analytic ability and metalinguistic rule rehearsal ability) predicted learning but only for their long-lag condition (ISI-7). On the other hand, Kasprowicz et al. (2019) found language analytic ability to be a significant predictor of scores for young learners regardless of ISI.

Another potential source of learner-related difficulty could be the learner's general L2 proficiency. Learners of higher L2 proficiency can be expected to experience less difficulty in learning a new L2 form than those with lower proficiency. Previous findings might also be explained by this learner-related difficulty, which might have led to shorter lags being more beneficial for learners with lower L2 proficiency (e.g., the beginner-level learners in Suzuki & DeKeyser, 2017a and Suzuki, 2017) and longer lags for higher proficiency levels (e.g., the intermediate learners in Bird, 2010 and Rogers, 2015).

A third cause of learner difficulty may be age. In a classroom setting, adolescents over the age of 12 tend to learn foreign languages faster than children (Muñoz, 2006, 2007, 2008). This has been attributed to superior cognitive abilities, including organization, selective attention, decision making, and working memory, due to neurobiological processes such as myelination (Bathelt et al., 2018; Yurgelun-Todd et al., 2002) that begin at adolescence. Lower scores overall may therefore be expected from children in cognitively demanding tasks. Regarding lag effects, children's lower

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

short-term memory capacity (Fandakova et al., 2014) would lead to more forgetting between sessions after longer lags. This would consequently lead to fewer successful retrievals at the beginning of a new session, meaning that more successful retrievals will come later in the session where the delay since feedback is only a few minutes, rather than days. Vaughn et al. (2016) conducted a study where participants learned items to criterion, meaning that items were dropped from the cycle after being answered correctly but were otherwise repeated in subsequent rounds. They found that successful retrievals on the first round were more effortful, based on first key-press latencies, and that the conditions that led to more effortful successful retrievals also produced more durable knowledge. Accordingly, if children experience fewer effortful successful retrievals as a result of forgetting between sessions, they may benefit less from the added difficulty of a longer lag as compared to older learners.

To the best of the authors' knowledge, no studies have directly compared lag effects in L2 learning among children and adolescents. However, studies of lag effects for L2 learning in children support the notion that longer lags might not be beneficial for this age group. In a study of learning French morphology through computer games, Kasprowicz et al. (2019) found minimal differences among learners aged 8–11 between ISI-3.3 and ISI-7, with a small advantage to the ISI-3.3 group. Similarly, research on vocabulary learning in primary school children has shown either no differences between shorter and longer lags or an advantage for the shorter lag, or less effortful condition (Goossens et al., 2016; Rogers & Cheung, 2020a, 2020b).

As a comparison, Küpper-Tetzel et al. (2014) found stronger lag effects among older children (aged 11–13). Küpper-Tetzel and colleagues taught English–German vocabulary pairs to students in an authentic classroom with ISIs of 0, 1, or 10 days. At RI-7, ISI-1 outperformed the other two conditions, whereas at RI-35 both the 1-day and 10-day ISI groups outperformed the massed group, with ISI-1 still best. It was concluded that the optimal ISI increases with RI, noting the importance of using multiple RIs in lag experiments. Their particular optimal ISI for RI-35 was shorter than for Cepeda et al. (2008)'s lab study with adults, where scores increased from 0 to 11 day ISIs. The discrepancy was explained by the differential working memory and forgetting rates of adults and children.

Of course, age-related cognitive differences are not the only factor that separates classroom studies with school-aged learners from lab studies like Cepeda et al. (2008). Firstly, an experiment in an authentic classroom setting with younger learners will undoubtedly involve countless extraneous variables and less control. This would make it difficult to isolate time distribution as a factor. Moreover, lab studies of undergraduate students are undertaken voluntarily by participants of a certain level of education, and probably a certain willingness to perform the study appropriately. Children in a classroom may have little interest in following instructions, or become easily distracted, and often have less choice as to their participation. Nevertheless, the small advantage to the shorter ISI in school classroom studies has been fairly consistent (Goossens et al., 2016; Kasprowicz et al., 2019; Küpper-Tetzel et al., 2014; Rogers & Cheung, 2020a; Serrano & Huang, 2018, 2021). Therefore, although the classroom context involves many variables, shorter lags seem to be preferable for this age group.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Digital flashcards

The tool of learning in the present paper was the digital flashcard app Quizlet. This app is typically used for paired-associate learning, whereby the target L2 item may be paired with its L1 translation or a definition, and learners can study target items selected by their teacher independently as well as create their own sets. Numerous studies have shown the use of flashcard apps to be an effective and motivating tool for enhancing vocabulary learning (Kornell & Bjork, 2008; Nakata, 2020; Wissman et al., 2012). Recently, Serfaty and Serrano (2020) also showed that flashcards can be successfully used for grammar learning by using whole sentences as items.

Quizlet in particular has been widely used in L2 classroom research (e.g., Andarab, 2017; Ashcroft et al., 2018; Dizon, 2016). As a research tool, it does not provide detailed data such as participants' actual responses on incorrect attempts or their response times, which other research platforms can provide (e.g., Gorilla, DMDX). However, it does bring a number of advantages. For example, L2 learners are generally already familiar with the tool and are motivated to use it (Franciosi et al., 2016; Korlu & Mede, 2018; Sanosi, 2018). It is also one of the top 10 most visited educational websites worldwide (Similarweb, 2021) with 60 million monthly users (Quizlet, 2021), bringing ecological validity to empirical research. Additionally, Quizlet is free to use, which allows for administration to large groups in a variety of settings (including low-resource settings). Finally, L2 learning through Quizlet can be considered more experimentally controlled than the average classroom study, since learning takes place individually while controlling for variables such as feedback style, instructor factors, and number of correct retrievals per participant.

Present study

Suzuki et al.'s DDF (2019) seems to plausibly account for the different results obtained in some of the L2 lag-effect studies presented in the previous section, but to the best of the authors' knowledge no previous studies have used the framework to examine how lag effects are related to other sources of difficulty in determining "optimal" practice conditions. The primary aim of this study is to investigate whether different sources of difficulty are related to lag effects in grammar learning.

The present study used Quizlet in the productive recall mode to manipulate grammar learning under a shorter and longer lag by comparing results at two RIs under different conditions of linguistic and learner-related difficulty. Two different grammatical structures were used to examine linguistic difficulty. For learner-related difficulty, three measures were used. Firstly, age differences were compared by including both children and adolescents. Secondly, general English proficiency was used to approximate prior L2 knowledge. Finally, time on task was used to measure the difficulty experienced by individual learners during training. More details can be found about these measures in the Methodology section.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Research questions and hypotheses

The following research questions (RQs) guided the present study. Each one may be broken down into subquestions, as follows:

RQ1: Are lag effects found in grammar learning with digital flashcards?

a. Is there an advantage to training at either a shorter (ISI-1) or longer (ISI-7) lag?
b. Are scores different at RI-7 and RI-28?
c. Is there an interaction between ISI and RI?

RQ2: Do lag effects depend on other sources of difficulty?

a. Does ISI interact with linguistic difficulty?
b. Does ISI interact with learner-related difficulty factors such as age, proficiency, and time on task?

Considering the results of previous studies involving difficult tasks that required productive recall (Suzuki, 2017; Suzuki & DeKeyser, 2017a), our hypothesis for RQ1 is that the shorter lag, ISI-1, will lead to better scores at both RIs, with overall lower scores at RI-28. Regarding RQ2, in line with Suzuki et al. (2019), it is hypothesized that the benefits of ISI-1 (easier practice condition) will be stronger for the more difficult linguistic structure and for learners experiencing more difficulty during training (children, lower proficiency, and learners that require more time to complete the training), while ISI-7 scores may be higher for the simpler structure, and for learners experiencing less difficulty during training.

Methodology

Participants

Participants were students in a Cambodian international school who study an English-language curriculum in addition to their local curriculum. Initially, all students in the secondary school, grades 6–11, were recruited for the study ($n = 230$), but due to sporadic school closures, absences during data collection points, or not following instructions, only around half ($n = 129$) could be considered for analysis. A further 12 participants who showed previous knowledge of the target grammar forms on a pretest were also excluded from analysis. The final sample comprised 117 participants, aged 10–18 ($M = 13$, $SD = 1.87$), including 63 females and 54 males. The school in which this experiment took place does not necessarily assign grade level by age, which is why some 10 year olds are included in this secondary school study.

Difficulty sources

This study manipulated several conditions of difficulty. A summary of variables can be found in Table 1.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Table 1. Summary of difficulty variables

Type	Measure	Lower difficulty	Higher difficulty
Practice condition	Intersession interval	ISI-1	ISI-7
Practice condition	Retention interval	RI-7	RI-28
Linguistic	Number of transformations and L1 similarity	A	B
Learner	Age	Adolescents	Children
Learner	Proficiency	Low Medium	High
Learner	Time on task	Faster	Slower

Practice conditions

Practice conditions were manipulated in terms of lags and RIs. The two lags chosen for comparison were ISI-1 and ISI-7, to be assessed at either RI-7 or RI-28. Two RIs were used due to evidence from prior research that the optimal ISI depends on the RI (Cepeda et al. 2006, 2009). The shorter ISI is assumed to be easier, considering the evidence that longer lags lead to more forgetting between sessions (e.g., Li & DeKeyser, 2019; Suzuki, 2017), and the shorter RI is assumed to be easier because declarative knowledge is prone to decay after acquisition (Ullman & Lovelet, 2018). These intervals were chosen to allow comparison between this study and previous studies, as well as for practical purposes regarding data collection. Two sessions were used per structure because a similar study using digital flashcards (Serfaty & Serrano, 2020) reported a ceiling effect for a third of participants after three sessions.

Linguistic difficulty

Linguistic difficulty refers to any difficulty regarding the target form, which could include intrinsic complexity, differences from the L1, or task-specific difficulty such as the medium of input, frequency, and salience (Housen & Simoens, 2016; Spada & Tomita, 2010). Although both target structures were designed to be highly difficult, in order to make the task meaningful for students with high proficiency and to avoid previous knowledge, Structure B was intended as more difficult than Structure A in order to test the hypothesis that linguistic difficulty interacts with lag effects.

Structure A was the future perfect progressive (e.g., *I will have been studying for 3 hours by the time I see you*). Structure B was the past perfect conditional in the interrogative form (e.g., *What would you have done if you had found the money?*). Eight sentences per category were created for the pretest and training, and a further eight sentences each were created for the posttest. See Appendix A for all items.

