UNIVERSITAT DE BARCELONA

The effects of distribution, difficulty, and quantity of digital flashcard practice on language learning

Jonathan Serfaty

Doctoral Thesis

The effects of distribution, difficulty, and quantity of digital flashcard practice on language learning.

Jonathan Serfaty

Supervised by Dr. Raquel Serrano

UNIVERSITAT DE BARCELONA

The effects of distribution, difficulty, and quantity of digital flashcard practice on language learning.

Tesi presentada per optar al grau de doctor per la

Universitat de Barcelona

Programa de doctorat: Ciència Cognitiva i Llenguatge
Linea de recerca: Lingüística teòrica i aplicada

Jonathan Serfaty

Directora: Dra. Raquel Serrano
Tutora: Dra. Raquel Serrano

Dep. Llengües i Literatures Modernes i Estudis Anglesos
Facultat de Filologia i Comunicació
Universitat de Barcelona

UNIVERSITAT DE BARCELONA

September 2022

Abstract

According to research into second language (L2) practice, learners should repeatedly engage in output activities with feedback in order to develop accuracy. Digital flashcards can be used to prompt L2 output and deliver feedback when teacher instruction or peer interaction are not available. Traditionally used for vocabulary, this tool could also be used for grammar learning by providing exemplars of target structures. The difficulty, distribution, and quantity of practice have been shown to affect learning and retention in other domains, but little is known about how these variables affect productive L2 grammar practice. Research into these areas could deepen our understanding of L2 learning processes while providing useful guides for enhancing L2 practice. This thesis includes four papers. Chapter 1 is an introduction to the topic, covering the importance and theories of L2 practice, a primer on digital flashcards, and the research gaps to be addressed. Chapters 2-5 are research papers. The first paper, published in *System* as Serfaty and Serrano (2020), investigated how flashcards could be used for grammar learning in an environment where other forms of learning are unavailable. Simple structures were studied by 31 low-proficiency learners, aged 9-17, in a rural setting in Cambodia over two weeks. They were tested after one day, two weeks, and four months. Results showed that participants made large gains in their grammatical accuracy and maintained these gains over time. Scores were equivalent for trained and untrained items, showing that the exemplars used in training provided rules that were generalized to novel sentences. The second paper, published in *Applied Psycholinguistics* as Serfaty and Serrano (2022), investigated how this type of learning might be affected by the distribution of sessions. Two complex structures were studied at intervals of one day or one week, tested after one week or one month. Participants (N = 117) came from an international school in Phnom Penh, aged 10-18. The optimal lag was predicted by individual differences. Participants with slower times

and lower proficiency obtained higher scores from the shorter lag, whereas faster and more proficient learners benefited from a longer lag. Neither lag was better overall. The third paper, currently in review in *Language Learning and Technology*, repeated this methodology, using the same intervals, tools, and setting, with vocabulary items. Of the 96 participants analyzed in this study, 77 were also in the grammar analysis. This allowed for a comparison between grammar and vocabulary lag effects. This paper also aimed to ascertain whether a lag effect is found outside of lab conditions with secondary school students, which has not been found previously, and to explore whether lag effects are different for productive (form-recall) and receptive (meaning-recall) knowledge of learned words. Results showed a small but consistent advantage to the longer lag at both testing times, in contrast to grammar. The longer lag was particularly effective for retaining receptive knowledge at the 28-day posttest. The final paper, currently resubmitted to *Language Learning* after revisions, aimed to find the optimal amount of practice for the long-term retention of grammar knowledge. An artificial language was learned and then practiced on either one, two, three, or four relearning sessions on consecutive days, with 30 participants per condition (N = 129), aged 18-30. At a two-week posttest, it was found that average scores were significantly higher after a third relearning session. Accuracy during training peaked after the second relearning session, leading to a hypothesis that a threshold of knowledge is crossed after performing two sessions without errors. When re-coding the participants accordingly, this threshold was a stronger predictor of high posttest scores than the total number of sessions. The final chapter summarizes the findings from these papers, details implications from these findings on future theory, research methods, and pedagogy, and ends with some suggested future directions.

Resumen

Los estudios sobre el efecto de la *práctica* en una segunda lengua (L2) sugieren que es importante que los estudiantes participen repetidamente en actividades de producción en las que puedan recibir *feedback* o retroalimentación. Las tarjetas digitales (*digital flashcards*) se pueden usar con tal fin cuando el contexto de aprendizaje no facilite la recepción de *feedback* por parte de los estudiantes, ya sea por carecer de profesores formados, o por la imposibilidad de realizar actividades de producción en clase. Las *flashcards* tradicionalmente se han utilizado para el aprendizaje de vocabulario, sin embargo, también pueden utilizarse para el aprendizaje de la gramática. Se ha demostrado que la dificultad, la distribución temporal y la cantidad de la *práctica* afectan el aprendizaje y la retención de contenido en una L2, pero se sabe poco sobre cómo estas variables afectan la práctica productiva de la gramática. La investigación sobre estos temas podría facilitar nuestra comprensión de los procesos de aprendizaje de L2 al tiempo que proporcionar una guía útil para profesores y estudiantes sobre cómo mejorar la práctica de L Esta tesis incluye cuatro artículos. El Capítulo 1 es una introducción que presenta los temas claves investigados en la presente tesis. Los capítulos 2 a 5 incluyen las publicaciones. El primer artículo, publicado en *System* como Serfaty y Serrano (2020), investigó cómo las *flashcards* podrían usarse para el aprendizaje de la gramática en un entorno donde no hay otras formas de aprendizaje disponibles. Los resultados mostraron que los estudiantes lograron grandes avances en su corrección gramatical y mantuvieron estos avances a lo largo del tiempo. El segundo artículo, publicado en *Applied Psycholinguistics* como Serfaty y Serrano (2022), investigó cómo este tipo de aprendizaje podría verse afectado por la distribución temporal de las sesiones. Los estudiantes aprendieron dos estructuras complejas a intervalos de un día o una semana, y realizaron una prueba una

semana o un mes después. Los resultados indicaron que el intervalo óptimo depende de lasdiferencias individuales de los estudiantes. El tercer artículo, actualmente en revisión en *Language Learning and Technology*, repitió esta metodología, utilizando los mismos intervalos, herramientas y entorno, pero con vocabulario en lugar de gramática como objetivo de aprendizaje. Los resultados mostraron una pequeña pero consistente ventaja del intervalo más largo, en contraste con los resultados obtenidos para la gramática. El último estudio, actualmente reenviado a *Language Learning* después de revisiones, tenía como objetivo encontrar la cantidad óptima de práctica para la retención a largo plazo del conocimiento de la gramática. Los participantes aprendieron una lengua artificial y luego la practicaron en una, dos, tres o cuatro sesiones de "reaprendizaje" en días consecutivos. En una prueba posterior dos semanas después, se encontró que las puntuaciones eran significativamente más altas después de una tercera sesión de reaprendizaje. Después de analizar los resultados a nivel individual, se observó que los participantes mejoraban de forma significativa su corrección en la producción gramatical después de realizar dos sesiones sin errores, lo cual ofrece una guía para la práctica de gramática en L El capítulo final de la tesis resume e interpreta los resultados obtenidos en los cuatro estudios, detalla sus implicaciones para las teorías de aprendizaje de L2, métodos de investigación y enseñanza, para concluir con la presentación de algunas direcciones para futuros estudios.

Contents

List of Figures

List of Tables

Full List of Abbreviations

CALL	Computer assisted language learning
DDF	Desirable Difficulty Framework
ELL	English language learning
GJT	Grammaticality judgment test
IRR	Incidence rate ratio
ISI	Intersession interval
L1	First language
L2	Target language
MTS	Minimum-trials session
OR	Odds ratio
RI	Retention interval
RQ	Research question
RS	Relearning session
S	Session (+ number)
SAT	Skill Acquisition Theory
SLA	Second language acquisition
SRT	Skill Retention Theory
TS	Training session

Acknowledgments

First and foremost, this thesis would not be possible without the immense and consistent support from my doctoral supervisor, Raquel Serrano. During the first year of the MA course, I told Raquel about my idea to use digital flashcards for grammar and to implement this in a remote village in Cambodia. Raquel was immediately supportive of the idea and suggested that time distribution might be an interesting avenue to pursue. Through her support, I was authorized to pursue my research remotely, before remote learning became the norm. The next 3-4 years were characterized by regular emails, video calls and copious comments on Google Docs. Each time I asked for feedback, Raquel invested her time and energy into the task. She prompted me to question my ideas, from theoretical arguments to ambiguous word choices, guiding me to change and improve each paper through countless drafts. I could not have asked for a more committed, approachable, and patient guide through this process.

Secondly, I would like to thank my follow-up committee members Tatsuya Nakata, John Rogers, and Yuichi Suzuki. It was a memorable experience to have three of the biggest experts in my topic discussing my ideas for nearly two hours. Each one of them offered valuable suggestions that were eventually incorporated into the thesis plan. In addition, over the course of the PhD, they also provided encouraging feedback and helpful resources, which I was very grateful to receive. I would also like to thank members of the GRAL research group, who allowed me to present my studies and offered feedback. In particular, I would like to mention my MA professors, Joan-Carles Mora, Sara Feijoo, Maria Luz Celaya, Julia Baron, Elsa Tragant and Imma Miralpeix. Each one of them indirectly contributed to my thesis by teaching me, in the order just listed, how to use SPSS, how to conduct valid research, how cross-linguistic differences affect the acquisition of certain structures, research tools, how to conduct classroom research, and issues in vocabulary research. A special thank you goes to Roger Gilabert, who introduced me to the theoretical frameworks that this thesis is based on, Carme Muñoz, whose course on age and individual differences has allowed me to make policy changes in the schools I currently manage, and to María Andriá, whose CLIL course is the foundation of the curriculum I am currently implementing.

Further support was obtained from Oliver Valero Coppin, who answered all my questions on statistical models. Without this guidance, I would not have been able to analyze my data. The staff at Gorilla also deserve acknowledgment. Over the course of

many weeks, I developed the coding that would be used in Chapter 5. I had a lot of requirements that had never been made before, and the kind staff at Gorilla worked to find work-arounds and coding tricks to meet my needs, answering my emails on a daily basis. This study could also not have been achieved without the Language Learning Dissertation Grant, and so I must acknowledge the committee that awarded me this prestigious grant and deemed this research worth pursuing. Of course, I must also thank the anonymous reviewers and editors that contributed to my publications and allowed them to be published.

On the ground, many people participated in these studies and without them, there would be no data to work with. The students of Green Village School in Kampong Cham chose to come for extra schooling every day of the week in order to improve their own futures, even without teachers. When offered the chance to learn English through smartphones, they not only agreed but even queued up to be next in line. I must also thank the volunteers that provided their phones for the experiment and helped to organize the students. Special thanks to Diogo, Marta, and Paolo for helping me to implement this study. The next thanks go to my own students, who participated enthusiastically in the Quizlet studies. As a result of these studies, many students continued to use Quizlet for their own exam revision in all subjects and were able to counteract the negative effects of distance learning. Additional thanks to Matthews Abuka for supervising some of the experiments, and to Sarah Thompson for supporting the process. Also, a huge thanks to the 30-40 people who volunteered to learn a language I invented, which became the pilot for Chapter 5. The results are coming soon, I promise! Next, I must thank all of my colleagues at Cambodian Children's Fund, from management to teachers, and of course all the students, for allowing me to implement digital flashcards into the curriculum. Over the past year, we have seen a huge improvement in students' learning abilities, grammar knowledge, and ICT skills. A special thanks to the many people who helped me to translate English flashcards into Khmer, which is by no means a simple task.

A final thank you goes to my friends, for putting up with me during this intense period of my life. I have no more excuses to stay home now, so let's celebrate!

Chapter 1: Introduction

Second language (L2) practice has been defined by DeKeyser (2007) as "specific activities in the second language, engaged in systematically, deliberately, with the goal of developing knowledge of and skills in the second language" (p. 8). The perception of L2 practice has suffered in recent decades from an association with early approaches to language learning (DeKeyser, 2010). In particular, the audiolingual approach was infamous for only allowing students to memorize, imitate or manipulate supplied target phrases and dialogues without progression to genuine interactions. In this approach, language was taught through mechanical habit formation, with no place for the analysis of underlying rules or the generation of original or communicative language (Richards & Rodgers, 2001). On the other end of the spectrum, Krashen (1985) insisted that language can *only* be acquired through input, subconsciously, following a natural sequence. Krashen has continued to be highly critical of claims that productive practice has any positive effect on language acquisition (e.g., Krashen, 1998). This is despite a body of evidence from immersion contexts (Swain, 1998) showing that meaningful input alone is not sufficient for acquiring accurate L2 use.

More recent approaches have prioritized productive and interactive L2 practice. In task-based learning, the emphasis is on communicating messages in order to complete tasks, with less emphasis on grammatical accuracy (Long, 1996). In this approach, grammar is taught inductively as a task goal rather than for its own sake. Thus, although it is now widely accepted that language

learners need to produce the language in order to learn it, the value of de-contextualised and non-communicative practice remains controversial.

This thesis will mainly, but not exclusively, deal with grammar practice. For clarity, grammar will be defined as the assembly of linguistic parts according to morphosyntactic norms that communicate an intended message. These norms are based on how a language is used by a target language community. For example, knowing to use “would have” rather than “would has” is important because this is how the language is currently used and understood by proficient English speakers. In essence, the term grammar is used in opposition to vocabulary. If vocabulary items are the building blocks, then the act of assembling and manipulating these blocks to form meaning is grammar. Note that this definition of grammar does not include knowledge of grammar terminology, but rather the effective use of the L2.

Grammar is especially difficult to master when one’s first language (L1) does not encode a particular feature that is required in the L2 (Crosthwaite, 2016; Öksüz et al., 2021; Schepends et al., 2020). This presents a particular challenge to speakers of non-European L1s from entering fields that require proficient English, which is the case for many employment opportunities and certainly for much of academia. It is therefore vital for learners to acquire some level of grammatical accuracy in their L2.

Despite some negative perceptions, it is clear that some form of practice is desirable in L2 learning. This chapter will present some factors that contribute to effective practice, introduce a tool for investigating these factors,

identify some specific research gaps, and finish with an outline of the studies conducted to address these gaps.

1.1 Theories of L2 Practice

While general rules of grammaticality could be learned as static factual knowledge, this thesis defines grammar as the skill of converting intended meanings into comprehensible forms without ambiguity. The difference between a factual knowledge of rules and a skill is that the latter can be performed with variable proficiency and success, and can be improved through practice.

The type of practice promoted by the audiolingual approach was imitation, which requires no effort or attention to underlying patterns. This could be useful for quickly acquiring useful phrases in an unknown language, but cannot lead to transferable knowledge. For example, memorizing "weər ɪz ðə ˈbɑːθru(ː)m" (*Where is the bathroom?*) does not enable the user to ask where other places might be. They would need to know which syllables express the question and which syllables refer to the location. Whereas teachers of the audiolingual method were instructed to praise learners after a successful imitation in order to subconsciously reinforce acceptable utterances (Richards & Rodgers, 2001), transferable practice should involve consciousness-raising feedback to promote deeper understanding of the underlying rules of the target language (Gass & Mackey, 2006).

The expression of language formulated from one's own mind, as opposed to reciting or imitating, is known as output. Swain's Output Hypothesis

(1985, 1988, 1995), formulated in direct opposition to Krashen's input-only approach, highlights three key roles of output. The first is noticing, which is the idea that learners must be consciously aware of a form or rule in order to learn it. By producing a form in output, the speaker processes it more deeply than they would through passive comprehension. If, however, the learner is not able to produce the required form, this failure primes them to notice the form in subsequent input. The next function of output is hypothesis testing. By producing the L2, the learner is testing their current understanding of the target language. The final function of output is metalinguistic reflection, which is the learner's ability to reflect on feedback from their output in order to confirm their hypothesis or be prompted to change their hypothesis. These aspects of language learning have been integrated into many competing models of instructed second language acquisition (Leow, 2015). As the learner produces more language, receives more feedback, and restructures their internal rules for the L2, they progressively improve their skill proficiency.

This gradual qualitative improvement is described by Skill Acquisition Theory (SAT; DeKeyser, 2020), which is a key theoretical framework in this thesis. According to SAT, learners first acquire declarative knowledge, which is the knowledge of what they are supposed to do (e.g., knowledge that regular verbs in English are expressed by adding *-ed*). This could be achieved through direct instruction from a teacher, for instance. As the learner begins to implement these rules, procedural knowledge also develops. Procedural knowledge consists of internal production rules (e.g., if "walk" + past, add *-ed*) that can be implemented. Procedural knowledge, once sufficiently developed, allows for faster and less effortful performance. Through extensive practice, this

knowledge should eventually become automatised. Automatised knowledge can be performed instantaneously and effortlessly (Segalowitz & Segalowitz, 1993), resulting in fluent use of the L2. The process of proceduralization is traditionally measured by response times and error rates (e.g., DeKeyser, 1997; Ferman et al., 2009; Suzuki, 2017). Both of these measures improve sharply after some initial practice and then continue to improve gradually with further practice until no more improvement can be detected (Kim et al., 2013).

While the Output Hypothesis tells us that output is needed as a component of practice, and SAT deals with the improvement trajectory of skill proficiency through extensive practice, there is still the issue of *how* the L2 should be practiced. For this, insights can be taken from the Desirable Difficulties Framework (DDF; Bjork, 1994). According to this framework, activities that cause difficulty during practice may lead to more errors in the training phase, but can often lead to better long term retention and transferability of the practiced knowledge or skills (Bjork, 1999; Schmidt & Bjork, 1992). This contrasts starkly with the audiolingual method, which discouraged communication due to the risk of producing incorrect utterances. Under the DDF, errors are viewed as an opportunity to notice gaps in knowledge and process the target material on a deeper level. It would also be suboptimal for conditions to be too difficult for learners to successfully acquire the target knowledge. In these cases, difficulty should be reduced.

Suzuki et al. (2019) adapted the DDF to L2 practice, suggesting that the desirable level of difficulty for a practice condition depends on the inherent difficulty in the feature being learned and the subjective difficulty for the

learner. One of the simplest ways to manipulate difficulty is to change the amount of time between practice sessions. More time would lead to forgetting, allowing learners to identify knowledge that has not yet been mastered, whereas less time would allow learners to practice more difficult skills without needing to relearn key concepts.

In sum, output practice serves several purposes in L2 development. Through output, learners notice and acquire target forms, which become more accessible through repeated practice. The best type of practice should be challenging in order to induce deep processing.

1.2 Digital Flashcards as a Research Tool

As stated above, practice should be productive, intensive, and challenging. Authentic communicative tasks would be ideal because practice is most effective when it matches the end goals (Lyster & Sato, 2013). However, there are a number of issues with relying on this type of practice. Firstly, most language learners are not exposed to situations in which they can authentically communicate in the L2. Much of the world's language learning takes place in a foreign language classroom in a secondary school, with English alone being a compulsory subject in 142 countries (Ives, 2022). Communicative practice can be manufactured through well-designed tasks, but planning and implementing tasks effectively requires a high level of expertise (Van den Branden, 2016) that many educators simply do not possess. Therefore, in many cases, authentic communicative practice is not a viable option. Secondly, although interaction can elicit useful feedback (Gass & Mackey, 2006), it is not guaranteed. In some

cases, the interlocutor may wish to be polite or may understand from context rather than from language, or may not be aware of the error. In these cases, the learner might not receive feedback, which is necessary to prompt metalinguistic reflection (Long, 1996). Thirdly, it can be difficult to design a task in which a specific structure is guaranteed to be needed extensively and learners might avoid it altogether. Therefore, in cases where a specific problematic feature is targeted, or if a teacher wishes to expand their students' linguistic repertoire, practice may need to be more systematic, artificially repetitive, and with the guaranteed provision of feedback.

Digital flashcards are applications that are designed for independent practice of specific items. They are commonly used for learning vocabulary pairs or for memorizing facts before exams (Zung et al., 2022). They are based on paper flashcards, where the front has the question or cue, and the back has the target, allowing the user to test themselves. This form of practice is known as retrieval. The learner repeatedly attempts to retrieve information from memory in order to strengthen their access to that knowledge.

The digital versions of flashcards include a host of useful features for the learner (Ashcroft et al., 2018; Nakata, 2011; Zung et al., 2022). For example, software will automatically remove known items, allowing the learner to focus on unknown items. With a criterion of one, the user must answer every item correctly once before it is removed from the cycle. The criterion could also be increased. For example, Cram.com has a mode in which five correct responses of each item are required before a session ends. Increasing the criterion has been shown to increase posttest scores in the short term (Rawson

& Dunlosky, 2011) and could be useful for “cramming” before an exam. Most apps also record the learner's progress so that users can choose to practice items that were difficult in a previous session. Apps are also accessible from anywhere, allowing learners to take advantage of time in transit, waiting for an appointment, or during a break from work. Sets can be synced to be available offline, making them suitable for environments without an internet connection. Flashcards can be created in any written language and accompanied by audio in many widely-spoken languages.

Although flashcards have traditionally been used in L2 learning for vocabulary, the papers in this thesis will show that flashcards can also be used for grammar practice. Rather than retrieving single items, learners can practice formulating full sentences in the L2. They see a cue, which could be the L1 translation or an L2 scenario, and they type the L2 sentence (see Figure 1). Upon doing so, they see the target response, presented along with their attempt (Figure 2). This leads them to notice similarities and differences between the sentences and either hypothesize rules (inductive learning) or be reminded of rules that have already been taught (deductive learning). The next flashcard allows them to apply these rules in a new sentence that requires the same knowledge (Figure 3). Errors could be made around the target feature, but errors could relate to any aspect of the sentence, from the orthography of content words to a missing ‘s’ (Figure 4). Through this mechanism, they not only practice the target feature but also work on other linguistic features and general accuracy.

FIGURE 1

The learner tests their hypothesis

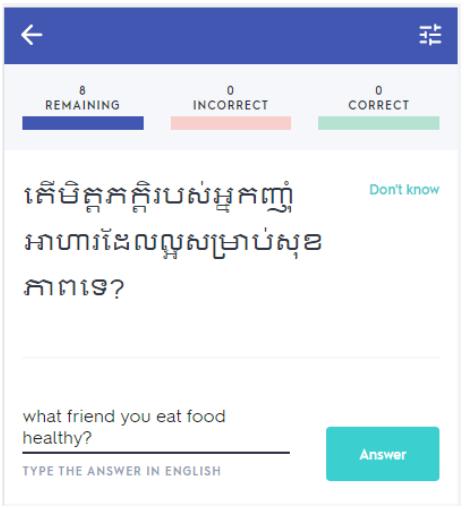


FIGURE 2

The learner sees feedback

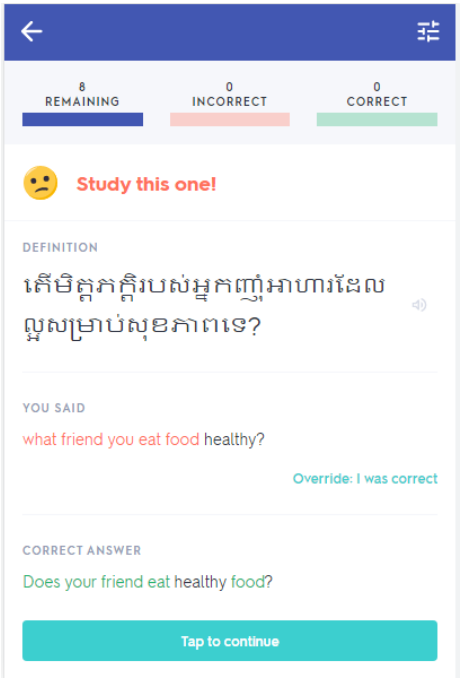


FIGURE 3

The learner overgeneralizes that all questions start with 'Does' and must adjust their hypothesis again. With fewer errors on screen, they notice that 'one' should be 'an'

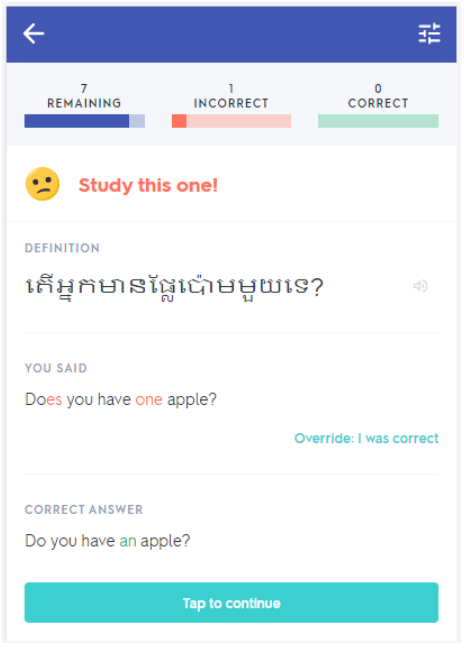
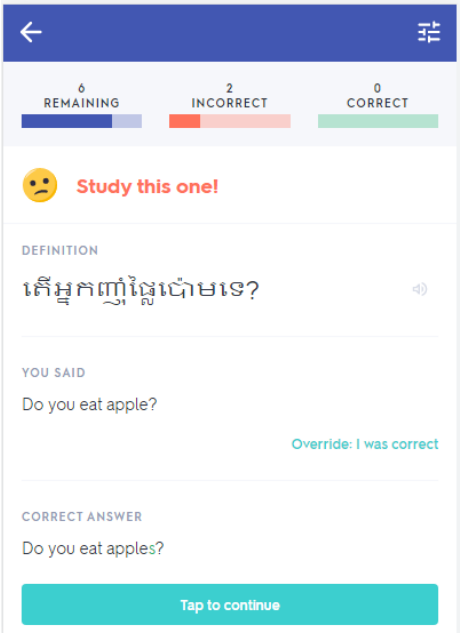


FIGURE 4

The learner successfully produces the target structure, but must now learn to add an 's'.



Digital flashcards were adopted as the focal tool for this thesis. The motivation for this was threefold. Firstly, from a research perspective, this tool can be used to control the nature of L2 practice in a precise manner. The exact input and feedback is predetermined, no instructor factors or peer-interaction factors are involved, and the exact amount of accurate output of each learner can be manipulated. These attributes make it easy to isolate specific practice variables while holding many other potentially confounding variables constant. Secondly, from a theoretical perspective, flashcards involve input, output, feedback, hypothesis testing, and extensive repetition, all of which have been highlighted as necessary ingredients for L2 practice. Sets of flashcards may be designed with a specific target feature in mind, but the participant must consider the accuracy of the entire sentence in order to remove an item. This contextualizes the target feature within general accurate language production, which is more desirably difficult and more like authentic language use than typical grammar exercises. For example, in a gap-filling exercise, the target feature is isolated while the learner can ignore the rest of the sentence. Thirdly, from a pedagogical viewpoint, flashcard apps are used by teachers in real L2 classrooms. No time limit is imposed on the input or output stages and there is no limit on the number of attempts for the participant, as in genuine L2 learning conditions. This makes research from digital flashcards ecologically valid and easily applicable for practitioners.

It should be noted that the goal of this thesis is not to advocate for the use of digital flashcard practice as an ideal or exclusive form of L2 practice. The purpose of using flashcards is only for learners to acquire knowledge of accurate forms and to increase their access to these forms. By prompting the

learner to independently induce rules for grammar or develop mnemonic strategies for vocabulary, the teacher avoids "explaining" the language metalinguistically, which can be confusing and ineffective for learners with low metalinguistic awareness, and avoids presenting lists of vocabulary items. In an ideal classroom setting with a highly competent teacher, the goal of flashcards would be to give every student the practice and time that they require outside of the classroom. On the one hand, flashcards could be assigned ahead of an upcoming activity. Teachers could save a lot of time in teaching a specific feature if the students have already undergone flashcard training at home. In doing so, the teacher would not need to focus on teaching rules, monitoring accuracy, or giving feedback. Instead, they could use their time to facilitate interactive activities in the L2 that focus on meaning. The other use of flashcards would be to collect and practice previously learned features. Language courses tend to be modular and may not repeat key vocabulary or grammar points with enough frequency for them to be well retained (Tschichold, 2012). By creating flashcard sets, the student can remind themselves of everything they have previously learned and still access it when desired in the future. In less ideal contexts, where a student must learn a language without access to proficient teachers, accurate textbooks, or online lessons, flashcard apps can provide an offline and accessible means for practicing the production of accurate L2 sentences.

This chapter has thus far sought to establish that L2 practice is worth investigating and that digital flashcards are a suitable tool for doing so. I will now specify the particular areas of L2 practice that will be addressed in this thesis.

1.3 Research Gaps

The first research gap that needs to be filled is the absence of studies on how written practice affects accuracy development at the sentence level. Repeated practice has been investigated for the development of accuracy in oral output (McDonough & Sato, 2019; Sato & McDonough, 2019) and for fluency without considering accuracy (de Jong & Perfetti, 2011; Suzuki, 2021; Suzuki & Hanzawa, 2021). However, many learners do not have reliable knowledge of their target language, either from lack of resources or from failing to notice L2 forms. Oral practice of erroneous language may lead to its fossilization in the learner's interlanguage (Han, 2012). Written practice, with feedback presented next to a learner's attempt, is especially helpful for noticing forms (Zalbidea, 2021). Moreover, oral practice is unrealistic for many learners that do not encounter the target language outside of their classroom or because they are studying autonomously. For all of these reasons, an investigation into the potential benefits of repeated written output is warranted.

The second gap is in research among underprivileged populations. Research on L2 practice tends to be carried out in wealthy countries among participants with access to educational resources (Collins & Muñoz, 2016). Very little research has been conducted for populations who lack access to formal education. This is surprising in that these populations have the greatest need for intervention. Learning English can be the ticket out of a cycle of poverty in many developing countries, opening doors to education and employment (Haidar, 2019; Hamid, 2016). The vast majority of techniques

found in the literature are simply not applicable to a large portion of language learners, who do not have access to reliable resources or proficient teachers.

The third gap relates to time distribution. When comparing longer or shorter distributions of study sessions, mixed results have been obtained. The types of target knowledge, learning activities, tests, and participants have varied too much for any clear conclusions to be made. It is clear that altering the distribution of practice has an effect, but the direction of this effect has varied depending on the study (Edmonds et al., 2021). It is necessary to perform more controlled and replicable research into time distribution, controlling for as many variables as possible. One specific question is to what extent distribution effects found in verbal learning studies from cognitive psychology (e.g., Cepeda et al., 2009) would apply to the skill of L2 grammar. A related question is whether distribution effects could be utilized for classroom learning. Lastly, little is known about how distribution effects may differ according to the type of knowledge under examination (grammar vs vocabulary; productive vs receptive).

The final and possibly most important gap concerns the quantity of practice. Previous research into multi-day learning has varied in how many sessions were involved. Conclusions have been made by comparing posttest scores between conditions with very little consideration to the overall scores. In reality, language learners do not engage in practice in order to obtain temporary knowledge or higher knowledge than someone from a different practice condition. They want to learn their target fully, and to remember it in the future. At the moment, we cannot advise a learner on how much practice is

needed before their knowledge is immune to forgetting, or how often it must be reviewed.

1.4 Thesis Outline and Overview

The present thesis uses digital flashcards to investigate L2 learning and practice. Three of the four studies target L2 grammar, as defined above. The flashcards are designed to target specific structures, but in all cases the entire sentence must be grammatical in order for an item to be removed from a practice set and learners are never instructed to focus on a specific feature. They are simply asked to translate a sentence into English, or formulate the sentence based on a given scenario. Four studies are presented.

Study 1 (Serfaty & Serrano, 2020) can be viewed as a proof of concept for using digital flashcards to learn grammar, which had not been researched previously. It also fills the first two research gaps by investigating the repeated practice of written language formulation and taking place among a resource-poor rural community trying to learn English without access to a teacher or other materials. This study included 31 participants with no previous exposure to native speakers or authentic language sources. They had already been learning English daily for a year from a revolving door of volunteers in an improvised school. The goal of the school was to supplement their public schooling, which in Cambodia only operates for half the day. In this time, the participants had acquired an impressive vocabulary but were failing to form simple sentences with recognisable grammaticality. Flashcard training was employed, using a range of target structures, including the present

simple, present continuous, there is/are, and their interrogative forms. Each structure had five exemplars, practiced three times over the course of eight days. The study examined two aspects in particular - the transferability of learned items to novel items and long-term retention. Tests required participants to translate full sentences from their L1 (Khmer) into English, for both trained and untrained items, using smartphones. The tests took place one day, two weeks, and 18 weeks after the final training session. This study established a baseline of success for the digital flashcard method.

Study 2 (Serfaty & Serrano, 2022) addressed the third research gap by investigating how time distribution affects retention from this type of grammar practice. Drawing on the Desirable Difficulty Framework for L2 proposed by Suzuki et al. (2019), this study was designed to primarily investigate the effects of using a longer or shorter intersession interval. Moreover, the study sought to determine which factors might influence this effect. Specifically, the study compared two grammatical structures, two age groupings, two ability levels (measured by time on task), and three levels of L2 proficiency. The participants were from an international school for wealthy families in Phnom Penh, with a much more privileged educational background and much higher English proficiency. Structures were typed from L2 scenarios, rather than L1 translations, with intervals of either one day or one week, tested after either one week or one month.

Further exploring time distribution, Study 3 (Serfaty & Serrano, in review) repeated the same methodology as Study 2 but using vocabulary items, with the aim of comparing lag effects between grammar and vocabulary

learning without task differences or participant factors. The study included 96 students with an overlap of 77 participants from the grammar experiment. Previous classroom studies had not reported better results from a longer intersession interval. However, these studies had not examined vocabulary retrieval training, which is the type of learning that has exhibited this effect under lab conditions. This study also included both productive and receptive tests, since these two dimensions of knowledge had never before been compared after different lags.

Finally, Study 4 (Serfaty & Serrano, resubmitted) addressed the final research gap relating to quantity of practice. It was observed that different scores and retention levels were achieved in the previous grammar studies, which had been different in terms of the quantity of practice. This study compared learning grammar on two, three, four, or five consecutive days. An experiment was developed on Gorilla to simulate digital flashcards. Participants were recruited online through Prolific, aged 18-30 and with a range of linguistic and geographical backgrounds. An artificial language was developed in order to avoid any prior knowledge. Their achievement at the training stage was measured by the number of trials required to complete a session. Training performance was then compared to posttest performance in order to form a new hypothesis about how to predict high posttest scores at the training stage. This hypothesis was tested by regrouping participants according to their training performance and modeling their posttest scores accordingly.

The final chapter provides a summary of the preceding chapters and discusses their implications for theories of L2 practice, future research

methods, and pedagogy. In light of the findings, and the questions raised, ideas for future research are proposed.

In sum, this thesis seeks to present new insights into the optimal distribution, difficulty, and quantity of L2 practice. The evidence gained from these studies will be used to expand upon the Desirable Difficulties Framework and Skill Acquisition Theory. The major pedagogical implications of this research will relate to the type of learning that could be expected from digital flashcards, the optimal scheduling of sessions, and the amount of practice required for each student to achieve their goals.

References

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

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.

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

Collins, L., & Muñoz, C. (2016). The foreign language classroom: Current perspectives and future considerations. *The Modern Language Journal*, *100*(S1), 133–147. https://doi.org/10.1111/modl.12305

Crosthwaite, P. (2016). L2 English article use by L1 speakers of article-less languages. *International Journal of Learner Corpus Research*, *2*(1), 68–100. https://doi.org/10.1075/ijlcr.1.03cro

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533–568. https://doi.org/10.1111/j.1467-9922.2010.00620.x

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*(2), 195–221. https://doi.org/10.1017/S0272263197002040

DeKeyser, R. (2007). Introduction: Situating the concept of practice. In *Practice in a Second Language* (pp. 1–18). Cambridge University Press. https://doi.org/10.1017/CBO9780511667275.002

DeKeyser, R. (2010). Practice for second language learning: Don't throw out the baby with the bathwater. *International Journal of English Studies*, *10*(1), 155–165. https://doi.org/10.6018/IJES/2010/1/114021

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

Edmonds, A., Gerbier, E., Palasis, K., & Whyte, S. (2021). Understanding the distributed practice effect and its relevance for the teaching and learning of L2 vocabulary. *Lexis*, *18*. https://doi.org/10.4000/lexis.5652

Ferman, S., Olshtain, E., Schechtman, E., & Karni, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics*, *22*(4), 384–412. https://doi.org/10.1016/j.jneuroling.2008.12.002

Gass, S. M., & Mackey, A. (2006). Input, Interaction and Output. *AILA Review*, *19*, 3–17. https://doi.org/10.1075/aila.19.03gas

Haidar, S. (2019). The role of English in developing countries. *English Today*, *35*(3), 42–48. https://doi.org/10.1017/S0266078418000469

Hamid, M. O. (2016). The linguistic market for English in Bangladesh. *Current Issues in Language Planning*, *17*(1), 36–55. https://doi.org/10.1080/14664208.2016.1105909

Han, Z. (2012). Fossilization. In C. Chappelle (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford, UK: Wiley Blackwell. https://doi.org/10.1002/9781405198431.wbeal0436

Ives, P. (2022). Countries in which English Language is a Mandatory or an Optional Subject (interactive). Retrieved from https://www.uwinnipeg.ca/global-english-education/countries-in-which-english-is-mandatory-or-optional-subject.html

Kim, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science*, *14*(1), 22-37, https://doi.org/10.1080/1464536X.2011.573008

Krashen, S. (1985). The Input Hypothesis: Issues and implications. New York, NY: Longman.

Krashen, S. (1998). Comprehensible output? *System*, *26*(2), 175–182. https://doi.org/10.1016/S0346-251X(98)00002-5

Leow, R. P. (2015). Explicit learning in the L2 classroom: A student-centered approach. New York: Routledge. https://doi.org/10.4324/9781315887074

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of Second Language Acquisition* (pp. 413–468). San Diego: Academic Press.

Lyster, R., & Sato, M. (2013). Skill Acquisition Theory and the role of practice in L2 development. In M. del Pilar García Mayo, M. Juncal Gutiérrez Mangado, & M. Martínez-Adrián (Eds.), *Contemporary Approaches to Second Language Acquisition* (pp. 71–92). John Benjamins. https://doi.org/10.1075/AALS.9.07CH4

McDonough, K., & Sato, M. (2019). Promoting EFL students' accuracy and fluency through interactive practice activities. *Studies in Second Language Learning and Teaching*, *9*(2), 379–395. https://doi.org/10.14746/ssllt.2019.9.2.6

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*(1), 17–38. https://doi.org/10.1080/09588221.2010.520675

Öksüz, D. C., Derkach, K., & Alexopoulou, T. (2021). L1 and L2 typological distance effects on the learnability of articles in L2 English: A large-scale learner corpus analysis. *In Architectures and Mechanisms for Language Processing*. Paris. Retrieved from https://amlap2021.github.io/program/106.pdf

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302. https://doi.org/10.1037/A0023956

Richards, J. C., & Rodgers, T. S. (2001). Approaches and methods in language teaching. Cambridge University Press. https://doi.org/10.1017/CBO9780511667305

Sato, M., & McDonough, K. (2019). Practice is important but how about its quality? Contextualized practice in the classroom. *Studies in Second Language Acquisition*, *41*(5), 999–1026. https://doi.org/10.1017/S0272263119000159

Schepens, J., van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, *194*, 104056. https://doi.org/10.1016/j.cognition.2019.104056

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

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*(3), 369–385. https://doi.org/10.1017/S0142716400010845

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

Serfaty, J., & Serrano, R. (2022). Lag effects in grammar learning: A desirable difficulties perspective. *Applied Psycholinguistics*, *43*(3), 513–550. https://doi.org/10.1017/S0142716421000631

Serfaty, J., & Serrano, R. (in review). The optimal scheduling of Quizlet sessions for L2 vocabulary learning.

Serfaty, J., & Serrano, R. (resubmitted). Practice makes perfect, but how much is necessary? The role of relearning in L2 grammar acquisition.

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. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*, *71*(2), 285–325. https://doi.org/10.1111/lang.12433

Suzuki, Y., & Hanzawa, K. (2021). Massed practice is a double-edged sword for fluency development: An EFL classroom study. *Studies in Second Language Acquisition*, 1–26. https://doi.org/10.1017/S0272263121000358

Suzuki, Y., Nakata, T., & DeKeyser, R. (2019). The Desirable Difficulty Framework as a Theoretical Foundation for Optimizing and Researching Second Language Practice. *The Modern Language Journal*, *103*(3), 713–720. https://doi.org/10.1111/modl.12585

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Swain, M. (1988). Manipulating and complementing content teaching to maximize second language learning. *TESL Canada Journal*, *6*(1), 68. https://doi.org/10.18806/tesl.v6i1.542

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford: Oxford University Press.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), Focus on form in classroom second language acquisition (pp. 64–81). New York: Cambridge University Press.

Tschichold, C. (2012). French vocabulary in Encore Tricolore : Do pupils have a chance? *The Language Learning Journal*, *40*(1), 7–19. https://doi.org/10.1080/09571736.2012.658219

Van den Branden, K. (2016). The role of teachers in task-based language education. *Annual Review of Applied Linguistics*, *36*, 164–181. https://doi.org/10.1017/S0267190515000070

Zalbidea, J. (2021). On the scope of output in SLA: Task modality, salience, L2 grammar noticing, and development. *Studies in Second Language Acquisition*, *43*(1), 50–82. https://doi.org/10.1017/S0272263120000261

Zung, I., Imundo, M. N., & Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? *Memory*, *30*(8), 1–19. https://doi.org/10.1080/09658211.2022.2058553

Chapter 2:

Examining the potential of digital flashcards to facilitate independent grammar learning

Published as:

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

System 94 (2020) 102342

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

Examining the potential of digital flashcards to facilitate independent grammar learning

Jonathan Serfaty [a], Raquel Serrano [b, *]

[a] Universitat de Barcelona, Spain
[b] Department of Modern Languages and Literatures and English Studies, Universitat de Barcelona, Gran Via de Les Corts Catalanes, 585 08007, Barcelona, Spain

ARTICLE INFO

Article history:
Received 19 March 2020
Received in revised form 3 July 2020
Accepted 11 August 2020
Available online 15 August 2020

Keywords:
Digital flashcards
Grammar learning
Output
Noticing
Underprivileged settings
Independent language learning

ABSTRACT

Digital flashcards are widely used and studied for vocabulary, but no previous research has examined this tool for grammar learning. This paper addresses this gap by asking whether full-sentence flashcard training could cause learners to notice and accurately use grammatical patterns. The participants ($N = 31$), school-aged independent English learners in rural Cambodia, underwent 8 sessions of typing translations from their first language, Khmer, to English using the smartphone app Cram.com Flashcards, with items disappearing only when answered without errors. Their performance on trained and untrained items was assessed before treatment, immediately after treatment, and 2 and 18 weeks after treatment. Before the final delayed post-test, one group ($N = 14$) underwent a single refresher session in order to observe its effects on retention. Results showed high gains for all participants (M = 82%) and minimal losses at the second delayed post-test. Equal gains between trained and untrained items demonstrated that participants had indeed inferred grammar rules from the training, and the refresher fully mitigated losses. Further research into digital flashcards for grammar is recommended, to ascertain which factors determine success.

1. Introduction

Grammatical accuracy not only aids in communication but is an important skill in the academic and professional contexts. Even teaching approaches which emphasize the importance of focusing on communicative skills in the second language (L2) class, such as task-based language teaching, also include a focus on grammatical forms when necessary (Bindileu, 2019; Robinson, 2011). Fortunately, grammar-focused technology has opened up a new realm of engaging learning activities. While the benefits of computer assisted language learning (CALL) are widely known, most software caters only for languages with many native speakers, such as English or Spanish. Language learners with less prominent first languages (L1s) do not have the same access to online tools. However, online flashcards can be created by the user in many more languages and distributed freely to those who cannot afford paid software. This makes digital flashcards a promising solution for under-represented language learners. Second language acquisition (SLA) research into flashcard training has, until now, solely focused on vocabulary learning. Findings have been largely positive (e.g. Andarab, 2017; Dizon, 2016; Nakata, 2020; Sanosi, 2018), despite

* Corresponding author.
E-mail address: raquelserrano@ub.edu (R. Serrano).

https://doi.org/10.1016/j.system.2020.102342
0346-251X/

the associations with behaviourism still prevalent among some scholars, but vocabulary is only one component of language. Especially for learners with typologically distant first and second languages, connecting words to communicate meaning in a more grammatically nuanced language can be a real challenge.

This longitudinal study will explore the possibility of using flashcards to improve grammatical accuracy by using exemplary sentences of grammatical patterns as flashcard items. Throughout this paper, the term "grammar" will denote generally accepted morphosyntactic norms, without implying any deeper metalinguistic knowledge. Participants are school-aged learners of English in a rural village in Cambodia, whose need to learn English is great, but who lack the resources to do so.