The determinants of linguistic difficulty examined in the present study include some of the factors that have been considered in previous research, namely the

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Table 2. Transformations required for each target structure and differences with respect to L1

	Structure A	Structure B
Cue	I will start studying at 3pm. I will see you at 6pm. (I will continue to study)	You didn't find the money, so you did nothing. But imagine a different past. Hmmm
Target	I will have been studying for 3 hours by the time I see you.	What would you have done if you had found the money?
Transformations Cue → Target		Declarative → interrogative Clause 1 ⇄ Clause 2
	Clause 1: "I will start" + V-ing (+Object/ Complement) + Time → "**I will have been"** + V-ing (+Object/Complement) + ***for*** + Time (**duration**)	**Clause 1** (conditional clause): Subj + V past + Object → **Wh-** + **Aux** + Subj + **V cond. Perfect** • Object → Wh- pronoun (choose between *what, who, where, how*) • Move Wh- to the front • V past → V conditional perfect • Subject + V → Aux Subj V
	Clause 2: "I will" + V + Object + Time Adjunct" → **"by the time I"** + V + Object	**Clause 2** (if clause): "Subject + V past + Object/ Complement" → "**if** + Subject + **V past perfect** + Object/Complement" • Change tense to past perfect • If V in cue is affirmative → negative If V in cue is negative → affirmative
L1 differences		Conditional tense Wh- fronting Interrogative subj-verb inversion

number of transformations required to arrive at the target form, and similarity to L1 features (Spada & Tomita, 2010). Accordingly, Structure B, the more difficult structure, involved more transformations and was less similar to the participants' L1 (see Table 2). For both structures, the participant must combine two sentences into a single sentence with two clauses and conjugate the verbs into complex tenses involving auxiliaries. However, for Structure B the participant must also produce an interrogative sentence from a declarative cue, swap the order of clauses, replace the object with a fronted Wh- word, and change an affirmative clause to a negative clause (or vice versa). In contrast, for Structure A, the conjugation is simplified by using "chunks" that are the same in every example, which means that the participants only need to remember to start each sentence with "I will have been" and then use the same verb in *-ing* form as in the cue. Similarly, the verb after "by the time I" is also in the same form as in the cue. Additionally, the participants' L1, Khmer, does not use an interrogative inversion, Wh- fronting, or express the conditional tense grammatically, whereas Structure A follows a similar syntax to that of the L1. Therefore, Structure B can also be considered more difficult from this perspective.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Linguistic difficulty was confirmed by performance measures during training, which Suzuki et al. (2019) propose as a measure of L2 difficulty (see Results section).

Learner-related difficulty

This study used three separate measures that tap different potential sources of difficulty within the learner, namely age, proficiency, and time on task[1]. In terms of age, in the present study 10–12 year olds were classed as children ($n = 52$) and 13–18 year olds were classed as adolescents ($n = 65$). It was expected that adolescents would experience less difficulty during treatment than children due to more developed cognitive abilities.

Although it is not easy to determine the exact onset of adolescence and it is well known that this varies among individuals, we followed the cut-off that has traditionally been used in the literature analyzing L2 learning in classroom settings (11–12), which roughly corresponds to the age at which different cognitive changes have been claimed to take place (Muñoz, 2007). We decided to choose 12 and not 11, first, in order to have a more balanced number of participants in the two groups, and second, because we observed that, in our sample, the performance of 12 year olds during training was similar to younger participants with a marked drop in time on task for 13 year olds. *T*-tests revealed nonsignificant differences in times on task between 11 and 12 year olds ($p = .498$), and between 13 and 14 year olds ($p = .612$), but a significant difference between 12 and 13 year olds ($p = .017$). Notably, a large majority of the 12 year olds in this study were in the same school grade as the 10 and 11 year olds.

A second measure of difficulty is proficiency level, because prior knowledge is expected to influence learners' ability to acquire target forms (Housen & Simoens, 2016). The participants' English proficiency levels were measured using the Oxford Quick Placement Test (UCLES, 2001), though 14 participants did not complete this test and were not included in this analysis. Since a large majority of participants achieved level B1, and levels A1 and C1 were represented by only three participants each, three new levels of proficiency were created for analysis: low, medium, and high. Low comprises A1 and A2 ($n = 31$), medium is equivalent to B1 ($n = 45$), and high denotes B2 and C1 ($n = 27$).

Lastly, a measure of task-specific learner difficulty was created based on observations during training. Previous research has used the number of trials to reach criterion (e.g., Bahrick & Phelps, 1987), or the first key-press latency (e.g., Pyc & Rawson, 2009) as a measure of difficulty on a per item basis. As grammar items are interrelated, a better measure for difficulty would be the total number of trials required to reach criterion or the accumulated first key-press latencies per session. Unfortunately, these data were not available through Quizlet, but participants did record their time on task. Longer time on task is a reflection of both more trials and more time spent on each trial, which are signs of difficulties experienced by the learners during the treatment. Additionally, time on task matched the first author's first-hand knowledge of students' academic abilities. However, time on task may be influenced by other factors, for example typing speed. Therefore, this measure constitutes only a rough indicator of difficulty and outcomes should be interpreted accordingly. Two groups were created using a *K*-means cluster analysis of

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Table 3. Breakdown of experimental groups by number of participants with ages in parentheses

Posttest RI	Group A	Group B	Total
RI-7	**26** ($M = 13.38$, $SD = 2.00$)	**36** ($M = 12.89$, $SD = 2.01$)	62
RI-28	**26** ($M = 13.19$, $SD = 1.88$)	**29** ($M = 12.90$, $SD = 1.59$)	55
Total	52	65	117

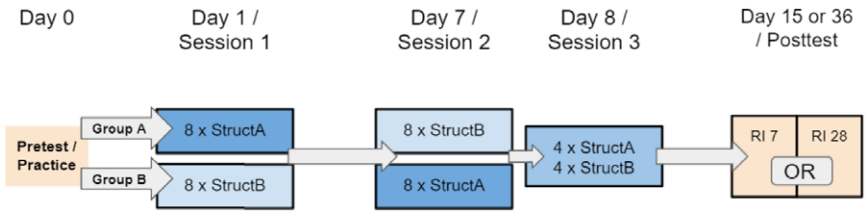


Figure 1. Experimental design.

participants' total time on task over the three training sessions: faster ($n = 60$, $M = 47.15$m, $SD = 14.17$m) and slower ($n = 43$, $M = 106.21$m, $SD = 23.71$m). Participants with missing data ($n = 14$) were not included in this analysis.

The three learner variables were moderately correlated (age*proficiency: $r = .389$, $p = <.001$; proficiency*time: $r = -.655$, $p = <.001$; age*time: $r = -485$, $p = <.001$), which may be interpreted as these variables being related but ultimately measuring different learner attributes.

Experimental design

The experimental design involved a pretest, treatment, and posttest (Figure 1). Students learned two structures at either ISI-1 or ISI-7, counterbalanced within subjects. The treatment consisted of three study sessions (S) in total, each using a single set of flashcards with eight items. S1 used items for ISI-7, S2 used items for ISI-1, and S3 combined them.

Learners were split alphabetically within each grade into two groups (Group A and Group B) that determined which grammatical structures would coincide with which ISI. Following the training phase, participants eligible for analysis were split into two distinct groups to be tested at either RI-7 or RI-28, manipulated for equal representation of the two treatment groups. RI was a between-subjects variable in order to avoid confounds caused by testing effects. By chance, Group B retained more participants. No experimental groups coincided with intact classes. The final breakdown of groups and age distribution can be seen in Table 3.

Independent *t*-test showed no differences in proficiency scores (/60) between treatment Group A ($n = 45$, $M = 34.1$, $SD = 8.6$) and Group B ($n = 57$, $M = 33.4$, $SD = 7.9$), $t[100] = 0.422$, $p = .674$, $d = 0.20$), or between testing groups RI-7 ($n = 56$, $M = 33.9$, $SD = 8.3$) and RI-28 ($n = 46$, $M = 33.5$, $SD = 8.0$), $t[100] = .227$, $p = .821$, $d = 0.08$.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Figure 2. Participants attempt to type the target response.

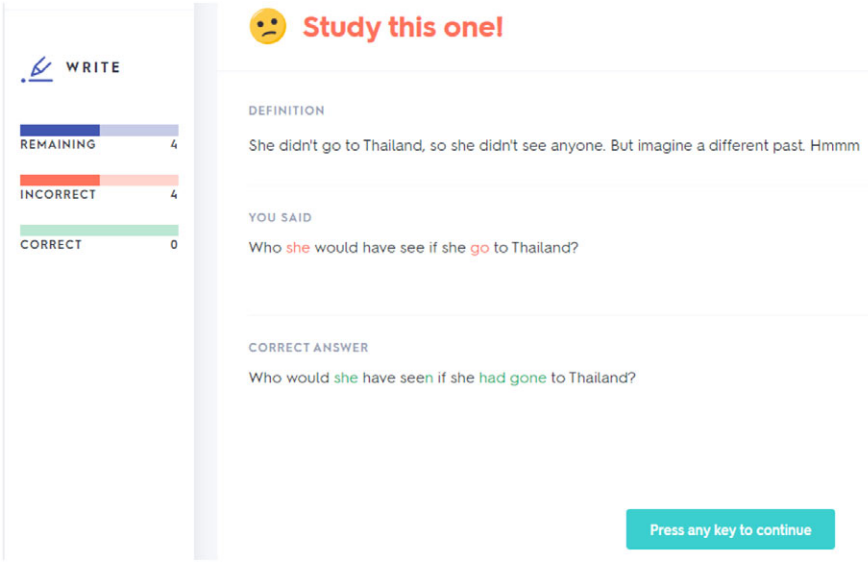


Figure 3. Participants receive feedback on incorrect responses.

Training

Training was performed using the *Write* mode of Quizlet with eight scenario-cues per target structure (16 target sentences in total). There was no instruction stage, but rather participants were presented with the cues and guessed the correct responses. As neither of the target structures can be expressed in isolation in the participants' native language, translations could not be used as cues. Instead, participants read a scenario in English (e.g., *I will start studying at 3pm. I will see you at 6pm.* [*I will continue to study*] for the target of *I will have been studying for 3 hours by the time I see you*). This approach also provided all the vocabulary within the cue, isolating grammar as the target.

After each incorrect response, the target response was presented alongside the participant's response (see Figures 2 & 3). Although it is possible to click "Don't know" and skip to the feedback, participants were strongly encouraged to always guess. Since each item used the same grammatical pattern, participants were expected to infer rules from the feedback as the training progressed

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

(Serfaty & Serrano, 2020). Any correctly typed responses were removed from the set, and training continued in rounds until all items were removed. The order of presentation within each set was randomized.

Tests

Productive cued recall tests were conducted using Google Forms. The pretest consisted of the 16 sentences from the training. Students were asked to write the sentences from the scenario cues. Questions for Structure B also provided the initial question word (see Appendix A).

The posttest comprised eight novel items for each structure, using cues written in an identical style as in the training and pretest (see Appendix A for all test cues). Novel items were used for the posttest to make sure that it was the structure and not the specific exemplars that were learned, following Serfaty and Serrano (2020). No time limits were imposed on tests. Cronbach's alpha showed high internal reliability for posttests: Structure A = .977; Structure B = .942.

Tools

As mentioned, the main tool of learning in this study was Quizlet. In accordance with the school's normal practice, Google Classroom was used to manage the experiment and students used their own devices. Each assignment included a Google Doc with a link to the relevant Quizlet activity and spaces for students to fill in their times as well as add screenshots. The reported times were corroborated with the screenshots, which included their device's clock, and Google Classroom's record of when each file was opened and submitted. The screenshots also served as the record for items answered correctly in the first round of each session (see Appendix B).

Procedure

Before training, two lessons were used for preparation activities, which included an explanation of the experiment, a brief presentation of the target concepts without revealing the target forms in English, a pretest to screen for prior knowledge, and two practice sessions in which participants learned to use Quizlet in the desired manner and record their progress. See Appendix C for a more detailed account of pre-experimental activities.