2. Literature review

2.1. CALL for grammar

Although the ultimate goal of second language (L2) learning should be automatic and fluent language use in meaningful situations, some researchers have claimed that a focus on language forms might be necessary in meaning-oriented L2 classroom contexts if the goal is to promote learners' accuracy as well as fluency (Lightbown, 2000; Spada, 1997). In an environment lacking in teachers to oversee controlled practice of forms, CALL provides a solution. Research into using CALL for grammar instruction (Abu Naba'h & Abdallah, 2012; McEnery, Baker, & Wilson, 1995; Mohamad, 2009; Nutta, 1998) has shown it to be as effective or even more effective than teacher-led instruction. For example, Cerezo, Caras, and Leow (2016) used the Spanish "gustar" structure to compare beginner English-speaking Spanish learners using a maze-style video game versus traditional instruction from a teacher. The game provided guided instruction designed to prompt reflection on forms, without explicitly teaching rules. Results of translation post-tests, written and oral, showed considerable learning in both groups, but with significantly higher gains for the CALL group (written: 83% vs. 63.2%; oral: 91.3% vs. 60.2%) and far higher retention on the two-week delayed post-tests (written: 72.6%, vs. 32.8%; oral: 81.6% vs. 39.7%). They concluded that CALL could replace teacher-led instruction and create more class time for communicative activities.

CALL has also been used for oral practice. Penning, Cucchiarini, Strik and Hout (2019) assessed the use of computerised corrective feedback on oral responses among 68 learners of Dutch from high, medium, and low education backgrounds. The software Greet showed users questions and required oral responses based on re-ordering given word blocks. One group received feedback on whether their response was correct, while the comparison group did not. The treatment was effective in both conditions for high and medium educated subjects, but in neither condition for those of low education.

It must be borne in mind that the approach to CALL in the studies above is different from digital flashcards, as they included games and oral responses, which probably contributed to students' motivation. Furthermore, these studies employed software designed specifically for certain rules in certain languages, which is not accessible to learners of low-resource environments with underrepresented L1s because software is simply not being produced for these languages, let alone in the form of free mobile apps. However, free, customizable digital flashcards apps represent a possible solution. Flashcards are, by design (see section 2.2), a tool to promote learning of isolated items, which is why they have been widely used for vocabulary (i.e., one flashcard-one word) and not for grammar rules. Nevertheless, as illustrated in the next section, the principles that apply to vocabulary learning through flashcards could also be expected to apply to grammar learning.

2.2. Digital flashcards

Digital flashcards, used on either a website or mobile app, use a paired-associate learning paradigm that typically includes two modes: the presentation mode, in which target words are presented together with their L1 translation (e.g. *house-casa*), and the retrieval mode, which consists of two stages (the Two-Stage Framework, described in Kornell & Vaughn, 2016). In the retrieval stage, the learner sees an item (e.g. *house*) and attempts to produce the paired-associate (e.g. *casa*), while in the feedback stage, the target response is presented. Previous research and practice has employed digital flashcards in vocabulary learning, rather than grammar. The type of knowledge gained from these vocabulary activities is that of idiosyncratic, non-derivable and arbitrary associations (Ullman & Lovelett, 2018). Conversely, grammar learning entails the acquisition of morphosyntactic patterns to be applied to infinite combinations of vocabulary. With no known previous research into digital flashcards for grammar, this section will first review the evidence from vocabulary studies, followed by a theoretical outline of how flashcards may be expected to improve grammatical accuracy.

Firstly, vocabulary research has shown that retrieval is more beneficial for learning than presenting paired associates together. Carrier and Pashler (1992) found that recalling an English word from a Yupik (Eskimo language) cue strengthened conceptual associations more than seeing both words simultaneously. Barcroft (2007) saw better retention of Spanish words with a 12-s lag between image and word presentation than with simultaneous image-word presentation. Kang (2010) likewise found an advantage for retrieval practice over restudy in learning Chinese logographs from English cues. Additionally, Kang, Gollan, and Pashler (2013) compared retrieval practice to imitation for learning Hebrew vocabulary. The retrieval condition outperformed imitation in both receptive (selecting the target picture) and productive (saying the target word) measures.

Explanations for the benefits of retrieval over presentation can be found in the cognitive psychology literature. According to Bjork's (1994, 1999) Desirable Difficulties Framework, any training is optimized by adding complexity and effort. A key difference between flashcards and behaviourist imitation drills is that retrieval demands more cognitive effort than repeating

or reciting (Roediger & Karpicke, 2006). Pyc and Rawson's (2009) Retrieval Effort Hypothesis applied the principles of Bjork's framework to flashcard training, claiming that difficult successful retrievals are better than easier successful retrievals for long term memory. Manipulations of retrieval effort can enhance vocabulary learning; for example, adding spacing between target items (known as the spacing effect). For additional manipulations, see Nakata (2015; 2017).

Kornell and Vaughn (2016) describe how even unsuccessful retrieval attempts are beneficial. Moreover, the more confidently an incorrect response is given, the more effective the subsequent feedback. The testing stage causes the learner to pay more attention to the feedback stage (see Kornell, 2009; Roediger & Karpicke, 2006). This *testing effect* can be magnified by adding more testing stages before the feedback stage (Izawa, 1970).

The testing effect is particularly relevant to digital flashcard training, as many attempts will be unsuccessful due to the nature of the *drop-out schedule*. This is where items drop out of the cycle when answered correctly a predetermined number of times (the criterion), but stay in the cycle after an incorrect response. This means that if the criterion is one, the number of *correct* responses is equal to the number of items. For example, a set with 10 items will include 10 correct retrievals, but the learner may make any number of incorrect attempts. With no instruction or study stage, learners must start the process by trial-and-error. In a sentence-level item there are many opportunities for errors and repeated unsuccessful retrieval attempts are likely to occur. This repetition of attempts is key to the noticing of feedback in trial-and-error training. Strong and Boers (2019) compared trial-and-error to study-and-retrieval for the learning of phrasal verbs, with the latter condition outperforming. Participants in the trial-and-error group were presented with sentences containing a phrasal verb, with the preposition missing. They guessed the correct preposition once and subsequently received feedback. For this group, 70% of post-test errors were duplicates of responses given during the trial-and-error phase, prompting the authors to suggest that feedback is often ineffective. However, when using a drop-out schedule, the incorrectly guessed items cycle back, giving the learner repeated opportunities to recall the feedback until they produce a correct response, thus guaranteeing that each item of feedback is internalised and reproduced once during training. This therefore could provide a solution to the ineffectiveness of feedback reported by Strong and Boers.

This drop-out schedule paradigm guarantees the occurrence of *noticing*, defined as consciously registering a form (Schmidt, 2010), which, according to Schmidt (1990; 2010)'s Noticing Hypothesis, is crucial to grammar acquisition. Furthermore, the Output Hypothesis (Swain, 1993; 1995; 1998) claims that one function of output is to promote noticing. When unable to produce the target output, the learner "notices the hole" (Swain, 1998, p. 66) in their knowledge and is triggered to look for it in the input. Some studies have found that the easier it is for learners to compare their output to the target forms (in this case, "noticing the gap"), the more they will benefit from this "triggering" function of output (Izumi & Bigelow, 2000; Izumi, Bigelow, Fujiwara, & Fearnow, 1999). Research has supported the notion that output-plus-feedback training enhances noticing in language-focused tasks (Izumi et al., 1999; Khatib & Alizadeh, 2012, p. p173; Nobuyoshi & Ellis, 1993). Most studies examining this noticing role of output have involved meaning-oriented tasks, where feedback may go unnoticed. However, digital flashcards in the productive recall mode require written output for each cue and then provide a visual comparison of the learner's attempt with the target response. Furthermore, they guarantee that each item will receive the learner's attention through the repetitive nature of the drop-out schedule. In sum, digital flashcards effectively utilise output and feedback to prompt the noticing of grammatical forms, which should facilitate their acquisition.

In order to successfully use digital flashcards for grammar, the learner must be able to independently locate gaps in their knowledge, theorise possible rules or patterns governing their errors, and retain multiple ideas simultaneously. This type of learning is more complex than in digital flashcards for vocabulary (based on rote learning), which might also explain why flashcards have not been used for grammar learning. Moreover, considering the complexity of the task, the effects of noticing previously mentioned may not apply if low education participants are differently able to notice. Penning et al. (2019), as mentioned, found their treatment to be ineffective for learners of low education compared to medium and high education subjects. Additionally, Bigelow, Delmas, Hansen, and Tarone (2006) replicated a study of university students (Philp, 2003) using a sample of less educated learners (L1 Somali) on their ability to notice recasts. They found that low-literacy learners noticed fewer recasts compared with the previous study, and ability to respond to recasts was also related to literacy level. Noticing, according to Robinson (1995) takes place in working memory, which can be limited by low education (Juffs, 2006). In this regard, previous research may not be a good indicator of learning outcomes in the present sample.

Reviews of SLA sample demographics (Ortega, 2019; Plonsky, 2014), reveal that the vast majority of participants have been young adults in higher education institutions in North America or Western Europe. It is for this reason that several scholars have raised concerns about the generalizability of research findings pertaining to L2 learning. This paper may be seen as a step towards filling this research gap.

3. Research questions

The aim of this study is to investigate whether flashcards can be used at the sentence level to improve L2 grammar, motivated by the need to aid language learners, especially (though not exclusively) those without access to formal education or other language learning opportunities. The research questions (RQ) are as follows:

RQ1: Can full-sentence flashcard translation training lead to improved L2 grammatical accuracy?

RQ2: How effective will the training be (i) immediately after treatment? (ii) two weeks after treatment? (iii) 18 weeks after treatment?
RQ3: To what extent can a refresher mitigate potential losses in retention 18 weeks after the treatment?

The third research question was added assuming that losses would occur after an extended period, as was the case in flashcard studies for vocabulary learning (e.g. Ashcroft, Cvitkovic, & Praver, 2018; Franciosi Yagi, Tomoshige & Ye, 2016; Nakata & Webb, 2016). The refresher was a single training session for half the participants (see section 4.4) intended to re-expose them to knowledge acquired in the treatment.

4. Methodology

4.1. Setting

The participants for the current study come from a rural village in Cambodia, where the school system cannot provide the necessary education in English language needed to pursue further education, study abroad, or succeed in the job market. As a response, students meet daily for lessons with short-term volunteers in an improvised classroom. These volunteers, though not necessarily high-proficiency English users, teach mainly explicit rules while also spending time communicating with students outside of the classroom. After one year, students would still be classed as A1 under the Common European Framework of Reference, based on their output during this study. They are able to use high-frequency phrases with relative fluency, and communication is possible, but morphosyntax has not developed to a level that meets the learners' needs. See Appendix A for examples of the participants' knowledge at pre-test.

4.2. Participants

The improvised school has 10 classes distributed into six levels of age and ability. All students from levels four to six were recruited for this study ($n = 38$), while lower-level students, aged 8 and below, were considered too young to participate. Within the target classes, data from some students could not be included because they were not available for all sessions ($n = 3$) or because they scored 14/16 or over in the pre-test ($n = 3$) leaving little room for improvement. One participant was retroactively excluded from the data due to noticeably different cognitive abilities. The final sample included 31 participants. Gender was evenly split (Females = 16, Males = 15) and ages ranged from 9 to 17 ($M = 13.5$, $SD = 2.4$), with clusters around ages 12 and 15.

A control group from the same setting was recruited a year later. They were not available for all data collection points, but serve as a baseline for the amount of incidental knowledge gained in a two-week period, approximately corresponding to the treatment time as well as the delay between the first and second post-tests. The use of this control group was intended to increase confidence that any effects seen in post-tests were due to the treatment. The group originally contained 33 members, but, unexpectedly, their pre-test scores were lower than that of the experimental group. In order to maintain the comparability of the groups, participants scoring 0 or 1 in the pre-test were excluded. The final control group contains 19 members including 13 females and six males, ranging in age from 12 to 16 ($M = 13.9$, $SD = 1.4$).

4.3. Instruments

4.3.1. Tool

The tool was the free app *Cram.com Flashcards* from Cram.com LLC (2016), which allows users to create custom flashcards and access them from any connected device. Errors in letter case, punctuation, or spaces were set to be ignored so as not to interfere with the data. There was no presentation stage, allowing participants to learn solely via the testing effect (Kornell, 2009, 2016; Roediger & Karpicke, 2006). The app's "Memorize" mode was employed, whereby flashcards continue to cycle through the set but disappear once answered correctly. "Text-input" was activated, requiring written answers from participants.

Each flashcard has two stages. In the first stage (Fig. 1), participants see the item in Khmer and must type the English translation. In the second stage (Fig. 2), feedback is presented.

4.3.2. Items

The target items were full sentences, grouped into eight categories (although each item contained multiple grammatical features). The first four groups were declarative sentences: (a) present simple, (b) present continuous with *is*, (c) present continuous with *am/are*, and (d) *there is/are*. The remaining sets were the same items in the interrogative form. Each group contained five items, with 40 in total. The chosen items were simple sentences using vocabulary the participants already knew and used regularly, based on the first author's experience in the context. The rationale of using familiar vocabulary was to keep the focus of the study on grammar. Items provided the opportunity to practice common errors, again based on

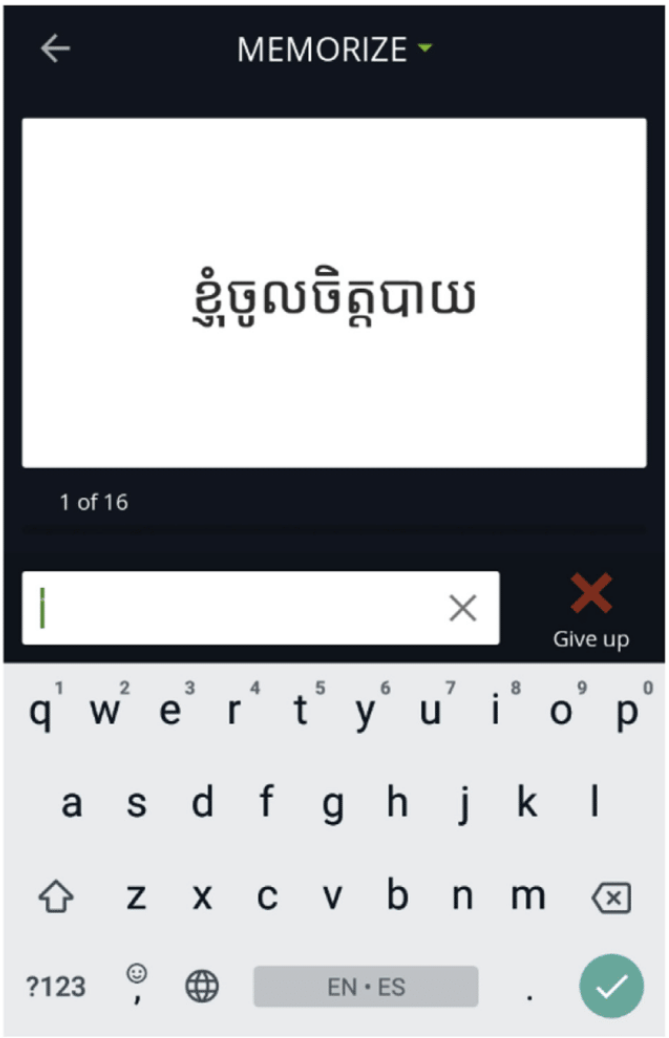


Fig. 1. Retrieval attempt.

experience in the context, such as conjugating the present simple and present continuous for 1st, 2nd, and 3rd person, using *there is* or *there are*, pronouns, articles, and plurals. Table 1 shows a breakdown of the items.

4.3.3. Item distribution

Flashcards were organized into eight sets, gradually introducing new groups. As items became more familiar, the size of the sets increased to keep the retrieval effort high, following the Retrieval Effort Hypothesis (Pyc & Rawson, 2009) and to allow for items to be repeated on different days. This corresponds with previous research which found that repetition led to higher retention (Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982; Pyc & Rawson, 2009; Rawson & Dunlosky, 2011). The distribution of groups and sets is shown in Table 2.

4.3.4. Tests

Two tests were used (see Appendix B). Test A (trained items) comprises 16 items, including two items from each group of flashcards, selected for maximum representation of the grammar points present in the treatment. Test B (untrained items) comprises an equivalent 16 items, using only vocabulary and grammatical structures found in the treatment, but in novel combinations. A test of untrained items was included to ensure that any gains made in the post-tests were not due to rote memorisation of trained items.

4.4. Procedure

The pre-test was administered on smartphones using Google forms. All three post-tests (immediate, two-week, 18-week) were carried out with pen and paper due to the logistics of testing many participants with limited available phones. As participants had no time limit and were encouraged to check answers thoroughly for all tests before submitting, this is not expected to have affected results.

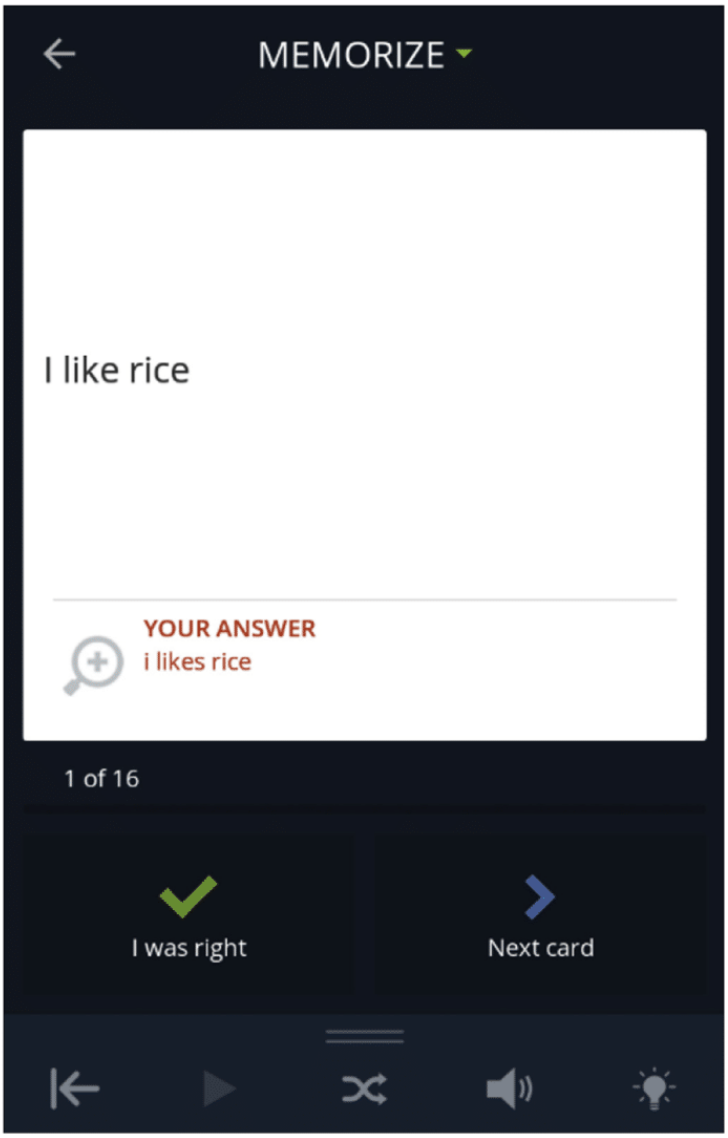


Fig. 2. Feedback.

Each test was coded as follows: '1' = pre-test, '2' = immediate post-test, '3' = two-week delayed post-test and '4' = 18-week delayed post-test. "A" denotes trained items and "B" denotes untrained items. For example, T2A represents trained items in the immediate post-test.

Participants first completed Test A (items to be trained) as a pre-test and started the treatment the following day. The test can be considered reliable according to Cronbach's Alpha ($\alpha = 0.852$). Participants completed each set individually, one set per day, using smartphones from volunteer teachers which had the app installed. Many participants missed a day and caught up by completing two sets on the next day. The treatment also coincided with a national election which caused a two-day interruption in the middle of treatment. Consequently, the 8 sets were completed during 10 days. The context dictated that tests and treatment were administered in an outdoor, communal area, monitored to ensure other students did not interfere. This served to increase the ecological validity of the study.

The day after the final treatment session, participants took Tests T2A ($\alpha = 0.744$) and T2B ($\alpha = 0.744$). They were not given advanced warning of delayed post-tests. The first delayed post-test (T3A: $\alpha = 0.733$; T3B: $\alpha = 0.669$) was given two weeks after treatment (as per Cerezo et al., 2016; Nutta, 1998), and the final post-test (T4A: $\alpha = 0.734$; T4B: $\alpha = 0.749$) was given 18 weeks after treatment. A refresher was included in this study in order to examine its effect on retention. A retention interval of 18 weeks was expected to show forgetting of previously learned material in 1-day intersession intervals (see Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008 or Rohrer & Pashler, 2007, for a discussion on how spacing between learning sessions should be distributed for optimal long-term retention). The refresher was a single set of flashcards containing the 16 items that appear in Test A. It was performed one week before the 18-week delayed post-test, aiming to remind students of previously acquired knowledge. The original sample ($N = 31$) was divided into two groups matched for age, gender and previous scores: Group-R ($n = 15$), which would take the refresher treatment and Group-NR ($n = 16$), which would have no extra

Table 1
Items of the treatment.

Group 1 (pres. simple)	Group 2 (pres. cont: is)	Group 3 (pres. cont: am/are)	Group 4 (there is/are)
I like rice.	He is playing volleyball.	I am eating.	There is a girl in my house.
You like chicken.	The boy is playing.	You are eating.	There are girls in my house.
He likes rice.	The girl is jumping.	The boys are eating.	There is a girl in the shop.
She likes chicken.	She is sitting.	The girls are eating.	There is a boy in the shop.
They like rice and chicken.	The chicken is eating.	I am playing volleyball.	There are boys in my house.
Group 5	**Group 6**	**Group 7**	**Group 8**
Do you like rice?	Is he playing volleyball?	Am I eating?	Is there a girl in my house?
Do you like chicken?	Is the boy playing?	Are you eating?	Are there girls in my house?
Does he like rice?	Is the girl jumping?	Are the boys eating?	Is there a girl in the shop?
Does she like chicken?	Is she sitting?	Are the girls eating?	Is there a boy in the shop?
Do they like rice?	Is the chicken eating?	Am I playing volleyball?	Are there boys in my house?

Table 2
Distribution of items across sets.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	TOTAL
Group 1	✓						✓		2[a]
Group 2		✓			✓		✓		3
Group 3		✓	✓				✓		3
Group 4			✓	✓				✓	3
Group 5				✓	✓			✓	3
Group 6					✓	✓		✓	3
Group 7						✓	✓		2[a]
Group 8				✓		✓		✓	3
Items	5	10	10	15	15	15	20	20	

[a] In order to avoid making sets larger than twenty items, two groups appear only twice, as opposed to three times for other groups. These groups were chosen because group 1 contained the simplest constructions and group 7 had been seen twice within the final three sets, and would thus benefit from recency effects.

treatment. However, due to five absentees from Group-R on the day of the refresher, four members of Group-NR were chosen at random to replace them. This created an imbalance, with Group-R having higher average gains at T2A and fewer members ($n = 14$) than Group-NR ($n = 17$). To address this imbalance and preserve the comparability of the groups, the three lowest scoring participants at T2A were excluded from Group-NR. The final number of participants per group was 14. Fig. 3 illustrates the experimental design.

A control group in the same context were given T1A (pre-test) and T2A (post-test) with a two-week delay. They received no flashcard training but, like the experimental group, they did receive daily instruction from volunteer teachers which incidentally covered many target forms of the study. This group's data was intended to indicate the amount of grammar learning that can take place in a two-week period in the context under analysis in order to support the claim that any gains among the experimental group can be confidently attributed to the treatment. They were not given Test B (untrained items) because this test served to establish whether correct responses among the experimental group were due to memorisation of language chunks or to learning of grammatical patterns, which was not an issue for the control group.

4.5. Scoring

Items were scored dichotomously, 1 point for each correct answer matching the target item exactly, with the following exceptions: (1) if a base word was spelled incorrectly but otherwise used correctly, for instance "gril" instead of "girl"; (2) if the wrong word was used, the only instance being the use of "football" instead of "volleyball"; (3) if the answer is an acceptable translation of the Khmer and still grammatically correct in English, for instance "Girls are eating" rather than "The girls are eating". The former two exceptions are because this study does not focus on vocabulary, and the latter exception is

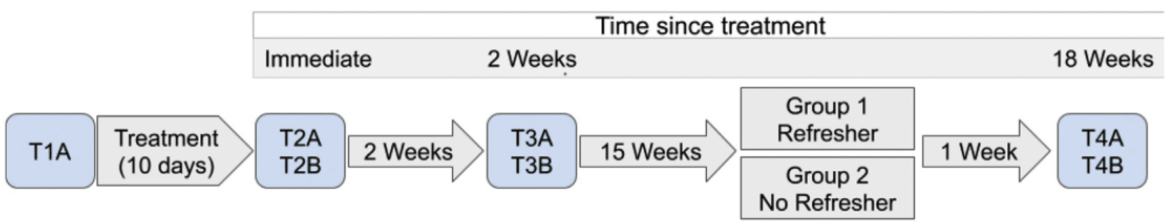


Fig. 3. Experimental design.

because the Khmer language does not differentiate between these two types of sentences, so without context both answers are fair translations. A second rater was instructed in the rubric and independently graded one test per participant at random (14.2% of total tests), with interrater agreement of 100%.

4.6. Analysis

As results were not normally distributed and the sample was small, the analysis was conducted using non-parametric tests. In order to answer RQ1, test scores for trained and untrained items were compared within subjects using the Related-Samples Wilcoxon Signed Rank Test for T2, T3, and T4, to check if results were due to rote learning of individual sentences or system learning. Next, RQ2 was addressed and Test A scores were compared between times using Wilcoxon Signed Rank Tests to establish the amount learned and retained (the same was not done for Test B, as there were no pre-test scores for this test). The same statistical analysis was performed for the control group between T1 and T2. Relative gains were also computed for Test A in order to more clearly present the effect of the treatment and allow results to be compared with other studies. The formula for this was (learned items/(total number of items - known items)) x 100 (Peters & Webb, 2018). Learned items are those which were incorrect in the pre-test and correct in the post-test, and known items are those which were answered correctly in both pre-test and post-test. For the 18-week delayed post-test, Groups −R and −NR were calculated separately in order to assess the refresher's effect (RQ3). Mann-Whitney *U* tests were performed to compare results between groups, as well as to confirm equal distribution of age, gender, and previous scores between Group R and −NR. The comparison between T2 to T4 and T3 to T4 scores constituted the measure of losses for each group.

5. Results

Table 3 displays the results of tests A (trained items) and B (untrained items) at each time: 1 = pre-test; 2 = post-test; 3 = two-week delayed post-test; 4 = 18-week delayed post-test for the experimental and control groups. Fig. 4 focuses on the experimental group and shows boxplots for T1, T2, T3, and T4, with the latter tests split by group.

The descriptive statistics show substantial gains after the treatment, and, even though there is evidence of decay across time, the performance in the delayed post-tests was quite accurate, especially for Group-R. Additionally, the results of Test A (trained items) and Test B (untrained items) seem quite similar. In contrast to the experimental group, the control group show virtually no gains between T1A ($M = 4.32$, $SD = 3.43$) and T2A ($M = 4.74$, $SD = 2.38$).

5.1. Trained vs untrained items

Results for trained and untrained items were compared at T2, T3, and T4 using a Related-Samples Wilcoxon Signed Rank Test.

At T2, Test A scores ($M = 13.77$, $SD = 2.38$, $Mdn = 14$) were significantly lower than Test B scores ($M = 14.39$, $SD = 1.96$, $Mdn = 15$), $Z = 2.44$, $p = .015$. At T3, Test A scores ($M = 13.39$, $SD = 2.55$, $Mdn = 14$) were again lower than Test B scores ($M = 13.94$, $SD = 2.14$, $Mdn = 15$), approaching statistical significance, $Z = 1.79$, $p = .074$. T4 was similar to T2, with Test A scores ($M = 12.42$, $SD = 2.83$, $Mdn = 13$) significantly lower than Test B scores ($M = 13.13$, $SD = 2.62$, $Mdn = 14$), $Z = 2.57$, $p = .010$.

This advantage to untrained items was unexpected. To explain this difference, the sum of correct answers for each test item across all participants was compiled, and the differences between Test A and Test B were calculated by item. For example, if an item was answered correctly by 7 participants on T2A and its corresponding item was answered correctly by 8 participants on T2B, the difference would be 1. The mean difference between tests was low at T2 ($M = 1.19$), T3 ($M = 1.06$), and T4 ($M = 1.37$). However, three items were outliers in how many times an item was answered incorrectly in Test A, but correctly in Test B. These were items *6* (differences: T2 = 8; T3 = 5; T4 = 10), *12* (differences: T2 = −1; T3 = 7; T4 = 8) and *13* (differences: T2 = 14; T3 = 11; T4 = 13). Looking at these items, the cause of the disparity seems to be in errors relating to the complexity of item subjects. Test A items with "The boys", "the chicken", and "the girls" are paired with Test B items containing "You", "She",

Table 3
Descriptive statistics raw scores (maximum = 16) for experimental and control groups.

		Experimental ($n = 31$)		Control ($n = 19$)
		Test A (trained)	Test B (untrained)	Test A
T1 (pre-test)	Mean (SD)	5.71 (3.42)	–	4.32 (3.43)
	Median	5	–	3
T2 (immediate post-test)	Mean (SD)	13.77 (2.38)	14.39 (1.96)	4.74 (3.85)
	Median	14	15	4
T3 (2-week delayed post-test)	Mean (SD)	13.39 (2.55)	13.94 (2.14)	–
	Median	14	15	–
T4 NR (18-week delayed post-test, no refresher)	Mean (SD)	11.79 (2.52)	12.64 (2.27)	–
	Median	11	12.5	–
T4 R (18-week delayed post-test, refresher)	Mean (SD)	14 (1.96)	14.43 (2.21)	–
	Median	14.5	15.5	–

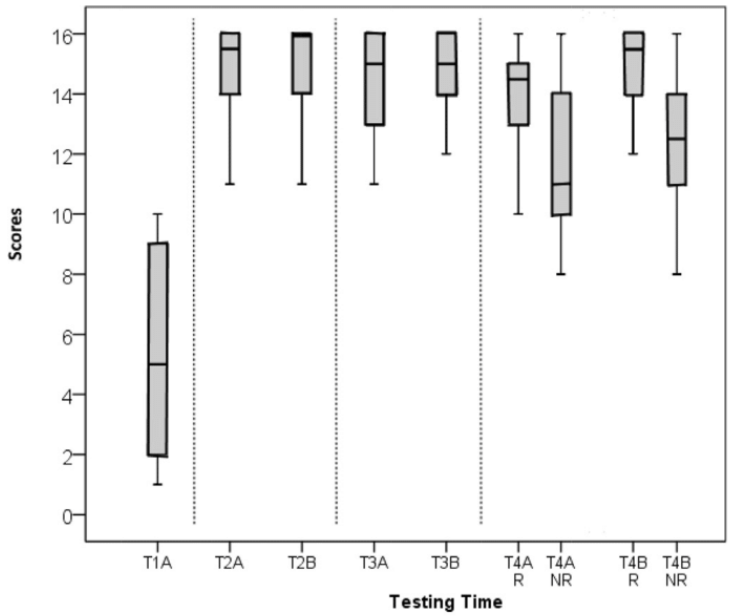


Fig. 4. Boxplots of raw scores (maximum 16) for trained (A) and untrained (B) items at all testing times.

and "I". The former create more opportunity for error, by omitting an article ("Chicken is eating") or a plural *-s* ("Are the girl eating"). Consequently, Test A had more opportunity for error here than its counterpart. When only the biggest outlier (item *13*: Are the girls eating?) is removed from the data, no significant differences are found between trained and untrained items for T2 ($Z = 0.962$, $p = .336$), T3 ($Z = 0.775$, $p = .439$), or T4 ($Z = 1.083$, $p = .279$).

In sum, the data show that participants were able to apply grammatical knowledge acquired during the digital-flashcard training to translate both trained and untrained items.

5.2. Gains and retention

The question of the treatment's effect at different testing points was addressed by comparing test A scores. As tests A and B were shown to be equivalent at all testing points, it would be redundant to present separate analyses of tests A and B. Of the two, test A was chosen because it is the one used for the pre-test and for the control group.

In order to examine gains from pre- to immediate post-test, T1A scores ($M = 5.71$, $SD = 3.42$, $Mdn = 5$) and T2A ($M = 13.77$, $SD = 2.38$, $Mdn = 14$) scores were submitted to a Related-Samples Wilcoxon Signed Rank Test and the difference in scores was significant ($Z = 4.874$, $p < .001$). When converted to relative gains (see Table 4 for mean relative gains), the immediate post-test (T2A) showed mean gains of 82.39% ($SD = 19.61\%$, $Mdn = 85.71\%$), ranging from 40% ($n = 1$) to 100% ($n = 10$). In contrast, the difference in scores among the control group at T1A ($M = 4.32$, $SD = 3.43$, $Mdn = 3$) and T2A ($M = 4.74$, $SD = 3.84$, $Mdn = 4$) was not significant ($Z = 0.492$, $p = .622$), and their relative gains were on average 12.91% ($SD = 15.88\%$, $Mdn = 7.14\%$). A Mann-Whitney test indicated that the two groups were matched for pre-test scores, $Z = 1.517$, $p = .129$.

In terms of retention, the two-week delayed post-test (T3A) produced a mean score of 13.39 ($SD = 2.55$, $Mdn = 14$), which in relative gains from pre-test is 76.99% ($SD = 17.68\%$, $Mdn = 80.00\%$). A Related-Samples Wilcoxon Signed Rank Test found no significant difference between T2 and T3 scores, $Z = 0.898$, $p = .369$, suggesting learners largely retained their knowledge from the treatment two weeks after the fact.

At the 18-week delayed post-test (T4A), Group-NR ($n = 14$) scored 11.79 ($SD = 2.52$, $Mdn = 11$). T3A scores for this subset ($M = 13.64$, $SD = 1.86$, $Mdn = 14$) were significantly higher than at T4, $Z = 2.363$, $p = .018$. Compared with their T2A scores ($M = 13.71$, $SD = 1.73$, $Mdn = 14$), the difference was also significant, $Z = 2.728$, $p = .006$. However, despite significant losses between the immediate/two-week delayed post-tests and the 18-week delayed post-test, the students showed overall relative gains of 61.56% (SD = 23.58%, $Mdn = 59.42\%$) with respect to their pre-test scores.

In contrast, Group-R's ($n = 14$) scores for T3A ($M = 13.93$, $SD = 2.76$, $Mdn = 15$) and T4A ($M = 14$, $SD = 1.96$, $Mdn = 14.5$) were not significantly different, $Z = 0.052$, $p = .958$. However, T2A scores for this subset ($M = 14.93$, $SD = 1.49$, $Mdn = 15.50$), were found to be statistically higher than their final T4A scores, $Z = 2.157$, $p = .031$. These results suggest that the learners largely retained what they had learned from the treatment not only two weeks later but also 18 weeks later, with a statistical difference only visible when the entire period is examined. This contrast between the two groups is salient considering that

Table 4
Relative gains (%) experimental and control groups.

		Experimental ($n = 31$)	Control group ($n = 19$)
T1-T2	Mean (SD)	82.39 (19.61)	12.91 (15.88)
	Median	85.71	7.14
T1-T3	Mean (SD)	76.99 (17.68)	–
	Median	80	–
T1-T4 NR	Mean (SD)	61.56 (23.48)	–
	Median	59.42	–
T1-T4 R	Mean (SD)	79.03 (13.32)	–
	Median	79.29	–

Groups −R and −NR were matched for distribution of T3A scores, $Z = 1.22$, $p = .246$. For Group-R, the final overall gains, with respect to T1, were 79.03% ($SD = 13.32\%$, $Mdn = 79.29\%$).

An Independent-Samples Mann-Whitney *U* Test revealed that the difference in gains between T1A and T4A between Group-R ($M = 79.03\%$, $SD = 13.32$, $Mdn = 79.29$) and Group-NR ($M = 61.56\%$, $SD = 23.58$, $Mdn = 59.42\%$) reached statistical significance, $Z = 2.233$, $p = .026$.

6. Discussion

This paper set out to explore whether digital flashcards may be used to improve grammatical accuracy. English language learners (ELLs) aged 9 to 17 in a low-resource, low-education context underwent eight sessions of full-sentence flashcard training in which they produced target language samples, prompted by translations from their L1 Khmer. They were required to type each item correctly once for it to drop from the set. Otherwise, participants were presented with the target response and the item returned to the cycle. After assessing participants' previous knowledge of the target forms (T1), the participants' performance was examined immediately after treatment (T2), two weeks after treatment (T3), and 18 days after treatment (T4) in order to assess their learning and retention of the target forms. Two tests were used to assess grammatical accuracy: Test A, comprising selected sentences from the treatment, and Test B, including the same target grammar but in novel sentences. Each research question will now be discussed in turn.

RQ1: Can full-sentence flashcard translation training lead to improved L2 grammatical accuracy?

The first research question asked whether the treatment would lead to improved grammatical accuracy. Gains in the post-test demonstrate that accuracy did improve, but to test whether participants achieved this by learning grammatical patterns, as opposed to memorising chunks of language, scores from items used in training (Test A) were compared to scores from equivalent items (Test B), using the same vocabulary and structures but in novel combinations. The results showed no significant difference between scores on trained and untrained items in the immediate post-test. This held true over time, even at T4 when half the group had been given extra practice on trained items (the refresher). Had the participants been memorising individual sentences, the trained items should have scored higher than untrained items, according to the experiment's rationale. Furthermore, if neither chunk nor grammar learning had occurred, then post-test scores would logically be similar to pre-test scores. Given that pre-test scores were low ($M = 5.71/16$) and that post-test scores were high (T2A: $M = 13.77/16$, T2B: $M = 14.39/16$) we can confidently conclude that the treatment helped participants to improve their grammatical accuracy of target forms. This is especially interesting as participants were never instructed to infer grammatical rules from the samples, nor that there would be a post-test of untrained items. No rule or instruction was provided with the samples. These results stand in contrast with the control group, which showed no significant change over an equivalent period of time following the regular school practice.

It would be reasonable to consider these findings as further evidence in support of the noticing function of output. If we assume that acquisition of forms is evidence of attention to forms (Schmidt, 2010), then this finding supports previous conclusions that output-plus-feedback promotes the noticing of grammatical structures (Izumi et al., 1999; Khatib & Alizadeh, 2012, p. p173; Nobuyoshi & Ellis, 1993) and that flashcard training seems to provide the necessary conditions for this to occur.

The success of the treatment is particularly salient given the low education background of the participants. It would appear that any disadvantage in their capacity to notice (as in Penning et al., 2019; Bigelow et al., 2006) was mitigated by ensuring that participants notice each item.

RQ2: How effective will the training be i) immediately after treatment? (ii) two weeks after treatment? (iii) 18 weeks after treatment?

The second research question concerned the extent of learning and retention through the treatment. The immediate post-test gains are indisputably high at over 80% ($M = 82.39\%$), and include 10/31 participants with 100%. Gains remained high after two-weeks ($M = 76.99\%$) and 18-weeks (Group-R: M = 79.03%; Group-NR: M = 61.56%). The drop in scores is statistically

significant for the no-refresher group when comparing scores for the immediate (M = 13.71/16) and two-week delayed (M = 13.64/16) post-tests with the 18-week delayed post-test (M = 11.79/16), though only by two items. The amount of knowledge that was retained in this study is quite impressive, considering the long retention intervals between the end of the treatment and both delayed posttests (Cepeda et al., 2008; Rohrer & Pashler, 2007).

Despite the different approach to CALL for grammar learning, these results are similar to Cerezo et al.'s (2016) results from videogame instruction among beginners, which also used written translation post-tests and reported gains of 83% at immediate post-test and 72.6% at two-week delayed post-test. In contrast, Ashcroft et al.'s (2018) study on flashcards for vocabulary items reported gains of 37% for beginners, and delayed post-test gains, three weeks later, dropped to 17%. Similarly, Nakata and Webb's (2016) study on short versus long spacing in vocabulary flashcard training showed immediate post-test results of 58% and 62%, which after a one-week delayed post-test was reduced to 8% and 20%. It seems, based on these data, that flashcard training may actually be more effective for grammar than for vocabulary, in terms of long term learning. There are several potential explanations for this phenomenon. Firstly, whole sentences may require more effort to reproduce than single word items, which would make the overall training more effective according to the Retrieval Effort Hypothesis. With more parts of a single item to remember, the likelihood of an incorrect retrieval is greater, giving the participant more opportunities for unsuccessful retrieval and feedback. Another factor could be that each item in the set complements the learning of the underlying rules. For example, if one item uses a first person subject, and the next uses a third person subject, the rule that only the third person requires an 's' is evident in both of these items. This is in contrast with a vocabulary-based design where each item is independent. Finally, grammatical structures may be more likely to be used communicatively, and therefore practiced, than specific vocabulary items. The structures used in this study are particularly ubiquitous in English. For example, an item in this study was "Do you like rice?", and so any time a student asked someone a question in the present simple, they would have the opportunity to practice this question inversion using "do". Likewise, they would have more opportunities to notice this structure in their input, as compared with vocabulary items. This additional communicative practice would naturally lead to higher retention. Nevertheless, it must be pointed out that communicative practice alone without the digital-flashcard training was not enough in the context under analysis for participants to acquire the target grammar forms from the input, as demonstrated by the results of the control group and the fact that the experimental group had, prior to the treatment, been studying for a year already with little progress, as shown in their pre-tests.

RQ3: To what extent can a refresher mitigate long-term losses in retention?

For the third research question, the sample was split into two groups of 14, in order to test the effect of recently reviewing the target items before the eighteen-week post-test. Group-R underwent a refresher set of flashcards containing only the 16 items of Test A, while Group-NR had no extra treatment. Results showed that while Group-R managed to maintain their previous knowledge, Group-NR showed small but significant losses over time. It should be noted that the overall retention was high in both groups (R: 79.03%, NR: 61.56%), and it remains unclear whether the refresher would have had the same effect in a scenario with greater overall losses. That said, in this study, it was indeed enough to prevent losses, demonstrating that minimal re-exposure to target forms through flashcard training aids in the retention of previously learned grammatical patterns. The higher performance of Group-R at T4 can be explained by a combination of extra training and a more recent exposure to the target items before testing. At the final delayed post-test, the retention interval was one week for Group-R compared with 18 weeks for Group-NR. Both of these explanations highlight the importance of continuously retrieving (or practicing) target structures in order to prevent forgetting.

7. Pedagogical implications

In light of our results, digital flashcards can be taken as a useful activity for allowing learners to practice producing accurate language with immediate and accurate feedback on every attempt. Even in settings with highly competent teachers, flashcards may be used to give students extra output practice outside of class, while perhaps freeing up time in class for more meaning-focused activities (in line with the suggestions made by Cerezo et al., 2016). Moreover, the current findings have shown that flashcards offer a solution for learners to study independently and receive feedback on their output in environments lacking in teachers and authentic input. Some may assert that a competent teacher and genuine interaction cannot be replaced, but such opportunities are not as ubiquitous as one would hope. With many NGOs focusing on training local teachers, who may themselves be undereducated, it may be wise to first invest in devices and internet connections so that students can access free learning resources in the short term. By doing this, learners will have access to reliable, consistent input, allowing them to study at their own pace, while being guaranteed quality feedback on their work. Additional support for the use of digital flashcards can be found in the literature comparing CALL with human instruction, in which, often, students learn at least as well or even better with CALL than with a teacher (McEnery et al., 1995; Nutta, 1998; Mohamad, 2009; Abu Naba'h, 2012).

This study focused on quite a homogenous group with participants of the same background, all of whom had been exposed to English for approximately one year before the treatment. It is therefore recommended that flashcards be more widely researched for the purposes of improving grammatical accuracy for learners of different proficiency levels, languages, and socioeconomic backgrounds. It would also be interesting to attempt to optimise the training by manipulating spacing,

frequency, item-type distribution and other factors. Furthermore, the only tests in this study were written translations from the L1, which leaves the question of how the treatment affected the students' other facets of language, such as spontaneous speech and open-ended writing.

8. Conclusions

This study investigated the use of flashcards for grammar learning and retention over four data collection points. Flashcards have previously been tested for their effectiveness in learning vocabulary, but the high gain scores of this study demonstrate that flashcards should also be investigated for grammar learning. This study is not without limitations. First of all, although the subjects in this study represented almost all members of the chosen population, the sample was small, especially for analysing the effect of the refresher. Secondly, the tests, which were designed for the purposes of this study, could be improved by ensuring equal difficulty between paired trained and untrained items.