The pretest was performed on the day before the training during class time. The three training sessions also took place during regular classes. The majority of sessions and tests happened under direct supervision of the first author or their teacher. Desks in classrooms were spaced according to COVID-19 guidelines, which helped to reduce communication between students during training. However, some sessions fell during periods of online learning. It was decided to continue the experiment unsupervised, based on evidence from Rawson, Dunlosky and Sciartelli (2013) that showed similar effects of distributed retrieval practice from supervised and unsupervised learners. In all cases, at least the two practice lessons and S1 were in-person, meaning that students knew what was expected of them. A general

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

baseline of possible performance was established from the 70+ participants that were fully supervised for every session by the first author. For example, times between sessions for the same participant should be similar and the pattern of learning should show a gradual reduction in the number of items with each round. Faster times reliably came from students from whom this was expected, based on their usual academic performance, and vice versa. In certain cases, data clearly did not match the expected pattern of learning and students were asked whether they had followed instructions. In all of these cases ($n = 30$), including one entire class ($n = 24$) who had not understood the goals of the task, students admitted to either not understanding the procedure or to intentionally cheating, and their data were discarded.

The posttest was conducted during regular classes on Google Forms, either 7 or 28 days after the last training session. Some tests (35/117) were completed during online learning, with no implausibly high or low performances. Posttests were not timed and took approximately 15 min to complete. The proficiency test was administered at different times according to student availability and on average it took around 20 minutes.

Analysis

Scoring

A two-point scale was used to score each sentence, one point for each of the two clauses. See Appendix D for examples of responses and criteria for scoring.

Every item was graded three times by the same rater on different days, in a randomized order. Of the 1872 total responses, 22 scoring differences were found and corrected on the second round, with no further differences found on the third round. A second rater marked 17 tests, corresponding to 15% of responses, with 98.5% interrater agreement. The discrepancy was resolved by discussion.

Statistical analyses

The program SPSS 27 (IBM, 2020) was used to perform the statistical analyses. *T*-tests were used to check for significant differences in training performance between groups.[2] Cohen's *d* was used as the effect size statistic, interpreted using the following benchmarks (Plonsky & Oswald, 2014) for independent samples: small ($d = 0.4$), medium ($d = 0.7$), and large ($d = 1.0$), and for paired samples, small ($d = 0.60$), medium ($d = 1.00$), and large ($d = 1.40$).

Generalized linear models for repeated measures with a binomial outcome were used to evaluate the proportion of correct scores in the posttests. This type of model is appropriate for data which does not meet assumptions of a normal distribution or homoscedasticity. Each test item is treated as an observation, and because the total score per item was two, this is equivalent to two binary opportunities for success per item. The lowest Akaike Information Criteria was used to determine the best data structure. Participants and items were the repeated measures, equivalent to random effects in mixed models, meaning that the model accounts for variability between participants and items. All models were built by first adding all possible two-way

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

and three-way interactions, and then removing nonsignificant interactions. In total, five models are reported. Model 1 includes only the key variables of ISI and RI. Each subsequent model includes a single added predictor variable as follows: Model 2 - structure; Model 3 - age; Model 4 - proficiency; Model 5 - time on task. Models 1, 2, and 3 included 1872 observations from all 117 participants. Models 4 and 5 excluded 14 participants, using a total of 1648 observations. A model containing all variables was not used due to the number of variables and possible interactions as well as the correlations between learner-related variables.

A significant *F* statistic for a statistical model indicates that it predicts outcomes better than a model without independent variables. Estimated marginal means with 95% confidence intervals were calculated. These estimated means, which will be labeled as scores for ease of exposition, represent the average proportion of correct responses in a given condition. For example, if ISI-1 scores are $M = 0.5$, this would indicate that a response in the ISI-1 condition has a 50% chance of being correct (in this case, of earning 2 points). The standard error (*SE*) represents the range of likelihood means within the population, so a smaller *SE* indicates better inferential strength to the general population. Odds ratios (*OR*) are used to measure the effect size for this type of analysis. They constitute the added relative likelihood of a correct response in comparison with another level of the predictor. For example, if ISI-1 scores are greater than ISI-7 scores with an *OR* of 1.5, it would indicate that a correct response is 1.5 times more likely, or 50% more likely, under the ISI-1 condition than the ISI-7 condition. As there are no standard guidelines in the field of applied linguistics for interpreting *OR*, we follow the benchmarks used by Kim, Skalicky and Jung (2020). Accordingly, *OR* will be interpreted as small if less than 3, moderate if between 3 and 10, and large if greater than 10. The alpha of *p* was set as .05. Accordingly, a significant effect indicates that the probability of no effect in the general population is less than 5%.

Results

Data files and syntax can be found online.

Training data

Firstly, in order to gain insights into learner-related and linguistic difficulty, time on task and the number of correct responses on the first round for each session were examined. Descriptive statistics are displayed in Table 4.

Paired samples *t*-tests showed that participants spent slightly less time on Session 2 than Session 1, $t[101] = 2.652$, $p = .009$, $d = 0.2$, and substantially less time on Session 3 than Session 2, $t[101] = 7.903$, $p < .001$, $d = 0.78$, where items were repeated from previous sessions. When analyzed by structure, Structure A took 26 min for both groups (Group A S1 & Group B S2), whereas Structure B, the more difficult structure, took 31 min for Group B and 25 min for Group A. Independent samples *t*-tests showed nonsignificant differences among groups for Structure A time ($t[101] = 0.149$, $p = .882$, $d = 0.03$) but time for Structure B was significantly higher for Group B ($t[101] = 2.068$, $p = .041$, $d = 0.42$), although the effect size is small. This may be because Group B started the treatment with the more difficult

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Table 4. Training data. Time in minutes and number of items correctly typed during round 1 (/8) with standard deviations in parentheses

		S1 Time	S2 Time	S3 Time	S3 Round 1 Correct (/8)
Treatment Group	Group A (S1: StrA; S2: StrB)	26 (15)	25 (13)	17 (11)	2.3 (2.1)
	Group B (S1: StrB; S2: StrA)	31 (14)	26 (14)	18 (11)	2.2 (1.8)
Age Group	Adolescents	22.3 (12.8)	20.9 (10.7)	13.8 (7.0)	2.9 (2.0)
	Children	37.5 (12.7)	32.3 (14.1)	23.3 (12.9)	1.3 (1.5)
Time on Task	Faster	18.7 (7.5)	17.6 (6.6)	11.7 (5.3)	3.2 (1.9)
	Slower	42.6 (10.3)	37.4 (11.5)	26.2 (11.1)	0.8 (0.9)
Proficiency	High	19.0 (10.4)	16.1 (7.0)	9.3 (4.1)	3.3 (1.7)
	Medium	27.9 (13.0)	25.6 (11.5)	16.9 (6.5)	2.3 (1.9)
	Low	38.6 (14.5)	34.0 (14.7)	25.2 (13.0)	1.2 (1.6)
Together		29 (15)	26 (13)	18 (11)	2.2 (1.9) ISI-7: 0.8 (1.0) ISI-1: 1.4 (1.3) StrA: 1.4 (1.2) StrB: 0.9 (1.2)

Structure B. In contrast, for Group A, the difficulty may have been offset by the practice effects of having already completed a training session for Structure A.

Comparing times on task between age groups, children spent significantly more time on all sessions compared with adolescents ($t[101] = 5.984$, $p < .001$, $d = 1.197$; $t[75.306] = 4.434$, $p < .001$, $d = 0.994$; $t[59.532] = 4.445$, $p < .001$, $d = 0.927$). Times on task were also significantly different for the three proficiency groups for all three sessions, with significant differences between high to medium proficiency (S1: $t[61] = 2.818$, $p = .006$, $d = 0.420$; S2: $t[60.759] = 4.087$, $p < .001$, $d = 1.001$; S3: $t[60.306] = 5.608$ $p < .001$, $d = 1.381$), medium to low proficiency (S1: $t[65] = 3.154$, $p = .002$, $d = 0.777$; S2: $t[44.277] = 2.489$, $p = .017$, $d = 0.643$; S3: $t[34.779] = 3.077$, $p = .004$, $d = 0.810$), and high to low proficiency (S1: $t[46.759] = 5.562$, $p < .001$, $d = 1.558$; S2: $t[36.637] = 5.569$, $p < .001$, $d = 1.563$; S3: $t[31.938] = 5.975$, $p < .001$, $d = 1.639$).

For S3, in which items from both structures were presented for the second time, paired samples *t*-tests showed that participants entered more correct responses in round one from Structure A than from the more difficult Structure B, $t[101] = 3.488$, $p = .001$, $d = 0.35$, regardless of ISI, and also more from ISI-1 than ISI-7, $t[101] = 4.854$, $p < .001$, $d = 0.48$), regardless of structure. This confirms that Structure B and ISI-7 imposed more difficulty at S3. Compared between faster and slower learners, the faster learners achieved significantly more correct retrievals on round one of S3 than slower learners, $t[92.285] = 8.410$, $p = <.001$, $d = 1.596$. This supports the notion that time on task was related to ability. As for age groups, adolescents entered significantly more successful responses in this round than

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Table 5. Posttest scores within participants

		RI-7 (/16)	RI-28 (/16)	Overall (/16)
ISI	1	8.40 (5.51)	4.35 (5.41)	6.50 (5.81)
	7	7.71 (5.68)	5.04 (5.50)	6.45 (5.73)
Structure	A	9.39 (5.49)	5.29 (6.11)	7.46 (6.12)
	B	6.73 (5.40)	4.09 (4.66)	5.49 (5.21)

Table 6. Posttest scores between participants

		RI-7 (/32)	RI-28 (/32)	Overall (/32)
Age	Adolescents	19.14 (8.68)	11.77 (9.63)	15.74 (9.79)
	Children	12.19 (9.05)	6.52 (8.35)	9.46 (9.10)
Proficiency	High	21.94 (7.09)	13.50 (11.12)	18.81 (9.54)
	Medium	18.05 (8.15)	10.04 (7.71)	13.96 (8.82)
	Low	8.00 (8.62)	4.64 (8.68)	6.48 (8.67)
Time	Faster	20.58 (8.29)	12.34 (9.83)	16.60 (9.90)
	Slower	10.13 (8.70)	7.20 (8.46)	8.77 (8.62)

children, $t[100.952] = 4.519$, $p = <.001$, $d = 0.880$. Finally, proficiency also predicted correct retrievals in this round with significant differences between high to medium proficiency ($t[61] = 2.061$, $p = .044$, $d = 0.548$), medium to low proficiency ($t[65] = 2.498$, $p = .015$, $d = 0.633$), and high to low proficiency ($t[48] = 4.510$, $p = < .001$, $d = 1.277$).

To summarize, the training data supports the rationale that the variables in this study imposed differing levels of difficulty during training. Fewer items were remembered at the start of S3 from the longer ISI (7 days) and from the more difficult structure (B). The latter also took more time to complete when it was presented as the first structure. Faster times on task were associated with more correct retrievals at the start of S3, and both older and more proficient learners performed better on time and retrieval measures. Effect sizes for comparisons of learner-related difficulty were medium to high, whereas for ISI and structure the effect sizes were low. No significant differences in overall training performance were found between randomly assigned treatment groups or RI groups.