The present participants are not the typical group included in most SLA studies (high education, high SES, mostly from Western countries) and CALL seems to have worked as well for them as for more privileged students. Through digital flashcards, these ELLs successfully improved their accuracy in trained grammatical forms and largely retained these gains after four months. Consequently, flashcards are recommended as a robust solution for learners without access to trained language teachers. It is hoped that more research will be carried out among different populations, outside the realm of western university students, in order to produce more generalisable data that better represents the diversity of learners and their needs.

CRediT authorship contribution statement

Jonathan Serfaty: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. **Raquel Serrano:** Supervision, Writing - review & editing.

Appendix A. Range of pre-test responses

Target	Highest Score (11/16)	Median Score (5/16)	Lowest Score (1/16)
I like rice	I like rice	I like rice	I like rice
He likes rice	He likes rice	He likes rice	He like rice
He is playing volleyball	He is playing volleyball	He is play volley ball	He playing volleyball
The girl is jumping	The Girl is jumping	She is jump	The girl jumping
I am eating	I am eating	I am eating	I eat
The boys are eating	The boys are eating	The boys is eating	New Boy eating
There are girls in my house	Have a girl stay in my house	In my house have one girl	Have girl in my house
There is a girl in the shop	Have a girl stay in my shop	In the shop have one girl	Have g
Do you like chicken?	Do you like chicken?	Do you like chicken	What do you like chicken?
Does he like rice?	Does he like rice?	Does he like rice	What he like eat Rice?
Is she sitting?	Is she sitting?	Does she siting	What she setdon?
Is the chicken eating?	Is chicken eating?	Does chicken eating	What chicken eating?
Are the girls eating?	Are girls eating?	Does the girls eating	What girl eating
Am I playing volleyball?	Am i playing volleyball?	Do i playing volley ball	What i playing volleyball?
Is there a girl in the shop?	Does a girl in the shop?	Does the girl in the shop	Who is girl in the market
Are there boys in my house?	Does have the boys in my house?	Does the boys in my house	What cheira Boy in my house?

Appendix B. Test items

	Test A		Test B	
	Cue	Ideal Response	Cue	Ideal Response
1	ខ្ញុំចូលចិត្តបាយ	I like rice	ខ្ញុំចូលចិត្តសាច់មាន់	I like chicken
2	គាត់ចូលចិត្តបាយ	He likes rice	នាងចូលចិត្តបាយ	She likes rice
3	គាត់កំពុងលេងបាល់ទះ	He is playing volleyball	គាត់កំពុងលេង	He is playing
4	ក្មេងស្រីកំពុងលោត	The girl is jumping	មាន់កំពុងលោត	The chicken is jumping
5	ខ្ញុំកំពុងញុាំ	I am eating	ខ្ញុំកំពុងលោត	I am jumping
6	ក្មេងៗប្រុសកំពុងតែញុាំ	The boys are eating	អ្នកកំពុងអង្គុយ	You are sitting
7	មានក្មេងស្រីជាច្រើននៅក្នុងផ្ទះរបស់ខ្ញុំ	There are girls in my house	មានក្មេងប្រុសម្នាក់នៅក្នុងផ្ទះរបស់ខ្ញុំ	There is a boy in my house
8	មានក្មេងស្រីម្នាក់នៅក្នុងហាង	There is a girl in the shop	មានក្មេងស្រីជាច្រើននៅក្នុងហាង	There are girls in the shop
9	តើអ្នកចូលចិត្តសាច់មាន់ទេ?	Do you like chicken?	តើអ្នកចូលចិត្តបាយដែរឬទេ?	Do you like rice?
10	តើគាត់ចូលចិត្តបាយដែរឬទេ?	Does he like rice?	តើនាងចូលចិត្តបាយដែរឬទេ?	Does she like rice?
11	តើនាងកំពុងអង្គុយឬ?	Is she sitting?	តើគាត់កំពុងលេងឬ?	Is he playing?
12	តើមាន់កំពុងស៊ី?	Is the chicken eating?	តើនាងកំពុងលោតឬ?	Is she jumping?
13	តើក្មេងស្រីៗកំពុងញុាំ?	Are the girls eating?	តើខ្ញុំកំពុងអង្គុយឬ?	Am I sitting?
14	តើខ្ញុំកំពុងលេងបាល់ទ:ឬ?	Am I playing volleyball?	តើអ្នកកំពុងលេងបាល់ទ:ឬ?	Are you playing volleyball?
15	តើមានក្មេងស្រីម្នាក់នៅក្នុងហាងទេ?	Is there a girl in the shop?	តើមានក្មេងប្រុសម្នាក់នៅក្នុងផ្ទះរបស់ខ្ញុំទេ?	Is there a boy in my house?
16	តើមានក្មេងប្រុសៗនៅក្នុងផ្ទះរបស់ខ្ញុំទេ?	Are there boys in my house?	តើមានក្មេងស្រីៗនៅក្នុងហាងទេ?	Are there girls in the shop?

References

Abu Naba'h, A. M. (2012). The impact of computer assisted grammar teaching on EFL pupils' performance in Jordan. *International Journal of Education and Development using Information and Communication Technology, 8*(1), 71–90.

Andarab, M. S. (2017). The effect of using Quizlet Flashcards on learning English vocabulary. In *113th the IIER international conference*. Frankfurt, Germany.

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

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Bigelow, M., Delmas, R., Hansen, K., & Tarone, E. (2006). Literacy and the processing of oral recasts in SLA. *Tesol Quarterly, 40*(4), 665. https://doi.org/10.2307/40264303

Bindileu, E.-I. (2019). Educational developments adjusted to current socio-economic demands: The communicative approach to second language learning. *Internal Auditing and Risk Management, 53*(1), 43–54.

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.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642. https://doi.org/10.3758/bf03202713

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

Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish gustar structures. *Studies in Second Language Acquisition, 38*, 265–291. https://doi.org/10.1017/S0272263116000139, 02.

Cramcom Flashcards. (2016). *Cram.com, LLC*. Retrieved from https://www.apk4now.com/apk/52746/cram-com-flashcards/download.

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

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

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83*(2), 340–344. https://doi.org/10.1037/h0028541

Izumi, S., & Bigelow, M. (2000). Does output promote noticing and second language acquisition? *Tesol Quarterly, 34*(2), 239. https://doi.org/10.2307/3587952
Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21*(3), 421–452. https://doi.org/10.2307/44486913
Juffs, A. (2006). Working memory, second language acquisition and low-educated second language and literacy learners. In I. van de Craats, J. Kurvers, & M. Young-Scholten (Eds.), *Low-educated second language and literacy acquisition: Proceedings of the inaugural symposium- tilburg 05* (vol. 6, pp. 89–104). LOT Occasional Papers. The Netherlands: Netherlands Graduate School of Linguistics.
Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38*(8), 1009–1017. https://doi.org/10.3758/MC.38.8.1009
Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review, 20*(6), 1259–1265. https://doi.org/10.3758/s13423-013-0450-z
Khatib, M., & Alizadeh, M. (2012). Output tasks, noticing, and learning: Teaching English past tense to Iranian EFL students. *English Language Teaching, 5*(4), p173–p187. https://doi.org/10.5539/elt.v5n4p173
Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*(9), 1297–1317. https://doi.org/10.1002/acp.1537
Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation, 65*, 183–215. https://doi.org/10.1016/BS.PLM.2016.03.003
Lightbown, P. (2000). Anniversary article. Classroom SLA research and second language teaching. *Applied Linguistics, 21*(4), 431–462. https://doi.org/10.1093/applin/21.4.431
McEnery, T., Baker, J. P., & Wilson, A. (1995). A statistical analysis of corpus based computer vs traditional human teaching methods of part of speech analysis. *Computer Assisted Language Learning, 8*(2–3), 259–274. https://doi.org/10.1080/0958822940080208
Mohamad, F. (2009). Internet-based grammar instruction in the ESL classroom. *International Journal of Pedagogies and Learning, 5*(2), 34–48. https://doi.org/10.5172/ijpl.5.2.34
Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning. *Studies in Second Language Acquisition, 37*(4), 677–711. https://doi.org/10.1017/S0272263114000825
Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition, 39*(4), 653–679. https://doi.org/10.1017/S0272263116000280
Nakata, T. (2020). Learning words with flash cards and word cards. In S. Webb (Ed.), *The routledge handbook of vocabulary studies*. London and New York: Routledge.
Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? *Studies in Second Language Acquisition, 38*(3), 523–552. https://doi.org/10.1017/S0272263115000236
Nelson, T. O., Leonesio, J. R., Shimamura, A. P., Landwehr, R. F., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(4), 279–288. https://doi.org/10.1037/0278-7393.8.4.279
Nobuyoshi, J., & Ellis, R. (1993). Focused communication tasks and second language acquisition. *ELT Journal, 47*(3), 203–210. https://doi.org/10.1093/elt/47.3.203
Nutta, J. (1998). Is computer-based grammar instruction as effective as teacher-directed grammar instruction for teaching L2 structures? *CALICO Journal, 16*(1), 49–62. https://doi.org/10.1558/cj.v16i1.49-62
Ortega, L. (2019). SLA and the study of equitable multilingualism. *The Modern Language Journal, 103*, 23–38. https://doi.org/10.1111/modl.12525
Penning de Vries, B. W., Cucchiarini, C., Strik, H., & van Hout, R. (2019). Spoken grammar practice in CALL: The effect of corrective feedback and education level in adult L2 learning. *Language Teaching Research, 136216881881902*. https://doi.org/10.1177/1362168818819027
Philp, J. (2003). Constraints on "noticing the gap. *Studies in Second Language Acquisition, 25*(1), 99–126. https://doi.org/10.1017/S0272263103000044
Plonsky, L. (2014). Sampling, power, and generalizability in L2 research (Or, why we might as well be flipping coins). In *Keynote presentation at the second language studies symposium*. Michigan: East Lansing.
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
Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140*(3), 283–302. https://doi.org/10.1037/a0023956
Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning, 45*(2), 283–331. https://doi.org/10.1111/j.1467-1770.1995.tb00441.x
Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning, 61*, 1–36. https://doi.org/10.1111/j.1467-9922.2011.00641.x
Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x
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. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129
Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, et al. (Eds.), *CLaSIC* (pp. 721–737). Singapore: National University of Singapore, Centre for Language Studies.
Spada, N. (1997). Form-focussed instruction and second language acquisition: A review of classroom and laboratory research. *Language Teaching, 30*(2), 73–87. https://doi.org/10.1017/S0261444800012799
Strong, B., & Boers, F. (2019). Weighing up exercises on phrasal verbs: Retrieval versus trial-and-error practices. *The Modern Language Journal, 103*(3), 562–579. https://doi.org/10.1111/modl.12579
Swain, M. (1993). The Output Hypothesis: Just speaking and writing aren't enough. *The Canadian Modern Language Review, 50*, 158–164. https://doi.org/10.3138/cmlr.50.1.158
Swain, M. (1995). Three functions of output in second language learning. In G. Cook, & B. Seildlhofer (Eds.), *Principles and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford: Oxford University Press.
Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty, & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64–81). New York: Cambridge 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

Chapter 3:

Lag effects in grammar learning: A desirable difficulties perspective

Published as:

Serfaty, J., & Serrano, R. (2022). Lag effects in grammar learning: A desirable difficulties perspective. *Applied Psycholinguistics, 43*(3), 513–550. https://doi.org/10.1017/S0142716421000631

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

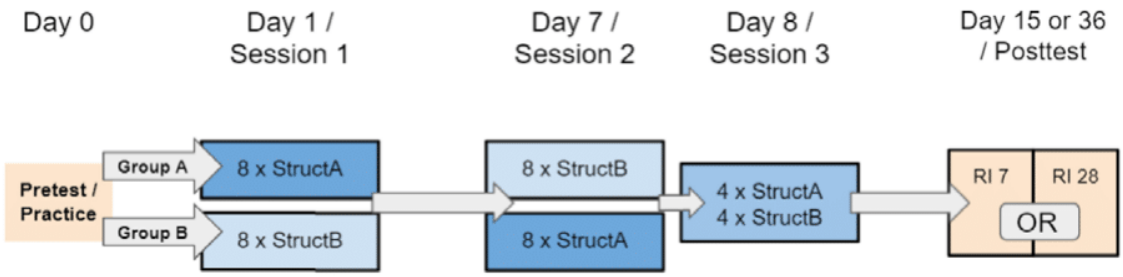


Figure 1. Experimental design.

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

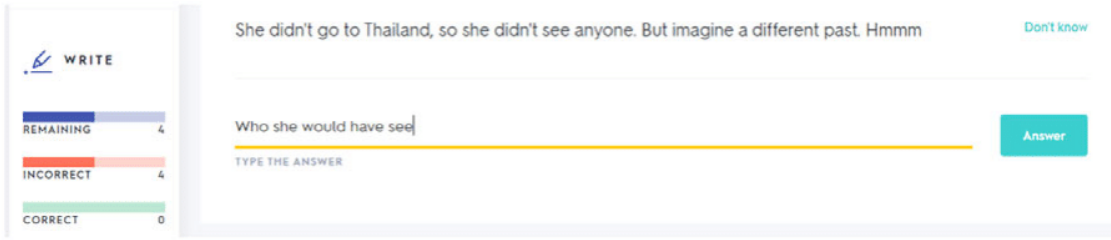


Figure 2. Participants attempt to type the target response.

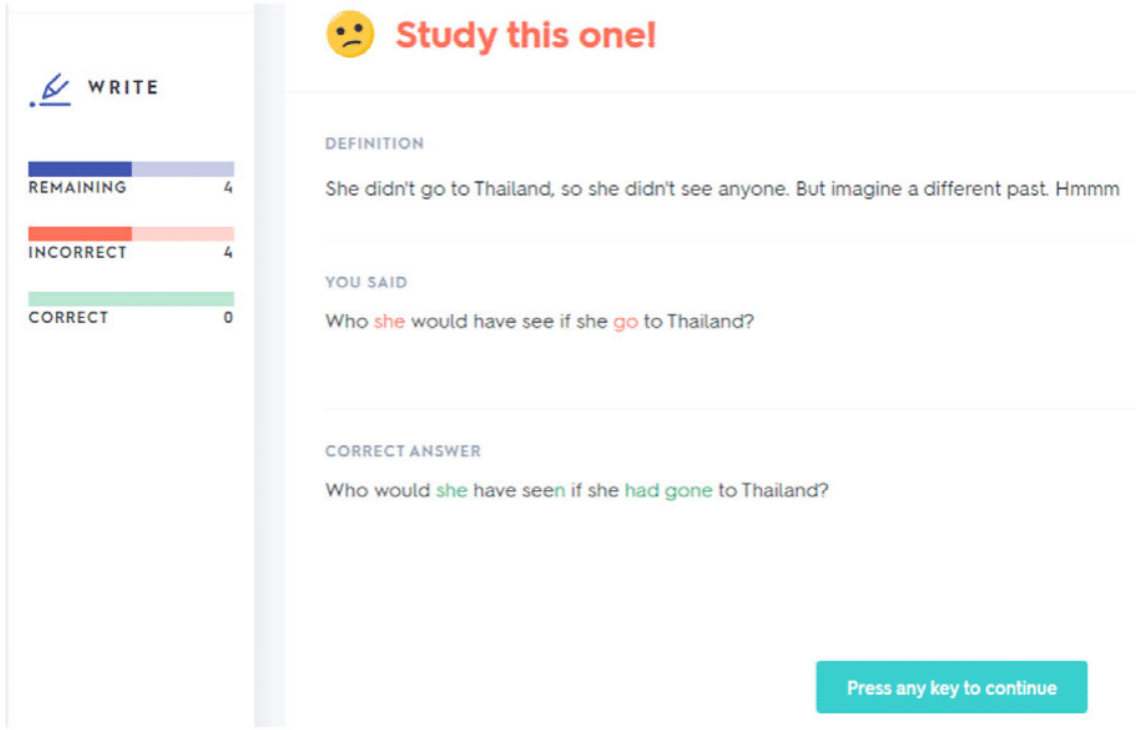


Figure 3. Participants receive feedback on incorrect responses.

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

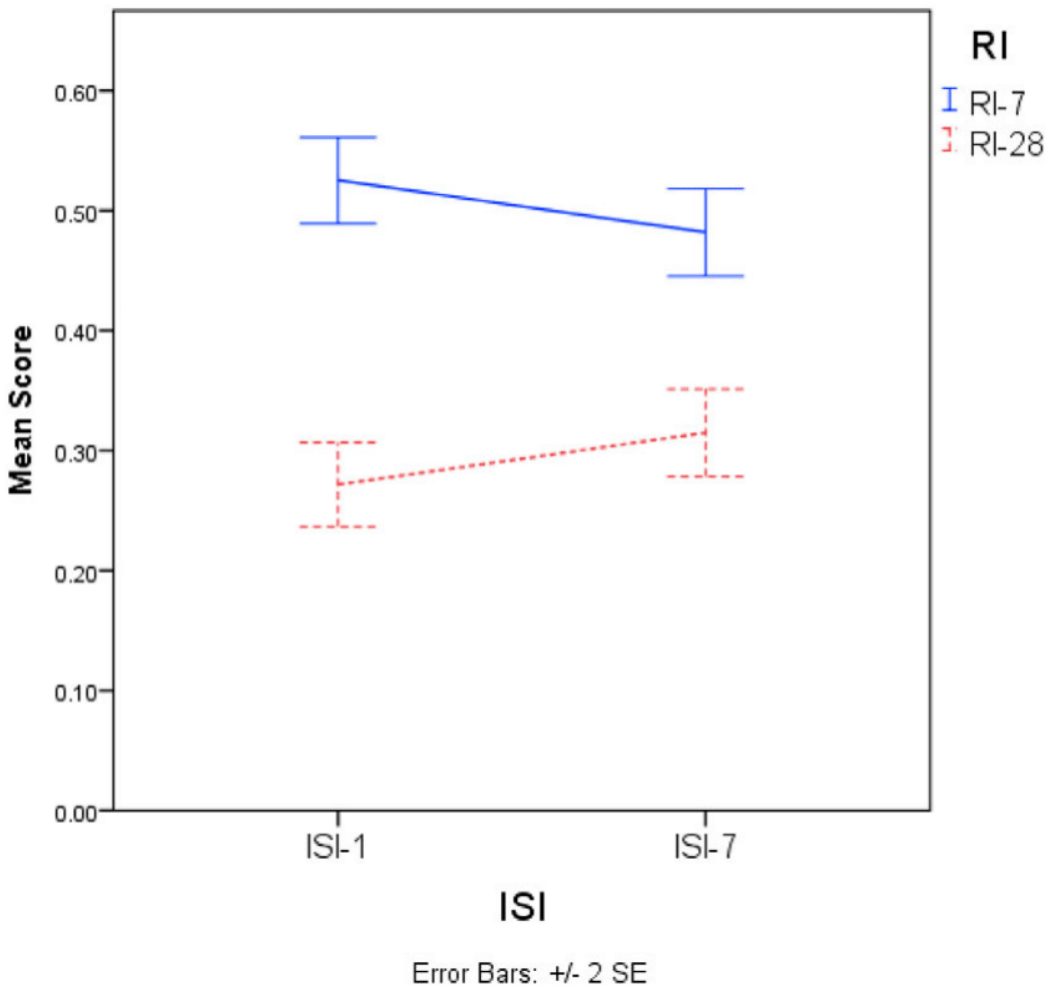


Figure 4. Model 1: ISI by RI interaction.

$M = .585$, $SE = .040$; 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

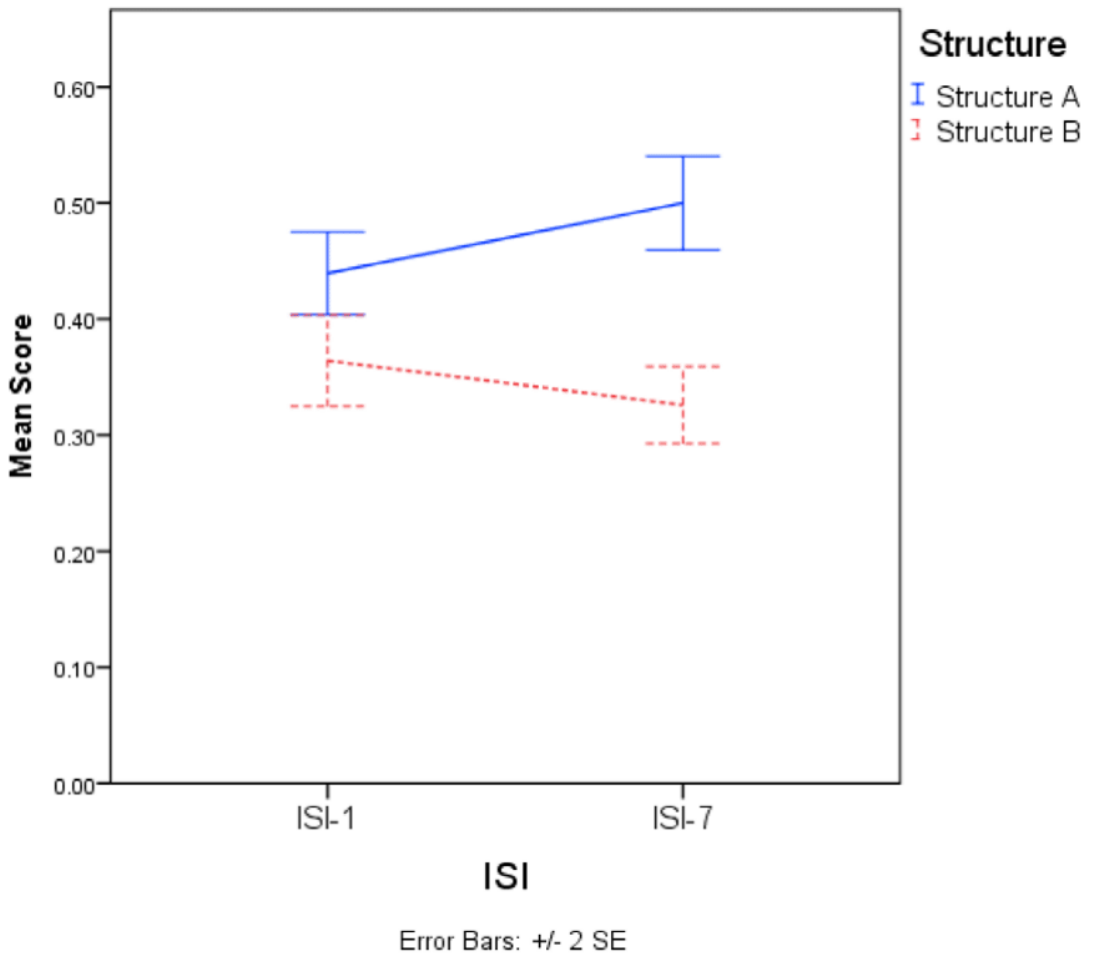


Figure 5. Model 2: ISI by structure interaction.

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

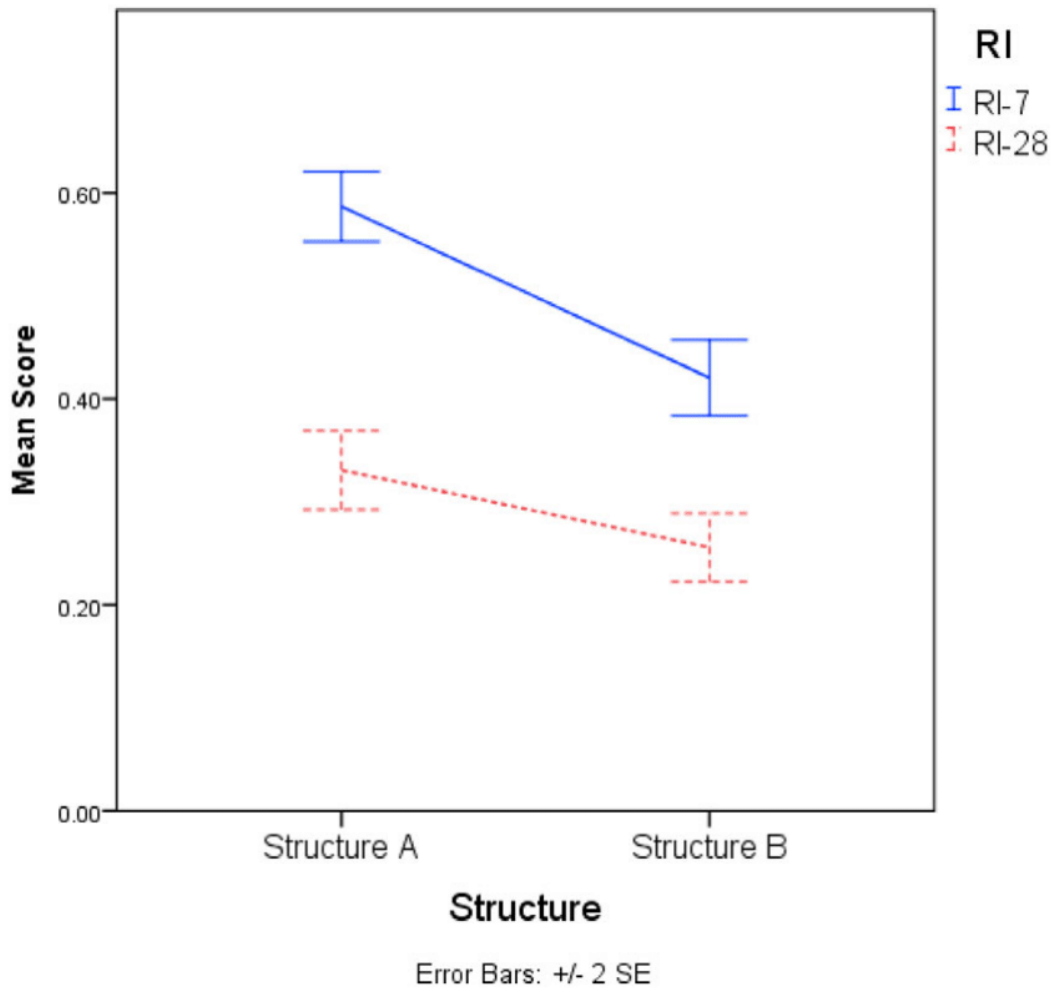


Figure 6. Model 2: RI by structure interaction.

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

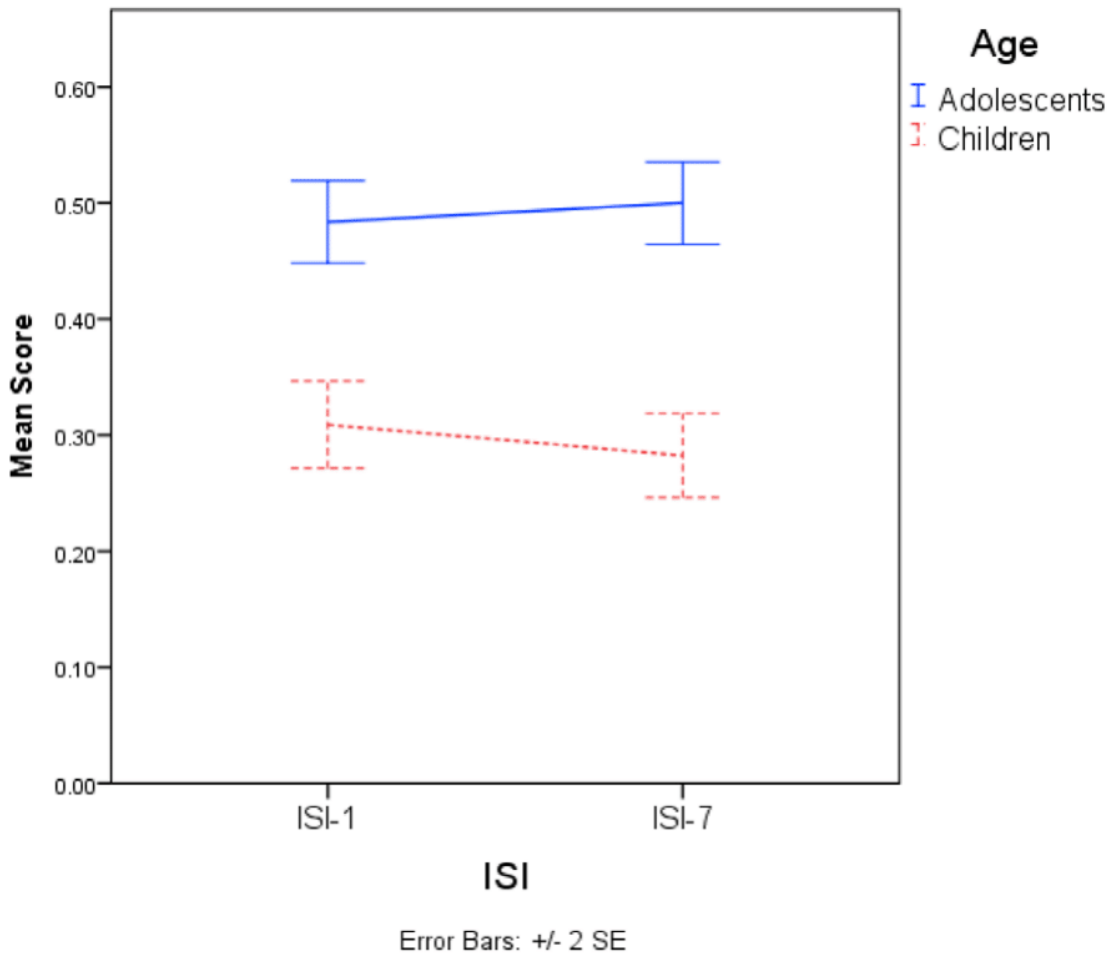


Figure 7. Model 3: ISI by age interaction.

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

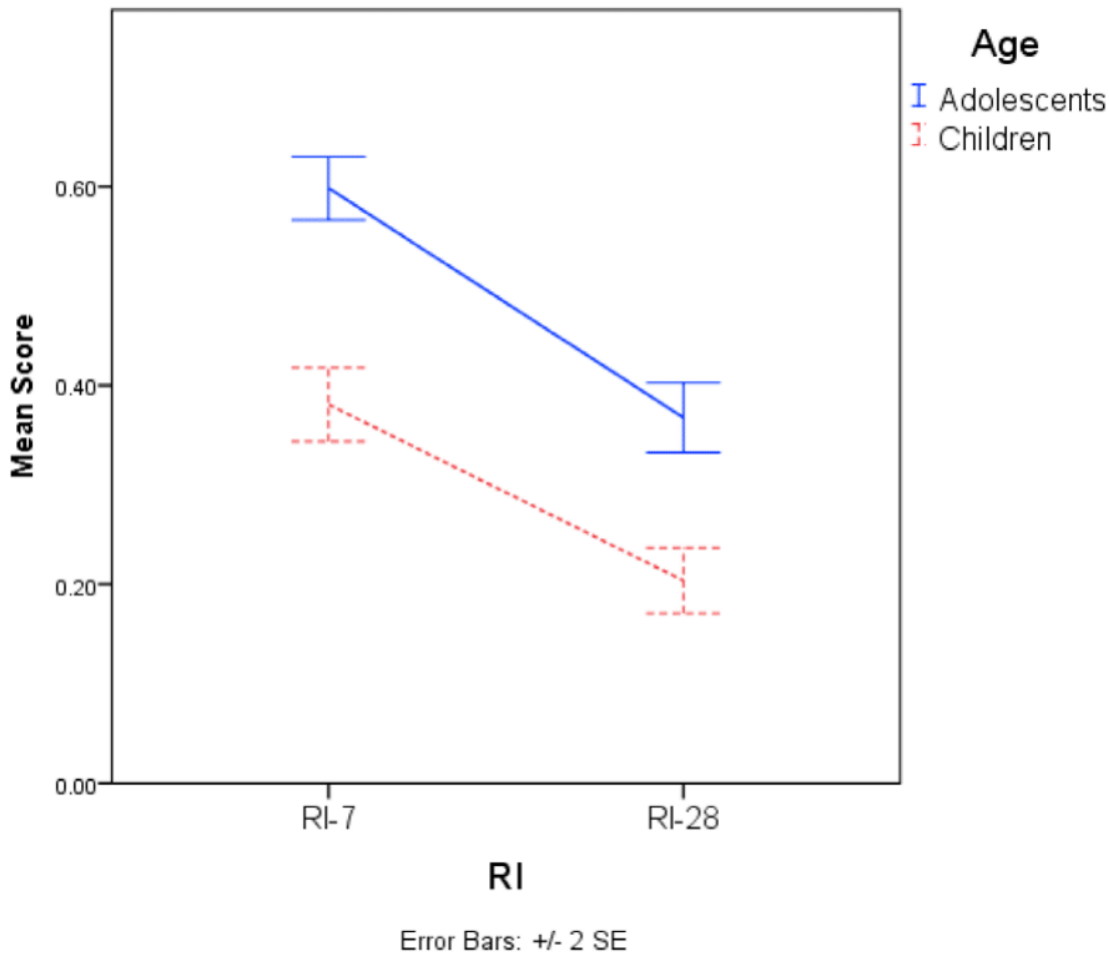


Figure 8. Model 3: RI by age interaction.

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

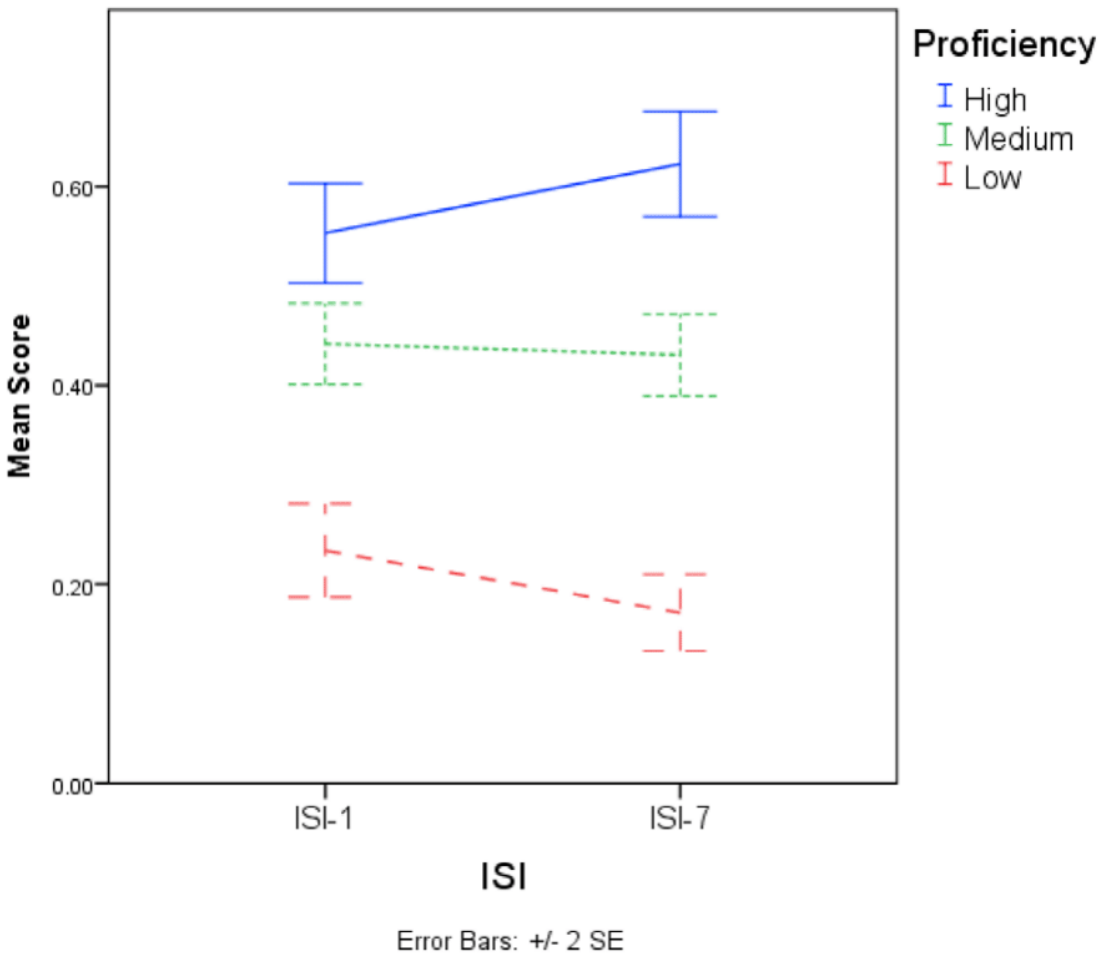


Figure 9. Model 4: ISI by proficiency interaction.

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

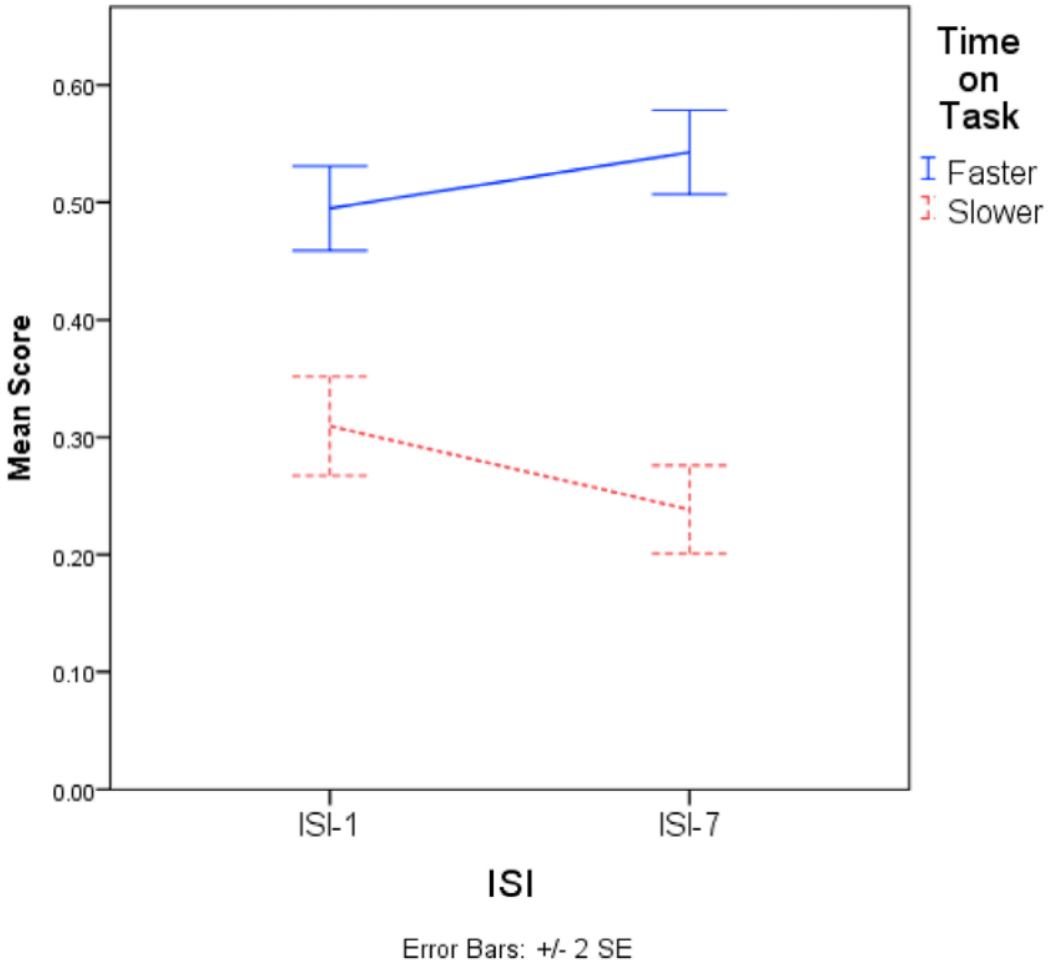


Figure 10. Model 4: ISI by time on task interaction.

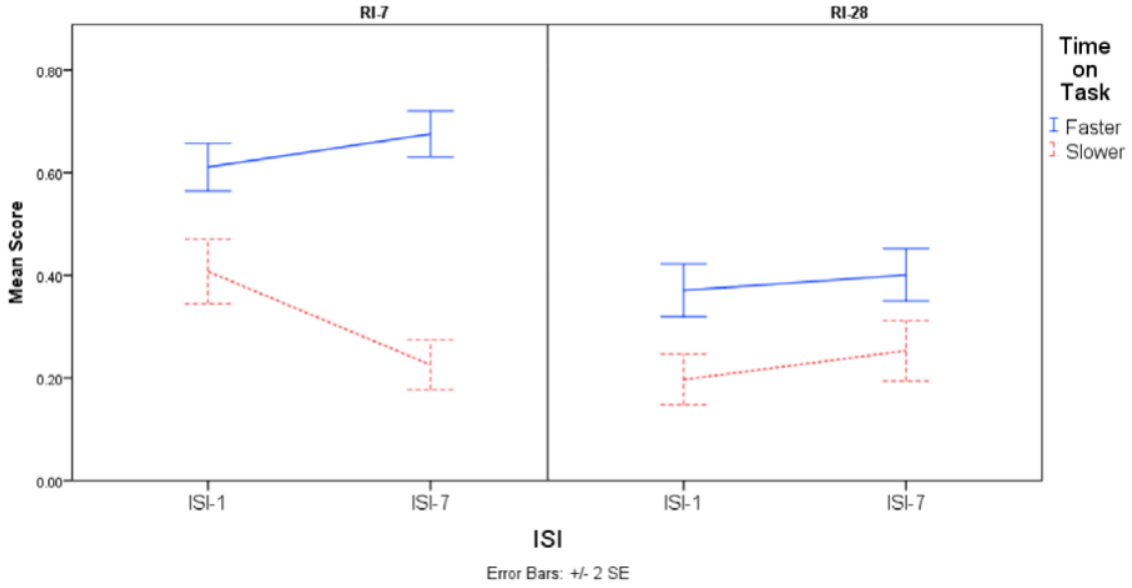


Figure 11. Model 5: ISI by time on task interaction at RI-7 and RI-28.

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 counter-intuitive. 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

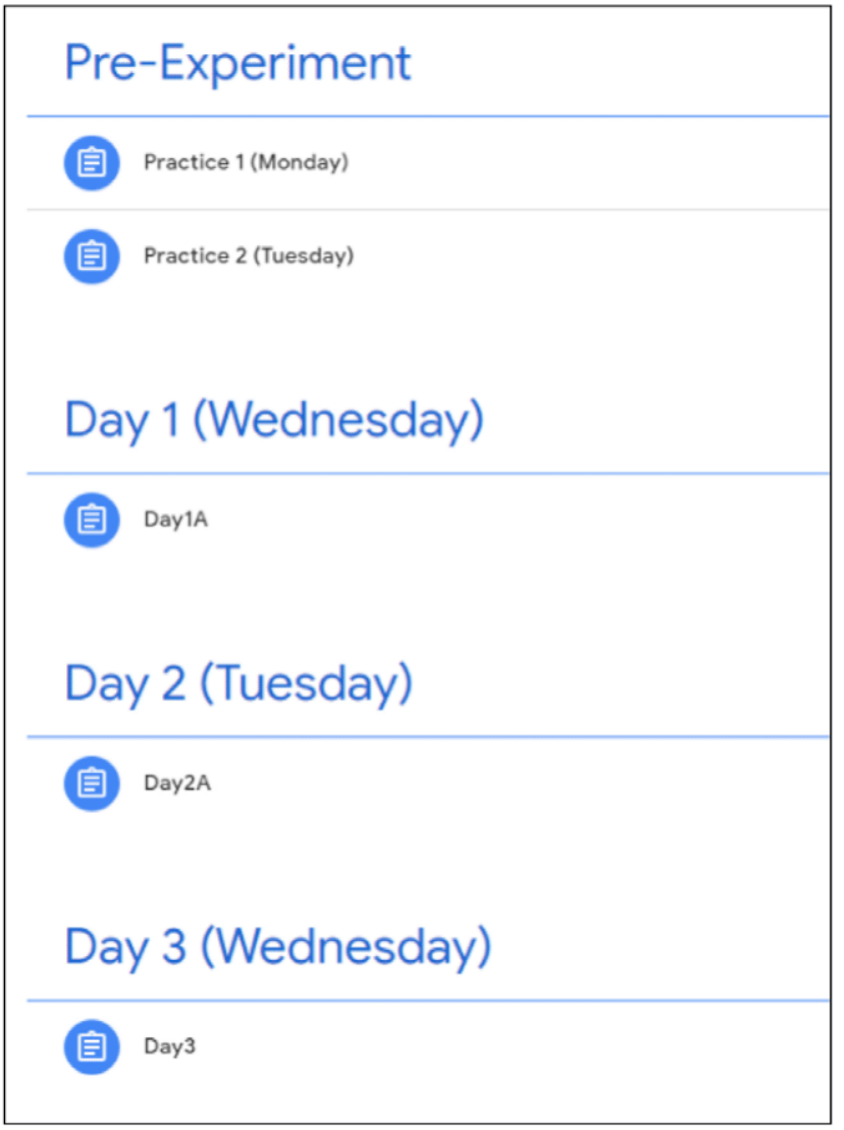
Test Items:

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

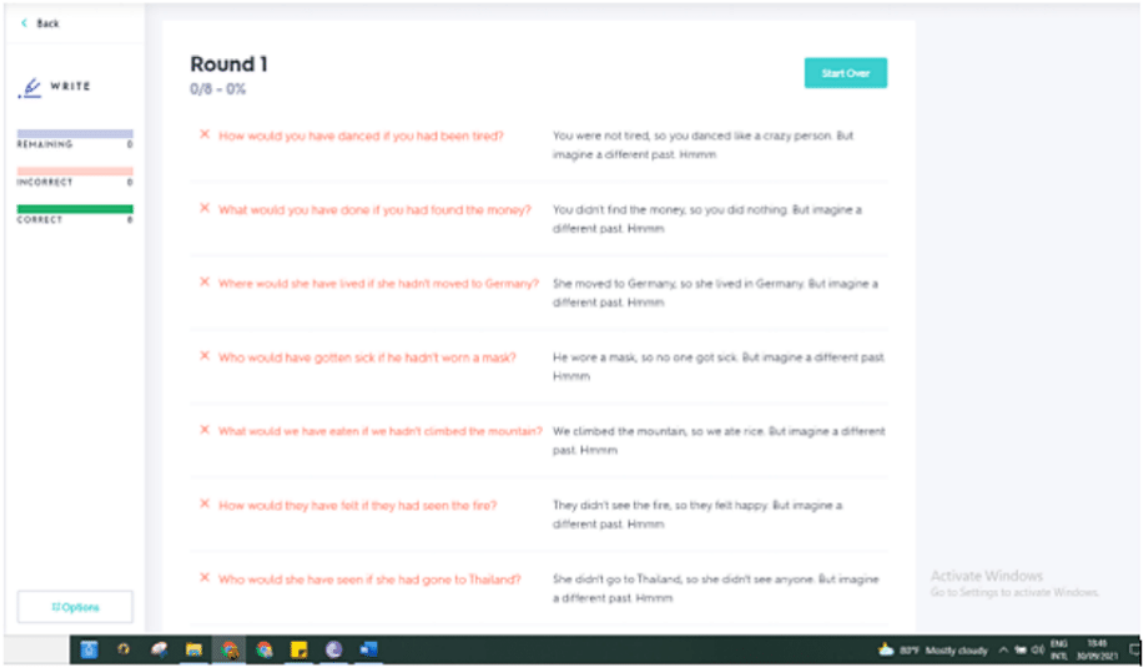
Participants saw their assignments in a Google Classroom. Each assignment only appeared at the appropriate time.