Posttest results

Table 5 shows the results for posttests for each ISI and structure, according to RI. Table 6 shows the breakdown of total scores by learner differences. Descriptively, participants at RI-7 scored higher than those at RI-28 in both conditions and both structures. Within each RI group, Structure A obtained higher scores than the more difficult Structure B, especially at RI-7. ISI-1 scores are slightly higher than ISI-7 scores at RI-7, but this is reversed at RI-28. Regarding learner differences

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Table 7. Summary of statistical models

MODEL	Predictors	*F*	*p*
Model 1	ISI, RI, ISI*RI	8.288	<.001
Model 2	ISI, RI, Structure, ISI*RI, RI*Structure	25.363	<.001
Model 3	ISI, RI, Age, ISI*RI	9.128	<.001
Model 4	ISI, RI, Proficiency, ISI*RI, ISI*Proficiency	9.540	<.001
Model 5	ISI, RI, TimeOnTask, ISI*RI, ISI*TimeOnTask, ISI*RI*TimeOnTask	10.134	<.001

(Table 6), adolescents obtained higher scores than children, faster participants achieved higher scores than slower participants, and scores increased with proficiency level. The large standard deviations in Table 5 indicate high variance among participants, with noticeably higher variance at RI-28. Table 6 shows that variance decreases considerably in favorable conditions (older, higher proficiency, faster), with much higher standard deviations in conditions of higher difficulty, relative to scores. This could be interpreted as lower difficulty conditions leveling the playing field.

Table 7 summarizes the statistical models, with a more detailed summary in Appendix E. Additional statistics for nonsignificant interactions and all estimated means with pairwise comparisons for each main effect and interaction can be found in Appendix S1 in the supplementary online materials.

Model 1: ISI and RI

Model 1 included ISI, RI, and their interaction. The main effect of ISI was not significant (ISI-1: $M = .391$, $SE = .003$; ISI-7: $M = .395$, $SE = .029$), $OR = 1.012$, $p = .829$, but RI-7 scores ($M = .504$, $SE = .039$) were significantly higher than RI-28 scores ($M = .293$, $SE = .038$), $OR = 2.451$, $p < .001$. The interaction (Figure 4) was also significant, though with a small effect size and overlapping standard errors. At RI-7, ISI-1 scores ($M = .525$, $SE = .040$) were higher than ISI-7 scores ($M = .482$, $SE = .040$), $OR = 1.189$, $p = .014$, whereas at RI-28, ISI-7 scores ($M = .315$, $SE = .040$) were higher than ISI-1 scores ($M = .272$, $SE = .038$), $OR = 1.234$, $p = .012$. The drop in scores from RI-7 to RI-28 was therefore more pronounced for ISI-1 items. To summarize, there was no main effect of ISI, but a small crossover interaction with RI was statistically significant.

Model 2: ISI, RI, and structure

Model 2 added the predictor of structure, with Structure B being more difficult than A. The main effect of structure was significant. Structure A scores ($M = .455$, $SE = .030$) were higher than Structure B scores ($M = .333$, $SE = .028$), $OR = 1.733$, $p < .001$. Although the interaction with ISI (Figure 5) was not statistically significant, $F = 1.164$, $p = .281$, there appears to be a trend towards higher scores for the easier structure with the longer lag. However, there was a significant interaction with RI (Figure 6), as the difference in scores at RI-7 (Structure A:

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

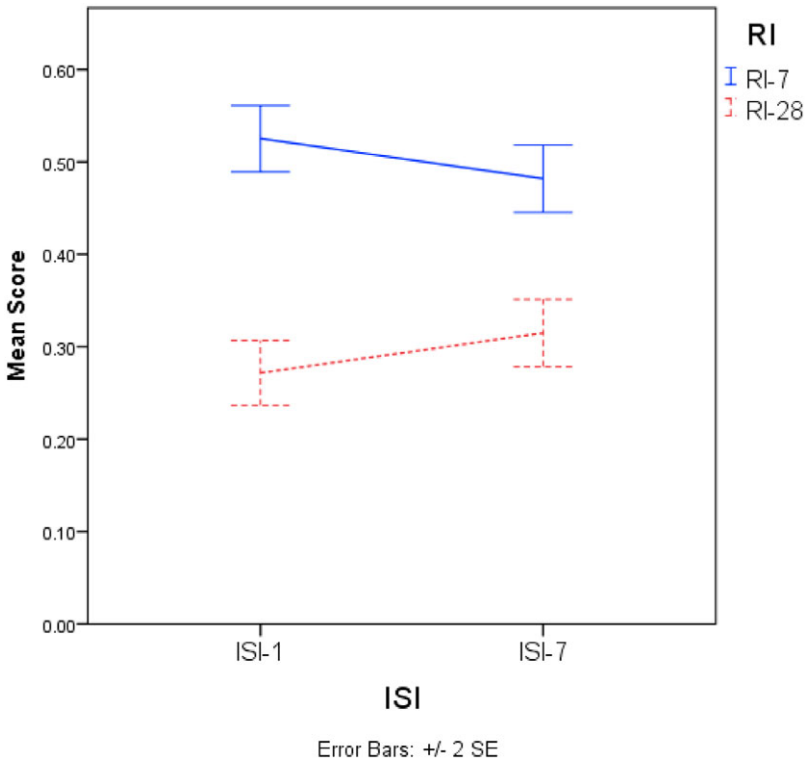


Figure 4. Model 1: ISI by RI interaction.

$M = .585$, $SE = .040$; Structure B: $M = .422$, $SE = .040$), $OR = 1.976$, $p < .001$, was more pronounced than at RI-28 (Structure A: $M = .331$, $SE = .040$; Structure B: $M = .254$, $SE = .037$) $OR = 1.440$, $p < .001$. However, all effects were small.

Model 3: ISI, RI, and age

The third model compared ISI and RI effects for the two age groups of children (ages 10–12) and adolescents (ages 13–18). The main effect of age was significant. Adolescents' scores ($M = .484$, $SE = .038$) were significantly higher than children's scores ($M = .281$, $SE = .038$), $OR = 2.358$, $p < .001$. Age did not interact with ISI (Figure 7) or with RI (Figure 8).

Model 4: ISI, RI, and proficiency

Model 4 included ISI, RI, and proficiency with three levels, as well as their significant interactions. The model produced a significant, moderate main effect for proficiency, where higher proficiency learners obtained higher scores (high: $M = .554$, $SE = .059$; medium: $M = .436$, $SE = .045$; low: $M = .185$, $SE = .041$). Low proficiency scores were significantly lower than high proficiency scores, $OR = 5.621$, $p < .001$, and medium proficiency scores, $OR = 3.372$, $p = < .001$.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

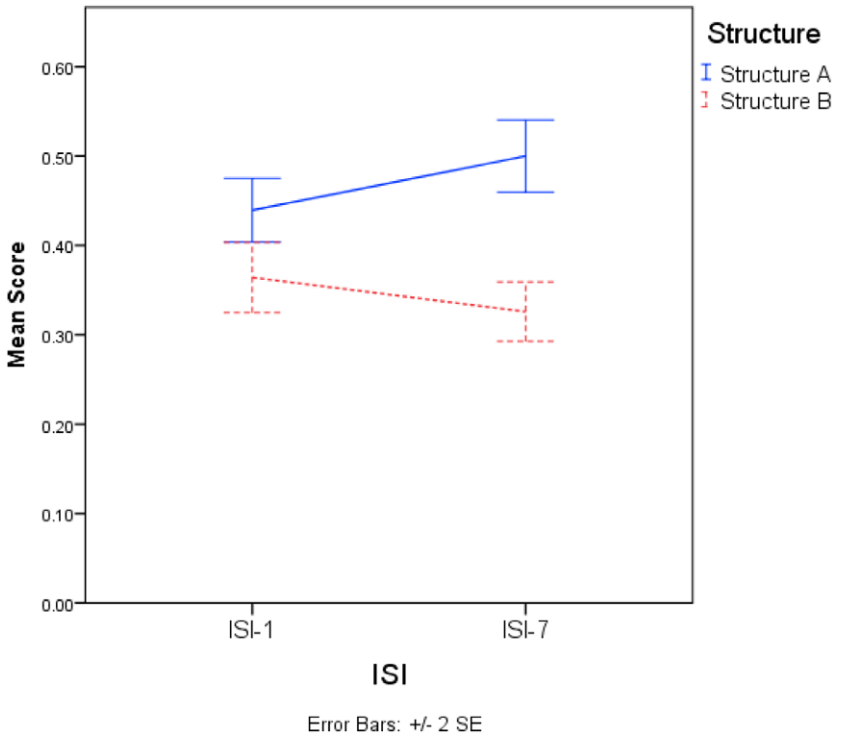


Figure 5. Model 2: ISI by structure interaction.

The difference between medium and high proficiency scores approached but did not reach significance, $OR = 1.667$, $p = .088$.

The interaction with proficiency and ISI (Figure 9) was also significant. With medium proficiency, there was no significant difference between ISI-1 ($M = .440$, $SE = .047$) and ISI-7 ($M = .431$, $SE = .045$) scores, $OR = 1.027$, $p = .675$. However, high proficiency led to significantly better scores for ISI-7 ($M = .599$, $SE = .059$) compared with ISI-1 ($M = .508$, $SE = .062$), $OR = 1.361$, $p = .010$. Conversely, low proficiency led to significantly better scores for ISI-1 ($M = .209$, $SE = .045$) compared with ISI-7 ($M = .164$, $SE = .040$), $OR = 1.499$, $p = .004$. Additionally, the difference between high and low proficiency scores was considerably larger at ISI-7, $OR = 8.696$, $p < .001$, than at ISI-1, $OR = 4.270$, $p < .001$.

Model 5: ISI, RI, and time on task

The final model included ISI, RI, and time on task, with their significant interactions. A significant, moderate main effect was found for time on task, whereby faster participants ($M = .516$, $SE = .040$) scored higher than slower participants ($M = .264$, $SE = .041$), $OR = 3.029$, $p < .001$. An interaction between ISI and time

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

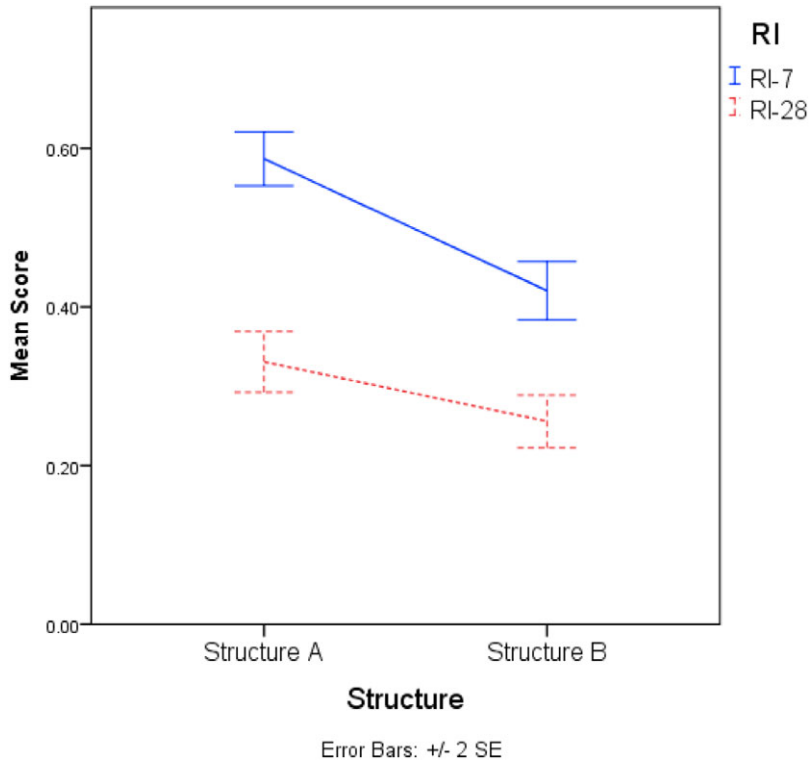


Figure 6. Model 2: RI by structure interaction.

on task (Figure 10) was also significant with small effects. For faster participants, ISI-7 scores ($M = .541$, $SE = .041$) were significantly higher than ISI-1 scores ($M = .490$, $SE = .041$), $OR = 1.228$, $p = .010$, but for slower participants, ISI-1 scores ($M = .291$, $SE = .045$) were significantly higher than ISI-7 scores ($M = .239$, $SE = .040$), $OR = 1.454$, $p = .001$.

Additionally, the difference between faster and slower participants was larger for ISI-7 scores, $OR = 4.495$, $p < .001$, compared with ISI-1 scores, $OR = 2.519$, $p = .001$.

A three-way interaction with ISI, RI and time on task (Figure 11) was also significant, with small effects. The interaction is evident among the slower participants, for whom ISI-1 scores ($M = .408$, $SE = .063$) were significantly higher than ISI-7 scores ($M = .226$, $SE = .053$) at RI-7, $OR = 2.364$, $p < .001$, but at RI-28, ISI-7 scores ($M = .253$, $SE = .060$) were slightly higher than ISI-1 scores ($M = .197$, $SE = .055$), though both ISI scores at this RI are very low and the difference only narrowly reaches significance, $OR = 1.383$, $p = .039$. As for faster participants, the longer lag was significantly better at RI-7 (ISI-1: $M = .611$, $SE = .054$; ISI-7: $M = .675$, $SE = .052$), $OR = 1.325$, $p = .009$, but not at RI-28 (ISI-1: $M = .371$, $SE = .055$; ISI-7: $M = .401$, $SE = .056$) $OR = 1.136$, $p = .245$. Another way to view this interaction is that RI-7 scores were always higher than RI-28 scores, apart from in the combination of slower participants and longer lag.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

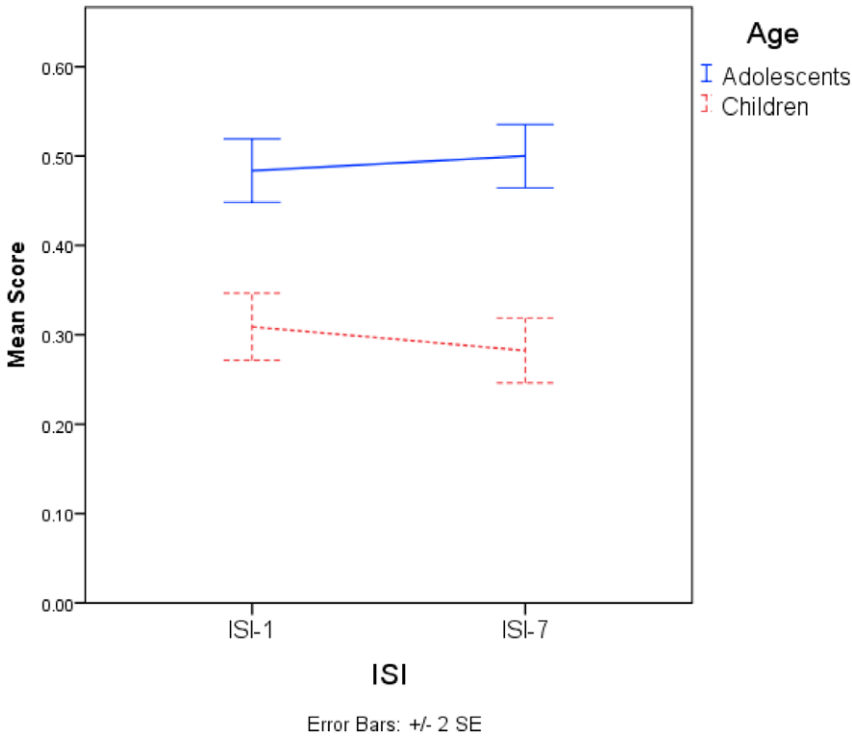


Figure 7. Model 3: ISI by age interaction.

Discussion

In the present experiment, 117 ELLs studied two grammatical structures of differing degrees of linguistic difficulty by retrieval and feedback on Quizlet, using a dropout criterion of one correct response, with two sessions per structure. These were counterbalanced over two different ISIs, 1 day and 7 days, and tested after either 1 week or 1 month. Results were also compared for learners of differing age, proficiency, and the time required to complete the training. We now present a summary of findings from this experiment and their implications for the DDF's account of lag effects in L2 practice.

RQ1

The first RQ concerned the overall effect of ISI measured at RI-7 and RI-28. Results showed no main effect of ISI in this experiment, contrary to our hypothesis that the shorter lag would lead to higher scores. However, there was a small but significant crossover interaction with RI, whereby a shorter lag was better for RI-7 and a longer lag was better for RI-28. This result is reminiscent of Rohrer and Pashler (2007)'s optimal ISI ratio of 10–20% of RI. The two combinations with higher scores had ratios of 14% (ISI-1:RI-7) and 24% (ISI-7:RI-28), compared with 100% (ISI-7:RI-7) and 3.5% (ISI-1:RI-28). However, this interaction is better explained after reviewing the rest of the findings.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

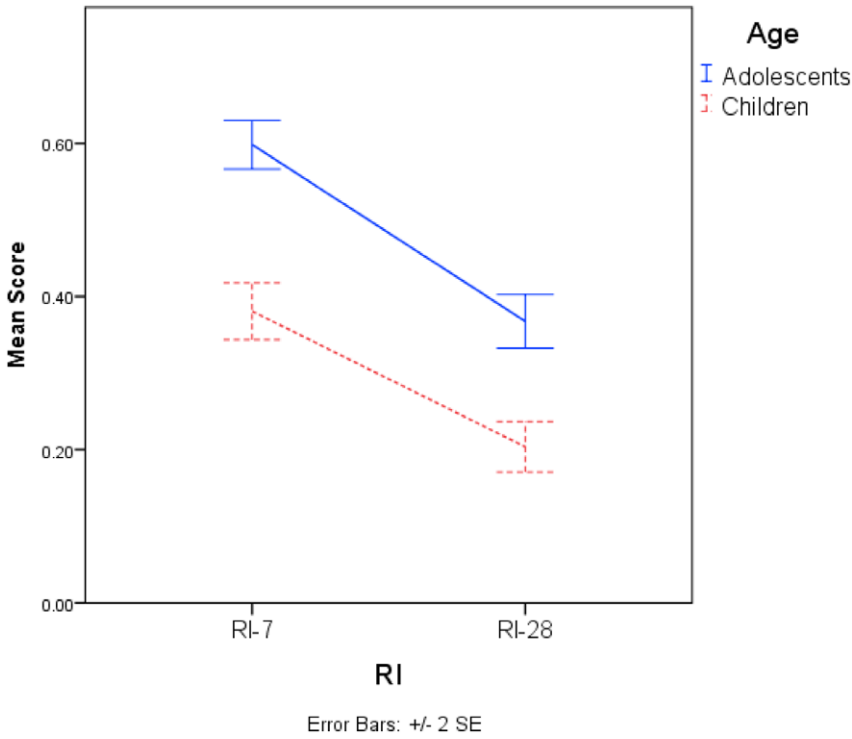


Figure 8. Model 3: RI by age interaction.

RQ2

The second RQ concerned how ISI may interact with other sources of difficulty. Firstly, two different grammatical structures were used to examine linguistic difficulty. Training data seemed to confirm the study's rationale that the more difficult structure (B) imposed more difficulty during training. A main effect was found, but contrary to our hypothesis there was no interaction with ISI. It seems that the effect of linguistic difficulty outweighed any effects of ISI, though the difference in scores was descriptively larger at ISI-7. Thus, these results are in line with Bahrick and Phelps (1987) and Suzuki (2017) in that lag effects did not significantly interact with linguistic difficulty. Based on the descriptive trend towards a greater difference at ISI-7, it may be expected that target forms with more extreme differences in complexity would have produced a significant interaction with ISI. Nonetheless, the hypothesized interaction between these two difficulty factors is not confirmed in this study.

Secondly, lag effects for adolescents and children were compared. Age was found to be a significant moderator of scores, with adolescents outperforming children as a whole. This is unsurprising given that they were learning the same complex, cognitively demanding materials. Training data also confirmed that children experienced more difficulty during training. However, as with structure, the hypothesized interaction with ISI was not found. Shorter lags were not better for children and longer

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

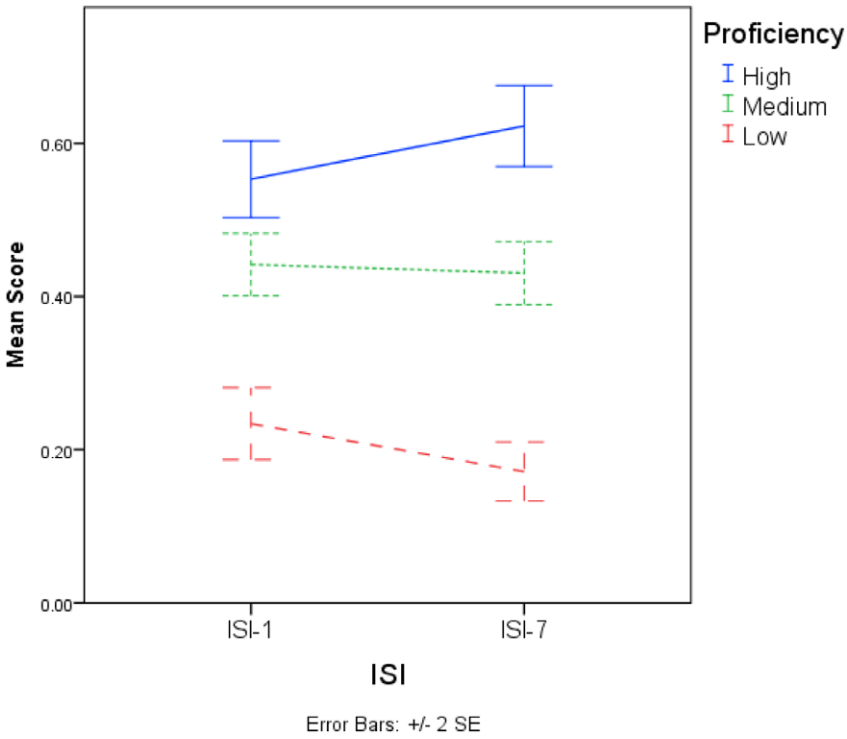


Figure 9. Model 4: ISI by proficiency interaction.

lags were not better for adolescents. A significant advantage to ISI-1 for children was expected at RI-7 based on Küpper-Tetzel et al. (2014), who demonstrated this effect for 11–13 year olds. The results of the current study do not show a significant effect, but do show a trend in the same direction. The present results are more similar to Kasprowicz et al. (2019) and Rogers and Cheung (2020a, 2020b) who found minimal differences in ISI conditions for young children using similar lags.