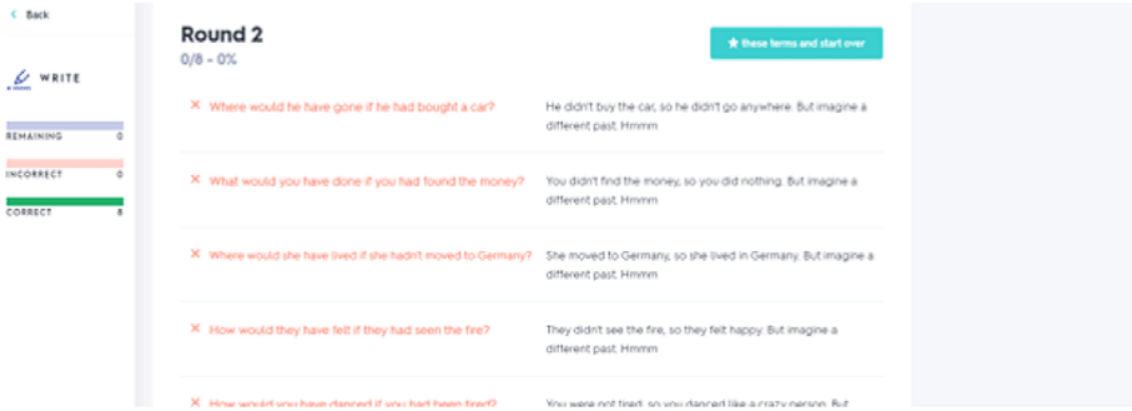


https://doi.org/10.1017/S0142716421000631 Published online by Cambridge University Press

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

Chapter 4:

The optimal scheduling of Quizlet sessions for L2 vocabulary learning

In review as:

Serfaty, J., & Serrano, R. (in review). The optimal scheduling of Quizlet sessions for L2 vocabulary learning. *Language Learning and Technology.*

The optimal scheduling of Quizlet sessions for L2 vocabulary learning

ABSTRACT

Digital flashcard apps allow students to learn and practice foreign language vocabulary independently and efficiently, allowing more time for communicative activities in the classroom. However, words learned this way are at risk of being forgotten. Previous lab studies have shown that vocabulary retrieval practice can be optimized for long-term memory by employing longer intersession intervals, but this effect has not been shown in classroom conditions. The present study investigated the optimal scheduling of independent vocabulary study using Quizlet. Secondary-school students (n = 96, mean age = 13.44) learned 16 novel words in an unknown language over two sessions, spaced at either a 1-day or 1-week interval. Their productive and receptive knowledge was tested after 7 or 28 days. The results show that longer spacing was beneficial for vocabulary learning, contrary to previous findings reported in classroom settings that used a variety of different approaches. The effect was small, but significantly larger on receptive tests, suggesting that the lag effect is dependent upon the kind of knowledge.

INTRODUCTION

One challenge in learning a foreign language (L2) is that the number of hours of exposure tends to be limited, and often restricted to the classroom (Lightbown, 2014). Moreover, L2 vocabulary knowledge is susceptible to forgetting if not sufficiently practiced (Pavlik & Anderson, 2005). Considering this, it is crucial to investigate how to optimize this limited time for the best long-term retention of knowledge.

Research has shown that paired-associate learning is an efficient way to learn new words quickly (Elgort, 2011; Fitzpatrick, et al., 2008; Nation, 2001; Webb, 2009). In second language acquisition (SLA), paired-associate learning is most commonly associated with vocabulary flashcards. Users see a cue, such as a first language (L1) translation, and attempt to retrieve the L2 word from memory, or vice versa. In contrast to traditional paper flashcard drills, digital flashcards offer a wide range of features to foster deeper processing. For example, they can be used to elicit written output with tailored feedback, test items until they have been produced correctly within a session, provide audio to clarify pronunciation, and motivate students through gamification. Flashcard sets can be assigned as homework to reduce classroom time devoted to vocabulary teaching, facilitating fluency and comprehension in subsequent classroom activities that require the target words. Moreover, in contexts that lack well-trained teachers, digital flashcards apps can constitute a reliable source of L2 input and feedback (Serfaty & Serrano, 2020).

Even when engaging in this type of activity, students are likely to forget words learned in a single session. However, research has shown that repeating sessions on multiple days has a powerful effect on long-term memory, even when controlling for the amount of time on task (Rawson et al., 2018), and that the optimal distribution of these relearning sessions can enhance retention further (Gerbier & Toppino, 2015). Several studies from cognitive psychology have shown that longer intervals between sessions promote long-term retention more than shorter intervals (Cepeda et al., 2006; Cepeda et al., 2009), known as *the lag effect.* However, this lag effect has not been consistently found in SLA research. Some studies involving grammar

learning (Suzuki, 2017; Suzuki & DeKeyser, 2017) and all studies involving L2 learning in the classroom with children (Kasprowicz et al., 2019; Rogers & Cheung, 2020a, 2020b) and teenagers (Küpper-Tetzel et al., 2014; Serrano & Huang, 2018, 2021) have reported no advantage to a longer interval.

On the other hand, the lag effect has been reported for vocabulary learning from lab studies involving the retrieval of paired-associates (Bahrick, 1979; Bahrick et al., 1993; Li & DeKeyser, 2019), which is the method employed by digital flashcard apps. Therefore, it is feasible that vocabulary learning through digital flashcards would also be optimized with longer lags between sessions. However, no previous study has investigated whether paired-associate retrieval is subject to lag effects under ecologically valid conditions.

In order to shed light on this issue, we conducted a study in which secondary school students learned novel vocabulary pairs through Quizlet, a popular flashcard app already widely used in classrooms. Words were learned over two sessions, spaced either one day or one week apart, and tested after either one week or one month. The findings are expected to fill an important gap in our understanding of the lag effect in classroom settings while also providing guidance as to the optimal scheduling of digital flashcards for vocabulary learning. This paper further contributes to the field by examining the difference in lag effects on productive and receptive knowledge, which has not yet been explored.

LITERATURE REVIEW

Digital Flashcards for Vocabulary Learning

Flashcards have traditionally been paper cards designed for self-testing. Retrieving information through testing is known to build memory more than re-reading the same information (Barcroft, 2007; Carrier & Pashler, 1992; Kang, 2010; Kang et al., 2013; Kornell & Vaughn, 2016). Online flashcard apps include useful features such as smart feedback that highlights the user's errors, helping them to notice the difference between their attempt and the target (Izumi & Bigelow, 2000; Izumi et al., 1999; Zalbidea, 2019) as well as audio to clarify the target pronunciation. Flashcard software commonly employs criterion learning, repeating items in a cycle until they are answered correctly. Consequently, more practice is automatically allotted to more difficult items. Survey data has shown that digital flashcards are popular with students in educational contexts (Altiner, 2019; Stroud, 2014; Zung et al., 2022). Quizlet in particular is widely used by both teachers and researchers (Franciosi et al., 2016; Korlu & Mede, 2018; Sanosi, 2018; Serfaty & Serrano, 2022; Stroud, 2014), with 60 million users (Quizlet, 2022).

The goal of flashcard assignments might be to familiarize students with useful words. For example, in order to comprehend a text without assistance, most words should already be known (98% according to Hu & Nation, 2000). Teachers could assign content-specific vocabulary in preparation for an upcoming reading or listening passage (Webb, 2009), allowing more classroom time for comprehension activities or communicative language practice. Alternatively, learners could focus on the most frequently occurring

words of the L2. Nation (2006) estimated that 3,000 word families would cover 95% of spoken English. Similarly, students could study academic words in preparation for L2-medium academic studies (Coxhead, 2000).

A second reason for using flashcards would be to collect and practice previously encountered words. Without maintenance, declarative knowledge such as L2 vocabulary decays quickly (Kim et al., 2013; Ullman & Lovelett, 2018) and many curricula do not adequately recycle vocabulary (Tschichold, 2012). Using digital flashcard software, the learner or teacher could cumulatively add words to sets as they are encountered. As long as these sets are periodically practiced, this would counteract the lack of repetition in the curriculum and prevent the forgetting of under-used vocabulary items (Nakata et al., 2021).

A third use for digital flashcards would be to provide individualized work. In mixed-level classes, students can be assigned different sets depending on their abilities and interests. Teachers may also allow faster students to practice flashcards while slower students receive more attention from the teacher. In cases where teachers cannot be present, for example during the recent online learning periods due to COVID-19, students could engage in output practice with reliable feedback without a teacher. This is especially important for students in under-developed educational systems, where teachers may not have a reliable L2 proficiency (Serfaty & Serrano, 2020).

Research suggests that the most efficient way to use flashcards is to retrieve the L2 word from an L1 translation, referred to as productive recall (Nakata, 2020). Although the cue could be something different like a synonym or an image, L1 translations are the most effective (Joyce, 2018; Laufer &

Shmueli, 1997; Lotto & de Groot, 1998). The reverse direction, translation from the L2 into the L1, or receptive recall, is better in terms of words learned per minute. However, productive recall is best for overall gains, especially when knowledge is measured with productive recall tests (Griffin & Harley, 1996; Nakata & Webb, 2016; Webb, 2005, 2009). Productive practice can also reduce forgetting and retraining time over repeated sessions (Schneider et al., 2002). This may be related to the high levels of effort (Pyc & Rawson, 2009) or user-involvement (Hulstijn & Laufer, 2001) in productive practice. By requiring users to produce the L2 word, with perfect orthography, flashcard training guarantees that a significant level of attention has been paid to the target word, which is a crucial step towards long-term acquisition (Leow, 2015; Schmidt, 1990, 2010).

The Lag Effect in L2 Paired-Associate Learning Under Lab Conditions

As with other domains of learning, it has been shown that distributing the practice of L2 vocabulary over multiple sessions (spaced) is better for long term memory than the same amount of practice in a single uninterrupted session (massed), known as *the spacing effect*. The spacing effect has been reported for a wide variety of knowledge and skills (Cepeda et al., 2006; Donovan & Radosevich, 1999), including for L2 vocabulary (Koval, 2019; Nakata, 2015; Nakata & Elgort, 2021) and grammar (Miles, 2014). Within a single session, more spacing between repetitions of the same item has led to better scores in a posttest (Nakata & Webb, 2016). Whether more spacing *between* sessions, i.e. a longer intersession interval (ISI), leads to longer retention is less clear.

Very few studies have tested different ISIs for studying L2 vocabulary over multiple days. In a landmark study by Bahrick (1979), subjects learned vocabulary pairs to criterion at ISI-1 or ISI-30. The ISI-1 group remembered more during training, but after a retention interval (RI) of 30 days from the final training session, the ISI-30 group had retained considerably more knowledge. Bahrick et al. (1993) also found a lag effect for criterion learning over a scale of years. These studies were limited by their inclusion of few participants and very long intervals that would not generalize to authentic classroom procedures. Li and DeKeyser (2019) also demonstrated a lag effect for vocabulary retention, using more participants and more pedagogically relevant intervals. Studying Mandarin words at ISI-1 (daily) or ISI-7 (weekly), retention was similar when tested seven days after training (RI-7), but at RI-28, more words were remembered from the ISI-7 condition. In contrast, studies that have compared lags at a proportionately shorter RI have not found this effect. Bahrick and Hall (2005) found no difference between ISI-1 and ISI-14 at RI-14, and Cepeda et al. (2009), who used a range of ISIs from 0 to 14 days, found no significant differences between ISIs of one day or more when tested at RI-10. It has therefore been claimed that the advantage of a longer lag only emerges at a suitably long RI (e.g., Bird, 2010) and that the ISI should be around 10-30% of the RI (Cepeda et al., 2008).

Of the studies that used paired-associate learning, only Bahrick (1979) used productive recall for training and testing. When tested again eight years later (Bahrick & Phelps, 1987), the longer ISI was better on a productive recall test (L1-L2 translation), but not on a productive recognition test in which subjects saw an English word and selected from five L2 Spanish options.

These findings indicate that lag effects may differ depending on the type of knowledge being tested. Li and DeKeyser (2019)'s vocabulary test was also productive, using pictures as cues. The other studies (Bahrick et al., 1993; Bahrick & Hall, 2005; Cepeda et al., 2009) used receptive recall (L2-L1 translation) for both training and testing. This difference in practice and testing directions may confound comparisons between studies (Edmonds et al., 2021). To our knowledge, no study has used both productive and receptive tests after different ISIs.

Several different accounts have been put forward to explain the lag effect. The reminding account holds that more time between encounters makes retrieval more effortful (Pyc & Rawson, 2009; Koval, 2022). This effort provides desirable difficulty and enhances learning (Bjork, 1994; Suzuki et al., 2019). Alternatively, the reconsolidation account (Smith & Scarf, 2017) focuses specifically on multi-day ISIs and explains the advantage of a longer lag through a greater degree of consolidation. When retrieved, a more consolidated memory trace is more effectively reconsolidated. Both of these accounts hold that if an item is completely forgotten, knowledge cannot be reinforced. It is therefore desirable to schedule a second session with the longest possible ISI before an item cannot be retrieved. A shorter ISI may allow more items to be retrieved, but a longer ISI makes retrievable items more durable.

The Lag Effect in L2 Vocabulary Classroom Studies

Limited research has also addressed lag effects for L2 vocabulary learning in the classroom and, to our knowledge, no advantage to a longer lag

has been reported. Examining assisted repeated reading among 16 year olds, Serrano and Huang (2018) found equal results from ISI-1 and ISI-7 on incidental vocabulary learning and an advantage to ISI-1 in a partial replication involving intentional learning (Serrano & Huang, 2021). For the learning of vocabulary pairs among 11-13 year olds, Küpper-Tetzel et al. (2014) found ISI-1 and ISI-10 to both be better than massed learning, with no significant differences between the two ISI conditions at the delayed posttest. Rogers and Cheung (2020a, 2020b) examined the learning of L2 vocabulary among children aged 8-9. The studies found no benefit for the longer lag (ISI-8), with even a slight advantage to the shorter lag (ISI-1) in one study. All these studies used ISIs within 10-30% of the RI, so a lag effect could have been expected.

The above-mentioned classroom experiments differ from digital flashcard learning in several ways. Firstly, flashcards employ criterion learning. Incorrectly answered items remain in the cycle to be attempted again in a subsequent round of retrieval attempts. The session only ends when all items have been retrieved successfully. Therefore, a repeated session serves to remind learners of already-learned knowledge. In contrast, previous classroom studies controlled for the amount of practice time, but not for the achievement of the learner within a session. Words may not be fully learned within the first session, making it difficult to classify the second session as a relearning event. Secondly, classroom studies are interactive, involving multiple learners and an instructor, as opposed to online flashcards that involve one learner guided by software. Classroom studies have also used a variety of training tasks, even within studies (e.g., picture quizzes, animations) as well as a variety of testing

formats (vocabulary matching test, crossword puzzles). These human and task-related factors may lead to less experimental control and less comparability between studies.

The scheduling of vocabulary flashcard learning in a classroom has only been investigated for university students and only in terms of an expanding versus a uniform ISI, rather than the length of the ISI itself (Schuetze, 2015; Schuetze & Weimer-Stuckmann, 2011). No studies have yet provided insights into the optimal ISI for vocabulary flashcard training under classroom conditions or for secondary school learners. The only previous study in this area used full-sentence items as flashcards with the aim of learning grammatical accuracy (Serfaty & Serrano, 2022). In this case, the longer ISI-7 was only better than ISI-1 for students with high L2 proficiency or fast completion times. ISI-1 was better for participants who found the task more challenging. Essentially, the more difficult ISI-7 added desirable difficulty when the task was not already too difficult. There was also a trend towards ISI-7 being better for the simpler grammatical structure but detrimental to the more complex structure. While grammar learning involves applying rules for a single complex structure, vocabulary learning involves retrieving many independent items. Therefore, it is unclear to what extent findings from a grammar-learning experiment would apply to vocabulary learning.

THE PRESENT STUDY

The present study explored lag effects for L2 vocabulary learning through digital flashcards. Participants retrieved new words over two sessions,

with either one or seven days between sessions, and their retention was assessed after one week or one month. The study represents an important contribution to the field for several reasons. First, by using paired-associate vocabulary learning, this study clarifies whether the lag effect from lab studies among adults applies to a younger demographic and whether the absence of a reported lag effect in previous studies for this age group is due to task factors or age-related factors. Second, by using an ecologically valid tool, Quizlet, as it would be used in authentic classroom conditions or by independent learners, our findings can be used to provide recommendations as to the optimal scheduling of L2 vocabulary sets. Third, in contrast to previous research in this area, the present study used both a productive and receptive recall test. Inconsistencies in previous research may partially be due to the use of either productive or receptive tests. The former taps the ability to generate the L2 form in speaking or writing, whereas the latter only tests the ability to comprehend the L2 word when it is encountered through listening or reading. Receptive knowledge is known to develop before productive knowledge and therefore represents a lower level of mastery of the L2 word (González-Fernández & Schmitt, 2020). In order to disentangle this potentially confounding factor, it is imperative to explore how lag effects could vary between these two kinds of knowledge. Finally, although not a specific research question in this paper, our experiment used the same design as a grammar experiment (Serfaty & Serrano, 2022), in terms of the tool, ISI, RI, and number of items. Moreover, the two experiments were conducted in the same setting and many of the participants involved took part in both experiments. Consequently, a valid comparison of lag effects for grammar and

vocabulary learning can be made without the confounds of task differences. Our research questions (RQs) are as follows:

RQ1: Is there a lag effect for L2 vocabulary learning through digital flashcards under classroom conditions?

RQ2: Is the lag effect different for productive and receptive knowledge?

For RQ1, we hypothesized that a lag effect would be found, despite it not being found in previous classroom vocabulary studies, because paired-associate learning to criterion has produced a lag effect under lab conditions with the same intervals. No hypothesis was made for RQ2 since this issue has not previously been explored.

METHODOLOGY

Participants

Participants came from an international school in Cambodia. All students aged 11-18 were recruited for the training phase of this study on a voluntary basis, as part of a wider project about memory, aptitudes, and learning techniques. Around half of students missed at least one session due to unpredictable school schedules related to COVID-19, and any students that failed to document their learning as required were also excluded from analysis, leaving a total of 96 participants (51 female). The distribution of ages was as follows: $n11 = 20$; $n12 = 16$; $n13 = 15$; $n14 = 13$; $n15 = 16$; $n16 = 12$; $n17\text{-}18 = 4$.

Experimental Design

Target words. The priority in this experiment was to use target words that were previously unknown and to be sure that gains could be solely attributed to the experimental training. In the present sample, English vocabulary sizes varied greatly, necessitating that target words were from an unknown language. Hebrew was chosen because it contains many words with a CVCVC structure with phonology common to English and Khmer. The categories of *animals* and *food* were chosen for their high imageability and familiarity in English. Each category included eight two-syllable nouns of five or six letters (e.g. *kelev* - dog), transliterated into the Latin alphabet (Appendix A).

Training. Using the *Write* mode of Quizlet, participants saw an English cue (e.g. *dog*) with an image, and typed their response in Hebrew. Since there was no presentation stage, participants needed to guess incorrectly on Round 1 in order to see the target words for the first time as feedback. They then continued through the rounds until they had typed all items correctly once (see Appendix B for screenshots).

There were a total of three training sessions (S). Half the participants studied animals in the first session (S1) and food in S2, while the other half did the reverse. In S3, participants studied a combined set of all the target words. S1 and S3 were separated by one week (ISI-7) while S2 and S3 were separated by one day (ISI-1). Figure 1 shows the timing of each session.