In contrast, proficiency level significantly moderated the direction of lag effects. Training data confirmed that lower proficiency led to more difficulty during training. For participants with higher L2 proficiency (B2+), the longer lag added desirable difficulty, while for lower level participants (A1/2), the easier shorter lag was better. The difference between these groups was particularly apparent in the more difficult ISI-7 condition. No differences in ISI items were observed for the participants with a medium (B1) level.

Time on task also proved to be a significant moderator of lag effects. Faster participants benefited from a longer lag while slower participants did better with a shorter lag. Additionally, the three-way interaction with RI showed that for slower participants with ISI-7, scores were very low even for the short RI. As with proficiency, time on task also predicted results more strongly for ISI-7 than for ISI-1. This is similar to how aptitude scores from Suzuki and DeKeyser (2017b) and Suzuki (2019) predicted L2 scores at ISI-7 only. Taken together, this could

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

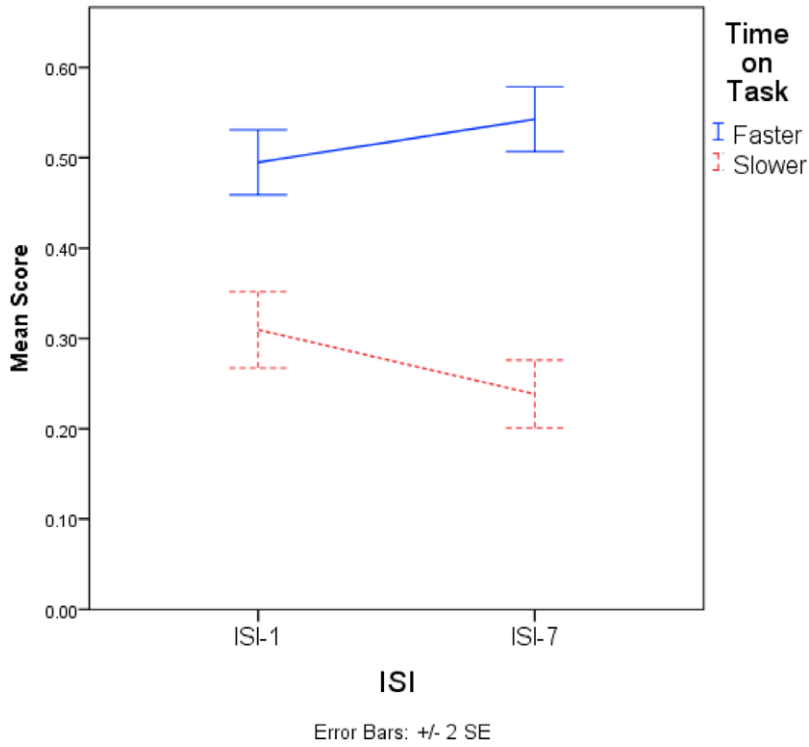


Figure 10. Model 4: ISI by time on task interaction.

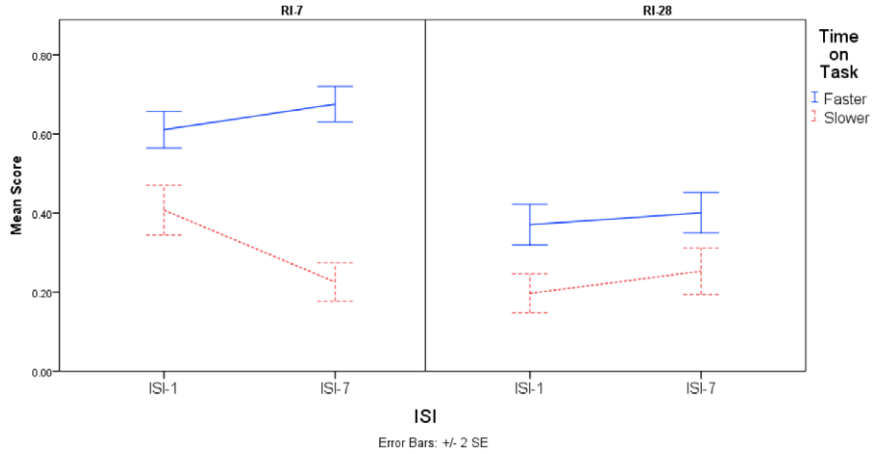


Figure 11. Model 5: ISI by time on task interaction at RI-7 and RI-28.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

indicate that learner-related differences play a stronger role in the more challenging ISI condition, which is typical for aptitude-treatment interaction research (DeKeyser, 2021).

Individualized pacing during training may have been expected to reduce variability between learners, given that all the participants learned to the same criterion of one successful retrieval, and the observed variability may therefore be counterintuitive. For grammar items, an advantage might even have been expected for slower participants, whose greater number of incorrect trials will have led to more practice and more feedback. A possible explanation for this might be found in the Retrieval Effort Hypothesis (Pyc & Rawson, 2009). Faster participants will have had more intervening items between each response, since faster participants were more likely to achieve correct retrievals in earlier rounds, and therefore their successful retrievals will have been more effortful. In contrast, participants that only achieved one or two correct responses per round experienced a continually decreasing number of intervening items. The feedback for more difficult items would then be more recent and more highly activated in working memory, and with each cumulative exposure to the correct response, the effort for the eventual correct retrieval would inevitably decrease. Therefore, successful retrievals that required more trials to achieve also required less retrieval effort, as a combination of higher activation and more practice. This reduced effort for successful retrievals is hypothesized to create weaker memory traces than more effortful successful retrievals.

Having reviewed the findings of RQ2, it is now clear that the significant interaction between ISI and RI is not applicable to all participants, but is rather the sum of different experiences. The ISI-1+RI-7 advantage comes from slower participants and those with lower proficiency. By RI-28, their scores drop and the ISI-7+RI-28 advantage emerges from participants with higher proficiency who better retained their knowledge and performed better with the longer lag at both RIs. Therefore, without taking learner differences into account, one could mistakenly conclude that ISI-1 is always best for RI-7, and ISI-7 is always best for RI-28. The present data demonstrate that the optimal ISI for each RI depends on the learner, and highlights the importance of considering these differences in future research.

Theoretical implications

The above findings partially confirm the predictions of the DDF for L2 practice (Suzuki et al., 2019). Firstly, difficulty is created by a combination of different sources. In this experiment, RI, structure, age, proficiency, and time on task all significantly affected outcomes. Higher scores were obtained at the shorter RI, for the easier target structure, for older learners, for higher proficiencies, and for faster times. In all of these comparisons, the higher scores were obtained for the condition with least difficulty. Put differently, adding difficulty to training was not desirable. This could indicate that the task of learning grammar through digital flashcards, as implemented in the current study, already involves high retrieval effort and therefore any further difficulty (e.g., more complex target forms or lower cognitive abilities) was not desirable.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

In contrast to other measures of difficulty, ISI had no main effect, and the direction of its benefit changed according to learner-related difficulty. Disadvantaged learners, as evidenced by their higher time on task or lower proficiency, were hindered by a longer lag, but for learners that found the task easier, the added difficulty of a longer lag proved to be desirable. In fact, ISI was the only variable to which adding difficulty was desirable. Based on these observations, linguistic difficulty and age had the most robust effect on scores, with no interactions with ISI. Next, the learner-related variables of proficiency and time on task had main effects but also interacted with ISI. Lastly, ISI only played a role as a moderator of learner-related difficulties, and its effect sizes were small. Therefore, ISI seems to have a comparatively small effect on learning outcomes. While this does confirm the prediction that lag effects depend on other sources of difficulty, it also highlights the greater importance of these other sources in determining outcomes. It is also noteworthy that linguistic difficulty did not interact with other variables, nor has it in prior research (Bahrick & Phelps, 1987; Suzuki, 2017). This leaves the question open as to whether linguistic difficulty could interact with lag effects, given the right conditions, for example if the structures were easier than in the present study or more different to each other.

Limitations

The present experiment is subject to certain limitations that should be addressed in future research. Firstly, the use of Quizlet as a tool brings many advantages, but prevents the accurate tracking of training metrics such as the number of trials to reach criterion and time per trial. A different platform might better elucidate the difficulties experienced by learners during training and provide a more refined measure of time on task. Secondly, highly complex target structures were chosen because all participants in this study were daily users of English for academic purposes. This complexity, together with the short training period, probably explains the low posttest scores overall, but especially in the case of younger learners. It would be interesting to use the same design with simpler structures or use more sessions in order to increase the amount of learning for all participants for both pedagogical and research purposes.

Finally, the unpredictable regulation changes related to COVID-19 necessitated that some sessions were performed online, without in-person supervision. While this may also add some ecological validity to the findings, it would be preferable from a methodological point of view to conduct a study where all sessions were supervised in person. A side-effect of this lack of in-person supervision was that some participants did not follow instructions as intended. In order to ensure that the data under analysis were valid, it was decided to conservatively exclude any participants that did not provide evidence of their correct adherence to the procedure, and a large majority of them came from the lower sets in their grade level. As a result, higher abilities are overrepresented in this study. Just as a majority of prior research has taken place among undergraduate students, with a certain academic ability and motivation to participate, there seems to be a natural bias in research against the types of learners that might benefit the most from better learning strategies. Future research should consider designing experiments to better include these learners.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Concluding remarks

To conclude, the DDF proposed by Suzuki et al. (2019) seems to be a promising framework to use to examine optimal L2 practice. Specifically, we have suggested that the conflicting results reported in the literature about lag effects for L2 grammar learning might be due to different degrees of difficulty with regards to practice conditions. Moreover, the results of our study suggest that learner-related sources of difficulty are crucial for understanding lag effects in grammar learning. When a task is less challenging, adding difficulty can be beneficial, and using a longer lag is one possible manipulation to enhance memory for easier tasks or for learners with higher abilities. However, the benefits found in this paper, although statistically significant, were small or moderate in terms of effect size. When applying this finding to an authentic classroom schedule, the advantages of adding a longer lag for grammar practice must be considered along with the risks of imposing too much difficulty on learners of lower ability. For those who found the treatment more challenging, the shorter lag was necessary to retain the acquired knowledge even at the 7-day posttest. Therefore, the small benefit of the longer lag for some is outweighed by its detriment to others, and a shorter lag would be more appropriate for a mixed-ability class. Of course, there is no one-size-fits-all best practice for choosing an ISI. Teachers should pay attention to the difficulty experienced by their students, and the time they require to complete a task seems to be a fair indication of this difficulty, at least as a relative measure to other students. It is hoped that researchers pay more attention to individual variability in future research as a predicting variable rather than as a factor to control for, as this paper has shown that individual ability not only influences the degree of outcomes, but the direction of outcomes as well.

Supplementary material. For supplementary material accompanying this paper visit https://doi.org/10.1017/S0142716421000631

Acknowledgments. This research was funded by the Spanish Ministry of Science and Innovation (PID2019-110536GB-I00). We would like to thank the editor and the three anonymous reviewers for their insightful comments.

Competing interests. The authors declare none.

Notes

1. Learner variables were analyzed as categorical rather than continuous variables for two reasons. Firstly, the other variables included in the study were also categorical (ISI, RI and linguistic difficulty). Secondly, and most importantly, the binary logit model in SPSS chosen for the statistical analyses would use a continuous variable as a control and would not provide estimated means or visual comparisons. The statistics would give only the effects from increments of the variable, for example the change in likelihood of a correct response by each additional minute on task, which does not answer our research questions well.

2. No corrections for multiple comparisons were made because each *t*-test was testing a different hypothesis.

References

Andarab, M. S. (2017). The effect of using Quizlet Flashcards on learning English vocabulary. *113th The IIER International Conference*.

Ashcroft, R. J., Cvitkovic, R., & Praver, M. (2018). Digital flashcard L2 vocabulary learning out-performs traditional flashcards at lower proficiency levels: A mixed-methods study of 139 Japanese university students. *The EuroCALL Review*, **26**(1), 14. https://doi.org/10.4995/eurocall.2018.7881

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, **52**(4), 566–577. https://doi.org/10.1016/j.jml.2005.01.012

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13**(2), 344–349. https://doi.org/10.1037/0278-7393.13.2.344

Bathelt, J., Gathercole, S. E., Johnson, A., & Astle, D. E. (2018). Differences in brain morphology and working memory capacity across childhood. *Developmental Science*, **21**(3), e12579. https://doi.org/10.1111/desc.12579

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, **31**(4), 635–650. https://doi.org/10.1017/S0142716410000172

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

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bjork, R. A. (2018). Being suspicious of the sense of ease and undeterred by the sense of difficulty: Looking back at Schmidt and Bjork (1992). *Perspectives on Psychological Science*, **13**(2), 146–148. https://doi.org/10.1177/1745691617690642

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, **74**(4), 245–248. https://doi.org/10.1080/00220671.1981.10885317

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**(4), 236–246. https://doi.org/10.1027/1618-3169.56.4.236

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science*, **19**(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

DeKeyser, R. (ed.) (2021). *Aptitude-Treatment Interaction in Second Language Learning.* Amsterdam, The Netherlands: Benjamins.

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

Dizon, G. (2016). Quizlet in the EFL classroom: Enhancing academic vocabulary acquisition of Japanese university students. *Teaching English with Technology*, **16**(2), 40–56.

Fandakova, Y., Sander, M. C., Werkle-Bergner, M., & Shing, Y. L. (2014). Age differences in short-term memory binding are related to working memory performance across the lifespan. *Psychology and Aging*, **29**(1), 140–149. https://doi.org/10.1037/a0035347

Franciosi, S. J., Yagi, J., Tomoshige, Y., & Ye, S. (2016). The effect of a simple simulation game on long-term vocabulary retention. *CALICO Journal*, **33**(3), 16. https://doi.org/10.1558/cj.v33i2.26063

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study. *Applied Cognitive Psychology*, **30**(5), 700–712. https://doi.org/10.1002/acp.3245

Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, **38**(2), 163–175. doi: 10.1017/S0272263116000176

IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp.

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. *The Modern Language Journal*, **103**(3), 580–606. https://doi.org/10.1111/modl.12586

Kim, Y. J., Skalicky, S., & Jung, Y. J. (2020). The role of linguistic alignment on question development in face-to-face and synchronous computer-mediated communication contexts: A conceptual replication study. *Language Learning*, **70**(3), 643–684. https://doi.org/10.1111/lang.12393

Korlu, H., & Mede, E. (2018). Autonomy in Vocabulary Learning of Turkish EFL Learners. *The EuroCALL Review*, **26**(2), 58. https://doi.org/10.4995/EUROCALL.2018.10425

Kornell, N., & Bjork, R. (2008). Optimising self-regulated study: The benefits-and costs-of dropping flashcards. *Memory*, **16**(2), 125–136. https://doi.org/10.1080/09658210701763899

Koval, N. G. (2019). Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics*, **40**, 1103–1139.

Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, **42**(3), 373–388. https://doi.org/10.1007/s11251-013-9285-2

Li, M., & Dekeyser, R. (2019). Distribution of practice effects in the acquisition and retention of L2 Mandarin tonal word production. *Modern Language Journal*, **103**(3), 607–628. https://doi.org/10.1111/modl.12580

Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System*, **42**(1), 412–428. https://doi.org/10.1016/j.system.2014.01.014

Muñoz, C. (2006). Chapter 1. The Effects of Age on Foreign Language Learning: The BAF Project. In C. Muñoz (Ed.), *Age and the Rate of Foreign Language Learning* (pp. 1–40). Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781853598937-003

Muñoz, C. (2007). Age-related differences and second language learning practice. In R. DeKesyer (Ed.), *Practice in a Second Language* (pp. 229–255). Cambridge University Press. https://doi.org/10.1017/CBO9780511667275.014

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, **29**(4), 578–596. https://doi.org/10.1093/applin/amm056

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/S0272263114000825

Nakata, T. (2020). Learning words with flash cards and word cards. In S. Webb (Ed.), *Routledge Handbook of Vocabulary Studies* (pp. 304–319). New York, NY: Routledge. https://doi.org/10.4324/9780429291586-20

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**(4), 559–586. https://doi.org/10.1207/s15516709cog0000_14

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

Pyc, M. A., & Rawson, K. A. (2009). Testing the Retrieval Effort Hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, **60**(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Quizlet (2021). About Quizlet | Quizlet. (n.d.). Retrieved September 25, 2021, from https://quizlet.com/mission

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review*, **25**(4), 523–548. https://doi.org/10.1007/s10648-013-9240-4

Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, **49**(4), 857–866. https://doi.org/10.1002/tesq.252

Rogers, J., & Cheung, A. (2020a). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*, **24**(5), 616–641. https://doi.org/10.1177/1362168818805251

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

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, **16**(4), 183–186. https://doi.org/10.1111/j.1467-8721.2007.00500.x

Sanosi, A. B. (2018). The effect of Quizlet on vocabulary acquisition. *Asian Journal of Education and E-Learning*, **6**(4), 71–77. https://doi.org/10.24203/ajeel.v6i4.5446

Schmidt, R. A., & Bjork, R. A. (1992). New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychological Science*, **3**(4), 207–217. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Serfaty, J., & Serrano, R. (2020). Examining the potential of digital flashcards to facilitate independent grammar learning. *System*, **94**, 102342. https://doi.org/10.1016/j.system.2020.102342

Serrano, R., & Huang, H. (2021). Time distribution and intentional vocabulary learning through repeated reading: a partial replication and extension. *Language Awareness*, 1–19. https://doi.org/10.1080/09658416.2021.1894162

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

Similarweb.com. (n.d.). Retrieved September 19, 2021, from https://www.similarweb.com/top-websites/category/science-and-education/education/

Spada, N., & Tomita, Y. (2010). Interactions Between Type of Instruction and Type of Language Feature: A Meta-Analysis. *Language Learning*, **60**(2), 263–308. https://doi.org/10.1111/J.1467-9922.2010.00562.X

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. https://doi.org/10.1111/lang.12236

Suzuki, Y. (2019). Individualization of practice distribution in second language grammar learning. *Journal of Second Language Studies*, **2**(2), 169–196. https://doi.org/10.1075/jsls.18023.suz

Suzuki, Y., & DeKeyser, R. (2017a). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, **21**(2), 166–188. https://doi.org/10.1177/1362168815617334

Suzuki, Y., & DeKeyser, R. (2017b). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics*, **38**(1), 27–56. https://doi:10.1017/S0142716416000084

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**(3), 713–720. https://doi.org/10.1111/modl.12585

Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. In B. H. Ross (Ed.), *The psychology of learning and motivation*: Vol. 60. (pp. 113–189). Elsevier Academic Press.

UCLES (University of Cambridge Local Examination Syndicate) (2001). *Quick Placement Test*. Oxford: Oxford University Press.

Ullman, M. T., & Lovelett, J. T. (2018). Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, **34**(1), 39–65. https://doi.org/10.1177/0267658316675195

Vaughn, K. E., Dunlosky, J., & Rawson, K. A. (2016). Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion? *Memory and Cognition*, **44**(6), 897–909. https://doi.org/10.3758/s13421-016-0606-y

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, **20**(6), 568–579. https://doi.org/10.1080/09658211.2012.687052

Yurgelun-Todd, D. A., Killgore, W. D. S., & Young, A. D. (2002). Sex differences in cerebral tissue volume and cognitive performance during adolescence. *Psychological Reports*, **91**(3 Pt 1), 743–757. https://doi.org/10.2466/pr0.2002.91.3.743

Appendix A. Training and Test Items

Training Items:

Target Forms	Prompts
Structure A	
I will have been studying for 3 hours by the time I see you.	I will start studying at 3pm. I will see you at 6pm. (I will continue to study)
I will have been living in Thailand for 2 months by the time I start my job.	I will start living in Thailand in March. I will start my job in May. (I will continue living in Thailand)
I will have been studying for 4 days by the time I meet my teacher.	I will start studying on Monday morning. I will meet my teacher on Thursday evening. (I will continue studying)
I will have been doing this test for 5 minutes by the time I understand what I need to do.	I will start doing this test at 12:05pm. I will understand what I need to do at 12:10pm. (I will continue doing it)
I will have been shopping for 20 minutes by the time I need to find my friend.	I will start shopping at 4pm. I will need to find my friend at 4:20pm. (I will continue shopping)
I will have been going to Southbridge for 4 years by the time I take my IGCSEs.	I will start going to Southbridge in 2016. I will take my IGCSEs in 2020. (I will continue to go to Southbridge)
I will have been sailing for 10 days by the time I reach Malaysia.	I will start sailing on June 2nd. I will reach Malaysia on June 12th. (I will continue sailing)
I will have been frozen in the ice for 100 years by the time Katara finds me.	I will be frozen in the ice in year 0. Katara will find me in year 100. (I will continue being frozen in ice for a few minutes after she finds me)
Structure B	
What would we have eaten if we hadn't climbed the mountain?	We climbed the mountain, so we ate rice. But imagine a different past.
What would you have done if you had found the money?	You didn't find the money, so you did nothing. But imagine a different past.
Where would he have gone if he had bought a car?	He didn't buy a car, so he didn't go anywhere. But imagine a different past.
Where would she have lived if she hadn't moved to Germany?	She moved to Germany, so she lived in Germany. But imagine a different past.
Who would have gotten sick if he hadn't worn a mask?	He wore a mask, so no one got sick. But imagine a different past.
Who would she have seen if she had gone to Thailand?	She didn't go to Thailand, so she didn't see anyone. But imagine a different past.
How would they have felt if they had seen the fire?	They didn't see the fire, so they felt happy. But imagine a different past.
How would you have danced if you had been tired?	You were not tired, so you danced like a crazy person. But imagine a different past.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

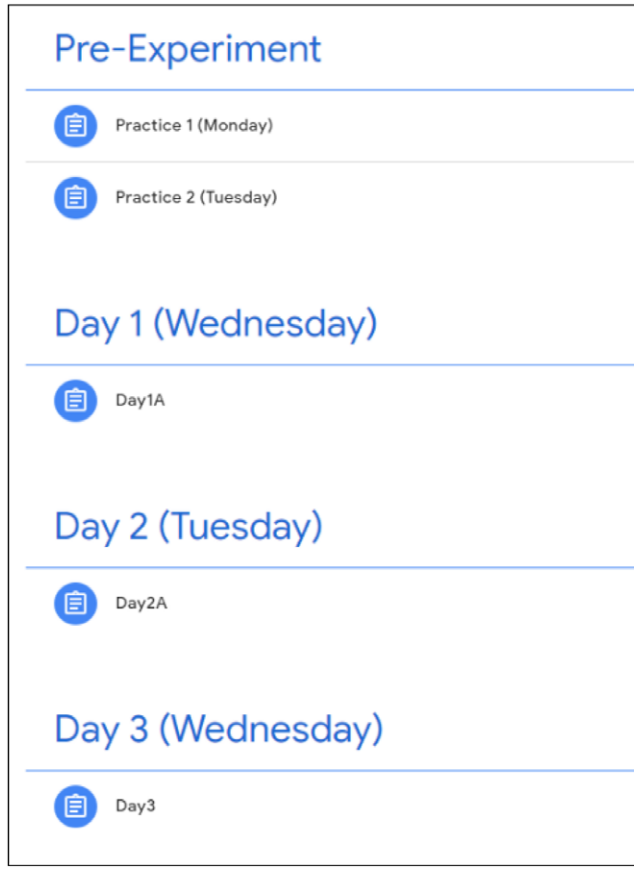
Test Items:

Example of Correct Response	Prompts
Structure A	
I will have been trying for 3 hours by the time I let you help me.	I will start trying at 3pm. I will let you help me at 6pm. (I will continue to try)
I will have been working there for 2 months by the time I meet my boss.	I will start working there in March. I will meet my boss in May. (I will continue working there)
I will have been fighting this war for 4 days by the time I learn to control my dragon.	I will start fighting this war on Monday morning. I will learn to control my dragon on Thursday evening. (I will continue fighting this war)
I will have been cutting my own hair for 5 minutes by the time I regret it.	I will start cutting my own hair at 12:05pm. I will regret it at 12:10pm. (I will continue cutting my own hair)
I will have been dancing for 20 minutes by the time I need to drink water.	I will start dancing at 4pm. I will need to drink water at 4:20pm. (I will continue dancing)
I will have been living in England for 4 years by the time I lose my accent.	I will start living in England in 2016. I will lose my accent in 2020. (I will continue living in England)
I will have been learning Chinese for 10 days by the time I know how to order a pizza.	I will start learning Chinese on June 2nd. I will know how to order a pizza on June 12th. (I will continue learning Chinese)
I will have been waiting for 100 years by the time I lose hope.	I will start waiting in year 0. I will lose hope in year 100. (I will continue waiting anyway)
Structure B	
What would you have worn if you hadn't felt happy?	You felt happy, so you wore orange. But imagine a different past.
What would I have found if I had looked in the box?	I didn't look in the box, so I didn't find anything. But imagine a different past.
Where would he have bought food if he had gone to Aeon Mall?	He didn't go to Aeon Mall, so he bought food at Kiwi Mart. But imagine a different past.
Where would she have stayed if she hadn't visited Angkor Wat?	She visited Angkor Wat, so she stayed at the Angkor Hotel. But imagine a different past.
Who would have done my work if I hadn't stayed home?	I stayed home, so someone else did my work. But imagine a different past.
Who would she have punched if she had been angry?	She wasn't angry, so she didn't punch anyone. But imagine a different past.
How would they have known about it if they hadn't asked?	They asked, so that's how they knew about it. But imagine a different past.
How would you have lived with yourself if you had eaten the puppy?	You didn't eat the puppy, so you have no problem living with yourself. But imagine a different past.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Appendix B. Google Classroom and Google Doc

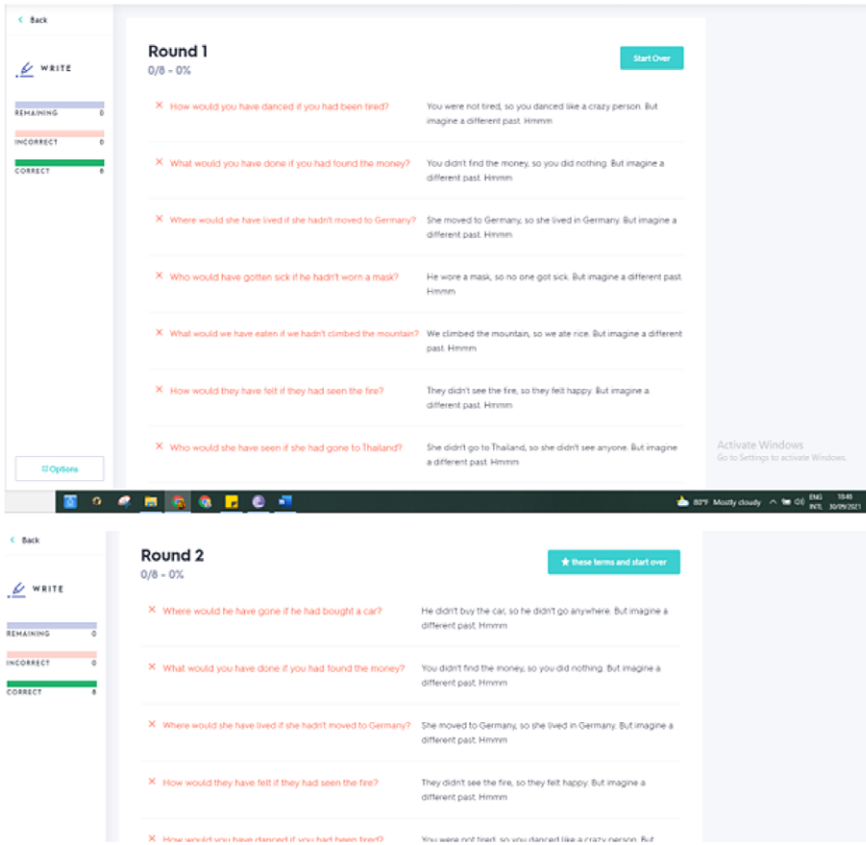
Participants saw their assignments in a Google Classroom. Each assignment only appeared at the appropriate time.



https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Participants recorded their progress in a Google Doc.

Time Started: 11:40
Set: https://quizlet.com/540667757/write
Time finished: 12:08
Total Time: 28 minutes



https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Appendix C. Details of pre-experimental procedures

Presentation of target structures:
It was explained as reporting the duration of an activity which has not yet started, but will continue after a certain future point in time. For this structure, they were told that they would be starting to learn Spanish next week and would visit Spain at Christmas, but would continue with Spanish classes after their trip. They then needed to think about the duration of their Spanish study from the point of view of their future trip. Structure B was the past perfect conditional in the interrogative (e.g., *What would you have done if you had found the money?*). This was explained to the students as wondering about a different past. To illustrate this, they were told that they had ordered fried noodles for breakfast but were wondering what they would have ordered if the restaurant had been out of fried noodles.

Practice activities:
The experiment was preceded with two preparation lessons. During these lessons, participants were shown a brief presentation about the target structures. Images showed events on a timeline to demonstrate the tenses conceptually, with the actual target forms omitted. Students then did their first practice, using Quizlet to answer five impossible-to-guess questions (e.g., "*How does your teacher take his coffee?*" [*Black*]). Through this, they learned to guess, look at feedback, and remember the answers. They also practiced taking screenshots, filling in their times, and submitting their documents. In the second preparation lesson they did their pretests and then another practice Quizlet set, this time using easy grammar materials. An example cue was "*Today, I didn't eat chicken, but tomorrow*" prompting them to type the end of the sentence in the past (*I ate chicken*) or future (*I will eat chicken*) tense, based on the use of "*tomorrow*" or "*yesterday*". They needed to work out what was required independently. Again, the emphasis was on the procedure of recording their progress correctly and using Quizlet in the intended manner.

Appendix D. Scoring criteria with examples

For Structure A, the points were for *I+will+have+been+gerund* and *for+time-period+by-the-time+I+present simple*. Examples of a 2-point, 1-point and incorrect response were, respectively, *I will have been living in England for 4 years by the time I lose my accent*; *I will have been living in England for 4 years by the time I will lose my accent*; *I will lose my accent in England 4 years after*. For Structure B, the points were awarded for *Question+would+subject+have+past participle* and *if+subject+had/hadn't+past participle*. Examples of a 2-point, 1-point and incorrect response were, respectively, *Where would she have stayed if she hadn't visited Angkor Wat?*; *Where would she have stayed if she haven't visited Angkor wat?*; *Where she have stay if hasn't visit Angkor Wat*.

The exact response could take any form, as long as the correct structures were used. For example, *What would you have worn if you hadn't felt happy?* and *How would you have felt if you hadn't worn orange?* were both correct answers to the prompt *You felt happy, so you wore orange. But imagine a different past*. Any unrelated mistakes, for instance missing a plural 's' or spelling a content word incorrectly, were ignored. A decision was taken to accept *wore* in place of *worn* because the past participle had not appeared in the training and the use of *wore* was highly frequent in posttests from participants that used past participles in every other response. This was put down to an incorrect assumption that the form would be known by all participants, and marking it as incorrect could produce misleading results. The response *What would you have been wearing* instead of *have worn* was also accepted, as it conveys an identical meaning to the target form.

https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

Appendix E. Summary of effects in statistical models

	Source	*F*	*df*1	*df*2	Sig.
MODEL 1: ISI & RI	**Corrected Model**	**8.288**	**3**	**1868**	**<.001**
	ISI	0.104	1	1868	0.747
	RI	13.856	1	1868	<.001
	ISI * RI	12.284	1	1868	<.001
MODEL 2: ISI, RI, & Structure	**Corrected Model**	**25.363**	**5**	**1866**	**<.001**
	ISI	2.01	1	1866	0.156
	RI	13.666	1	1866	<.001
	Structure	89.747	1	1866	<.001
	ISI * RI	7.282	1	1866	0.007
	RI * Structure	7.008	1	1866	0.008
MODEL 3: ISI, RI, & Age	**Corrected Model**	**9.128**	**4**	**1867**	**<.001**
	ISI	0.091	1	1867	0.763
	RI	15.135	1	1867	<.001
	Age	13.113	1	1867	<.001
	ISI * RI	12.186	1	1867	<.001
MODEL 4: ISI, RI, & Proficiency	**Corrected Model**	**9.540**	**7**	**1640**	**<.001**
	ISI	0.02	1	1640	0.886
	RI	14.278	1	1640	<.001
	Proficiency	11.749	2	1640	<.001
	ISI * Proficiency	7.556	2	1640	0.001
	ISI * RI	20.328	1	1640	<.001
MODEL 5: ISI, RI, & Time on Task	**Corrected Model**	**10.134**	**7**	**1640**	**<.001**
	ISI	0.237	1	1640	0.627
	RI	7.973	1	1640	0.005
	Time on Task	16.8	1	1640	<.001
	ISI * RI	15.6	1	1640	<.001
	ISI * Time on Task	13.144	1	1640	<.001
	ISI * RI * Time on Task	13.672	2	1640	<.001

Cite this article: Serfaty, J. and Serrano, R. (2022). Lag effects in grammar learning: A desirable difficulties perspective. *Applied Psycholinguistics* **43**, 513–550. https://doi.org/10.1017/S0142716421000631