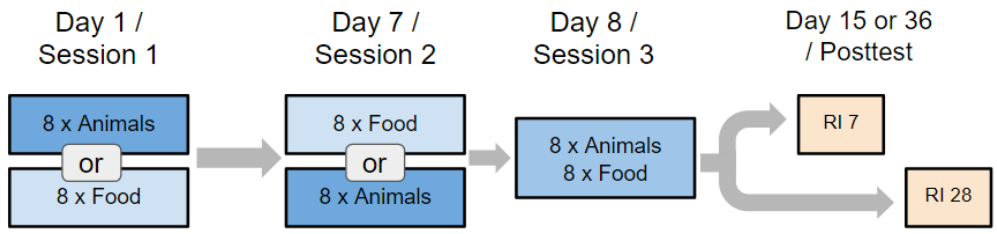
Tests

Participants were tested on all 16 items, firstly through productive recall (L1-L2 translation) and then through receptive recall (L2-L1

translation), as defined by Nakata (2020). Since the possible answers for the receptive test were used as cues in the productive test, some priming was unavoidable. A distraction round of three unrelated questions preceded the receptive test in order to reduce priming effects. Cronbach's alpha showed high internal consistency for *animals* and *food* for productive (.815; .830) and receptive (.746; .792) measures.

To avoid testing effects, RI was a between-subjects variable (Suzuki, 2017). RI-7 and RI-28 were chosen based on their relevance to real school schedules and for comparability with previous research using the same intervals. Based on claims that the optimal ISI is 10-30% of the RI (Cepeda et al., 2008), ISI-1 would be optimal for RI-7 and ISI-7 would be optimal for RI-28.

FIGURE 1
Experimental Design



Procedure

Participants were split alphabetically within grade levels to assign categories to ISIs. RI groups were manipulated after training so that the order of categories was equally represented at each RI, with no statistically significant differences in mean age or time on task (see Appendix C).

All sessions were conducted under conditions in which the students would normally engage in independent study, either in a classroom separated

according to COVID-19 guidelines or from home. Consequently, participants did not interact meaningfully with each other or with the instructor during training or testing. Students were supervised as in normal studying conditions, but were required to follow the prompts of the software independently.

Participants were already familiar with Quizlet. They recorded their time on task and added screenshots of their progress in Google Classroom (see Appendix D). Posttests were completed individually on their assigned days.

ANALYSIS

Posttests were scored one point for each correct response with a possible total of eight points. No ambiguous or partially correct responses were identified. Paired-samples *t*-tests showed no statistically significant differences in posttest scores between categories for productive (animals: M = 2.43, SD = 2.41; food: M = 2.16, SD = 2.38), $t[95] = 1.417$, $p = .160$) and receptive measures (animals: M = 3.79, SD = 2.37; food: M = 3.61, SD = 2.47), $t[95] = 0.775$, $p = .440$).

A generalized linear model with a binomial outcome was performed using SPSS 27 (IBM, 2020), which is suitable for data that is not normally distributed. Participant and item variations were included as random factors. Initially, individual differences of age and time on task were included as covariates, but they had no effect and were removed. The fixed predictors were ISI, RI, and test (productive, receptive), as well as their possible interactions.

The effect size for this model is the odds ratio (*OR*), representing the added likelihood of a correct response in one condition over another. For

example, *OR* = 2.000 implies that a correct response is twice as likely from the condition with the higher mean. The *OR* will be interpreted as small (1.68), medium (3.47), or large (6.71) following Chen et al. (2010). Significance tests were two-tailed and the alpha was set at $p = .005$ with sequential Bonferroni correction.

RESULTS

We first present training data in order to better interpret the results, followed by descriptive and inferential statistics for the posttests. All datasets and syntax can be found online.

Training

Two measures were used to examine participants' training performance: time on task during the learning (S1 and S2) and relearning sessions (S3), and accuracy at the beginning of S3 (see Table 1).

TABLE 1

Minutes on Task and Accuracy at Round 1 of S3 (maximum 8)

Minutes			**S3 Accuracy**	
S1	S2	S3	ISI-1 words	ISI-7 words
7.86 (4.08)	6.53 (6.17)	9.66 (5.09)	1.25 (1.77)	0.65 (1.53)

Time on task was highly variable between participants, based on the SD. On average, participants required less than one minute for each word. The time to complete S3, which included all words from S1 and S2, was less than

the sum of the previous two sessions, indicating that relearning was faster than learning. However, very few words were typed without errors on Round 1 of S3, averaging at just over one out of eight from ISI-1 and less than one from ISI-7. This difference was statistically significant but small, $t[89] = 3.809$, $p < .001$, $d = 0.362$.

Posttest Results

Table 2 displays the descriptive statistics for posttest scores. Overall, scores were quite low, which is unsurprising after only two sessions with relatively long RIs. As expected, receptive scores were higher than productive scores and RI-7 scores were higher than RI-28 scores. Crucially, ISI-7 words were better remembered than ISI-1 words.

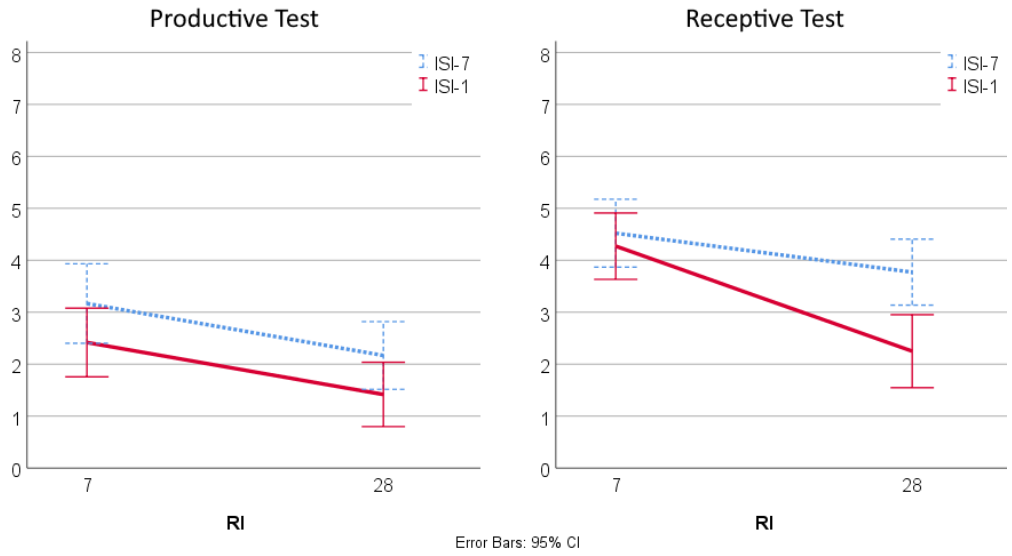
TABLE 2

Posttest Results from Productive and Receptive Tests by ISI (maximum score = 8) and together (maximum score = 16) at RI-7, RI-28 and Overall

	Productive			**Receptive**		
	RI-7	RI-28	Overall	RI-7	RI-28	Overall
ISI-1	2.42 (2.28)	1.42 (2.13)	1.92 (2.25)	4.27 (2.20)	2.25 (2.42)	3.26 (2.51)
ISI-7	3.17 (2.63)	2.17 (2.24)	2.67 (2.49)	4.52 (2.25)	3.77 (2.19)	4.15 (2.24)
All words	5.58 (4.60)	3.58 (4.02)	4.58 (4.41)	8.79 (4.07)	6.02 (4.09)	7.41 (4.29)

FIGURE 2

ISI-1 and ISI-7 Items at RI-7 and RI-28 for Productive and Receptive Tests.



Statistical Model

Full details of the model, including all non-significant means and effect sizes, can be found in Appendix S1. The GLMM produced statistically significant but small main effects for all variables. ISI-7 scores were significantly higher than ISI-1 scores ($p < .001$, $OR = 1.613$), RI-7 scores were significantly higher than RI-28 scores ($p = .004$, $OR = 1.972$), and receptive scores were significantly higher than productive scores ($p < .001$, $OR = 2.188$).

The interaction between ISI and RI was significant. While RI-7 scores were always higher than RI-28 scores, the drop was bigger in the ISI-1 condition ($p = .001$, $OR = 2.425$) but less pronounced in the ISI-7 condition ($p = .046$, $OR = 1.603$). Viewed differently, the difference between the ISI conditions was smaller at RI-7 ($p = .003$, $OR = 1.310$) but larger at RI-28 ($p < .001$, $OR = 1.984$). Thus, retention between the two RIs was better for words learned at ISI-7.

The other two-way interactions (ISI*test and RI*test) were not statistically significant in this model, but there was a significant three-way interaction between all predictor variables. For productive scores, the drop from RI-7 to RI-28 was consistent for words from ISI-1 (p = .017, OR = 2.012) and ISI-7 (p = .030, OR = 1.764). However, for receptive scores, only the drop for ISI-1 words was significant (p < .001, OR = 2.924). For ISI-7, the drop was not significant (p = .124, OR = 1.457). The advantage to ISI-7 words at the longer RI was therefore more pronounced in receptive scores than in productive scores.

DISCUSSION

The present study aimed to investigate the optimal scheduling of vocabulary learning with digital flashcards under conditions applicable to a classroom. Using Quizlet, secondary school students aged 11-18 learned 16 novel foreign words at either ISI-1 or ISI-7 (within-subjects), and were tested at either RI-7 or RI-28 (between subjects), on both productive and receptive measures.

Results showed a small but statistically significant difference between ISI conditions, according to which ISI-7 led to better retention at both RI-7 and RI-28. This contrasts with previous research in several important ways. Firstly, previous classroom research on L2 vocabulary using other types of tasks did not find a lag effect (Küpper-Tetzel et al., 2014; Rogers & Cheung, 2020a, 2020b; Serrano & Huang, 2018, 2021). Secondly, previous SLA studies have only found an advantage to a longer ISI at the longer RI (Bahrick, 1979; Bird, 2010; Li & DeKeyser, 2019), whereas our results showed the lag

effect to be consistent at the shorter and longer RIs for productive measures. Finally, a grammar-learning experiment using Quizlet with the same intervals (Serfaty & Serrano, 2022) found no global advantage to either condition. Instead, ISI-7 was beneficial under easier conditions, in terms of linguistic and learner-related difficulty, and ISI-1 was better for more difficult conditions. Grammar and vocabulary learning therefore appear to be differently affected by lag when controlling for task type and other methodological factors. To interpret this difference, it would be reasonable to assume that single items of vocabulary are simpler to remember than the complex rules of long sentences. This assumption is supported by the much shorter training times in the present study. Therefore, the more difficult ISI-7 would add desirable difficulty to the comparatively simple vocabulary learning task as compared to the complex grammar task. It is noteworthy that both experiments only used productive recall practice, which is more difficult than receptive recall practice. Following our rationale, it is possible that the added difficulty from ISI-7 would have a larger benefit on receptive practice.

The present study also compared productive and receptive vocabulary knowledge. Despite the fact that the training involved productive recall practice, significantly higher scores were obtained on the receptive test, in line with claims that receptive vocabulary knowledge develops earlier and is easier to attain than productive knowledge (Laufer & Goldstein, 2004; González-Fernández & Schmitt, 2020). The advantage to ISI-7 for productive knowledge is theoretically interesting but the difference in scores (18% vs 27% at RI-28) was small. However, in the receptive test, the difference in scores (28% vs 47% at RI-28) would be quite meaningful in an educational

context. Chen and Truscott (2010) also found bigger effects on receptive tests than on productive tests for different quantities of input.

A speculative interpretation could be that for many items, receptive knowledge was retained between sessions, whereas productive knowledge was not. This is similar to a finding from a study by Barclay and Pellicer-Sánchez (2021) in which form-recall knowledge decayed while easier form-recognition knowledge was retained. Only *retained* knowledge can be reinforced through relearning and so receptive knowledge may be more likely to be reinforced. Conversely, productive knowledge of the same words would need to be encoded anew. Some evidence for this is found in the training data, in that words were generally not retrievable productively at the start of S3, but that relearning was faster than learning. This implies that partial knowledge was retained, i.e., receptive knowledge. When the feedback from Round 1 was presented, this perhaps reminded participants of their retained receptive knowledge with a strengthening effect. This would be the case from both ISIs, but words from ISI-7 would be more effortful to retrieve, or better consolidated, leading to more effortful retrieval or stronger reconsolidation of ISI-7 words from the feedback. Therefore, as more words were retained receptively than productively between sessions, the lag effect was stronger for receptive knowledge.

The statistically significant advantage to ISI-7 in the productive test implies that the lag effect could be similar for productive knowledge if it was more developed or better retained between sessions. This could be achieved through adding more sessions (Rawson et al., 2018). It is clear that two sessions of Quizlet for L2 vocabulary, as an isolated activity, are not enough.

Nakata et al. (2021) showed the importance of cumulatively reviewing vocabulary over a long period of time and over multiple sessions in order to avoid forgetting. Rawson et al. (2018) have advocated for more studies involving several sessions that result in higher scores, pointing out that such low scores would not be useful to real students needing to pass exams.

One finding from the training data requires explanation. Current accounts of the lag effect (Koval, 2022; Smith & Scarf, 2017) emphasize that successful retrieval is necessary for a longer lag to have a facilitative effect on retention. However, words from ISI-7 were not typed correctly at the first round of S3 and therefore it could not be claimed that successful retrieval followed the longer lag. Despite this, a lag effect was detected.

One explanation is that through criterion learning, successful retrieval was achieved for all words within all sessions, and that the retrieval of ISI-7 words in S3 was still more effortful than for ISI-1 words, even if that retrieval came in Round 2 or Round 3. This conjecture is supported by the higher retrieval success rate from ISI-1 items in Round 1 of S3. Though only one item on average was recalled, this indicates that ISI-1 items were more accessible in memory, having been learned only one day earlier. Consequently, successful retrieval on Round 2 of S3 could still be more effortful for ISI-7 words, which would then promote better long-term memory.

Alternatively, it is probable that ISI-7 words required more retrieval attempts in S3 than ISI-1 words. Some viewpoints hold that unsuccessful retrievals are also beneficial, priming the learner to pay attention to feedback (Kornell & Vaughn, 2016). For Quizlet in particular, if a word is typed correctly, the user does not see it in feedback. Therefore, more incorrect

responses elicit more visual feedback and more retrieval attempts, which could have reinforced memory (Nakata, 2017; Webb, 2007). Barclay and Pellicer-Sánchez (2021) found that more attempts to reach criterion in productive flashcard learning resulted in better retention at their RI-28 form-recognition test. From another perspective, Bahrick and Hall (2005) argued that unsuccessful attempts prompt the learner to identify bad mnemonic strategies. If a word is easy to retrieve due to a short lag, a bad strategy might not be detected, but an unsuccessful retrieval attempt may prompt the learner to develop a better strategy.

LIMITATIONS AND FUTURE DIRECTIONS

The present study targeted L2 classroom learning with an aim of providing pedagogical recommendations. In order to maintain ecological validity, Quizlet was chosen as the tool under examination. However, Quizlet could not provide precise insights into the learning process. A partial replication of the present design using a research-focused tool, such as Gorilla, could facilitate deeper insights into learning processes by tracking learners' response times on successful retrievals and the number of trials required to reach criterion per word. These indicators could confirm our speculation that more effort was induced in retrieving words from the longer lag or that more retrieval attempts led to better retention.

A further limitation of the present study is that the receptive test came after the productive test, using the same items, and that participants only engaged in productive practice during training. It would be interesting to compare lag effects for both productive and receptive practice, using enough

target words to test both productive and receptive knowledge without repeating items. Additionally, a future study could conduct a productive and receptive test without feedback at the beginning of S3 in order to compare how much of each kind of knowledge was retained after different lags.

CONCLUSIONS AND PEDAGOGICAL IMPLICATIONS

The present paper has reported an experiment in which secondary school students used Quizlet to study unknown words over two sessions, using productive recall, with either one day or one week between sessions. The longer interval promoted better retention of learned items, especially on receptive measures. To our knowledge, this is the first study that has confirmed the lag effect for L2 vocabulary learning in a secondary school context and the first reported difference of lag effects for productive and receptive measures. These findings have direct pedagogical implications. Digital flashcards offer teachers a method of building a baseline of vocabulary knowledge for individual learners. Although scores were quite low in the present experiment, it would be expected, and recommended, that flashcards are reviewed more than twice in order to preserve memory for longer. Moreover, vocabulary sets should be used as a supplemental activity to meaningful language practice, either beforehand to pre-learn key vocabulary, or afterwards to prevent forgetting. Most importantly, our results suggest that teachers should schedule these sessions at multi-day intervals, rather than repeating the same set on consecutive days. This should help students to remember what they study for longer and reduce the amount of sessions required to build reliable and durable L2 vocabulary knowledge.

REFERENCES

Altıner, C. (2019). Integrating a computer-based flashcard program into academic vocabulary learning. *TOJET: The Turkish Online Journal of Educational Technology*, *18*(1).

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*(3), 296–308. https://doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*(5), 316–321. https://doi.org/10.1111/J.1467-9280.1993.TB00571.X

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

Barclay, S., & Pellicer-Sánchez, A. (2021). Exploring the learning burden and decay of foreign language vocabulary knowledge. *ITL - International Journal of Applied Linguistics*, *172*(2), 259–289. https://doi.org/10.1075/itl.20011.bar

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/J.1467-9922.2007.00398.X

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.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. https://doi.org/10.3758/bf03202713

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

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation*, *39*(4), 860–864. https://doi.org/10.1080/03610911003650383

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*(5), 693–713. https://doi.org/10.1093/APPLIN/AMQ031

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238. https://doi.org/10.2307/3587951

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. Journal of *Applied Psychology*, *84*(5), 795–805. https://doi.org/10.1037/0021-9010.84.5.795

Edmonds, A., Gerbier, E., Palasis, K., & Whyte, S. (2021). Understanding the distributed practice effect and its relevance for the teaching and learning of L2 vocabulary. *Lexis*, *18*. https://doi.org/10.4000/lexis.5652

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367–413. https://doi.org/10.1111/J.1467-9922.2010.00613.X

Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: A case study. *Language Learning Journal*, *36*(2), 239–248. https://doi.org/10.1080/09571730802390759

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

Gerbier, E., & Toppino, T. C. (2015). The effect of distributed practice: Neuroscience, cognition, and education. *Trends in Neuroscience and Education*, *4*(3), 49–59. https://doi.org/10.1016/j.tine.2015.01.001

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Griffin, G., & Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, *17*(4), 443–460. https://doi.org/10.1017/S0142716400008195

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430. https://doi.org/10.26686/wgtn.12560354

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Izumi, S., & Bigelow, M. (2000). Does output promote noticing and second language acquisition? *TESOL Quarterly*, *34*(2), 239-278. https://doi.org/10.2307/3587952

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition*, *21.* 421-452.

Joyce, P. (2018). L2 vocabulary learning and testing: the use of L1 translation versus L2 definition. *The Language Learning Journal*, *46*(3), 217–227. https://doi.org/10.1080/09571736.2015.1028088

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition*, *38*(8), 1009–1017. https://doi.org/10.3758/MC.38.8.1009

Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review*, *20*(6), 1259–1265. https://doi.org/10.3758/s13423-013-0450-z

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, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science, 14*(1), 22-37. https://doi.org/10.1080/1464536X.2011.573008

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., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation*, *65*, 183–215. https://doi.org/10.1016/BS.PLM.2016.03.003

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*(5), 1103–1139. https://doi.org/10.1017/S0142716419000158

Koval, N. G. (2022). Testing the reminding account of the lag effect in L2 vocabulary learning. *Applied Psycholinguistics*, *43*(1), 1–40. https://doi.org/10.1017/S0142716421000370

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

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*(1), 89–108. https://doi.org/10.1177/003368829702800106

Leow, R. P. (2015). Explicit learning in the L2 classroom: A student-centered approach. New York: Routledge. https://doi.org/10.4324/9781315887074

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

Lightbown, P. M. (2014). Making the minutes count in L2 teaching. *Language Awareness*, *23*(1–2), 3–23. https://doi.org/10.1080/09658416.2013.863903

Lotto, L., & de Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, *48*(1), 31–69. https://doi.org/10.1111/1467-9922.00032

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

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. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, *39*(4), 653–679. https://doi.org/10.1017/S0272263116000280

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

Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, *37*(2), 233–260. https://doi.org/10.1177/0267658320927764

Nakata, T., Tada, S., Mclean, S., & Kim, Y. A. (2021). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly*, *55*(1), 248–270. https://doi.org/10.1002/TESQ.596

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? *Studies in Second Language Acquisition*, *38*(3), 523–552. https://doi.org/10.1017/S0272263115000236

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2001). Learning Vocabulary in Another Language. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

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

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. (2022). About Quizlet | Quizlet. Retrieved September 25, 2021, from https://quizlet.com/mission

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology: Applied*, *24*(1), 57–71. https://doi.org/10.1037/xap0000146

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*, *43*(5), 1138–1156. https://doi.org/10.1017/S0272263120000236

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. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, I. Walker (Eds.), *CLaSIC* (pp. 721–737). Singapore: National University of Singapore, Centre for Language Studies.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*(2), 419–440. https://doi.org/10.1006/JMLA.2001.2813

Schuetze, U. (2015). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research*, *19*(1), 28–42. https://doi.org/10.1177/1362168814541726

Schuetze, U., & Weimer-Stuckmann, G. (2011). Retention in SLA lexical processing. *CALICO Journal*, *28*(2), 460–472. https://doi.org/10.11139/CJ.28.2.460-472

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

Serfaty, J., & Serrano, R. (2022). Lag effects in grammar learning: A desirable difficulties perspective. *Applied Psycholinguistics*, *43*(3), 513–550. https://doi.org/10.1017/S0142716421000631

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

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

Smith, C. D., & Scarf, D. (2017). Spacing repetitions over long timescales: A review and a reconsolidation explanation. *Frontiers in Psychology*, *8*(962). https://doi.org/10.3389/FPSYG.2017.00962/BIBTEX

Stroud, R. (2014). Student engagement in learning vocabulary with CALL. Paper presented at the CALL Design: Principles and Practice; Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands. https://doi.org/10.14705/rpnet.2014.000242

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

Suzuki, Y., Nakata, T., & DeKeyser, R. (2019). The Desirable Difficulty Framework as a Theoretical Foundation for Optimizing and Researching Second Language Practice. *The Modern Language Journal*, *103*(3), 713–720. https://doi.org/10.1111/modl.12585

Tschichold, C. (2012). French vocabulary in Encore Tricolore : Do pupils have a chance? *The Language Learning Journal*, *40*(1), 7–19. https://doi.org/10.1080/09571736.2012.658219

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

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/APPLIN/AML048

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*(3), 360–376. https://doi.org/10.1177/0033688209343854

Zalbidea, J. (2021). On the scope of output in SLA: Task modality, salience, L2 grammar noticing, and development. *Studies in Second Language Acquisition*, *43*(1), 50–82. https://doi.org/10.1017/S0272263120000261

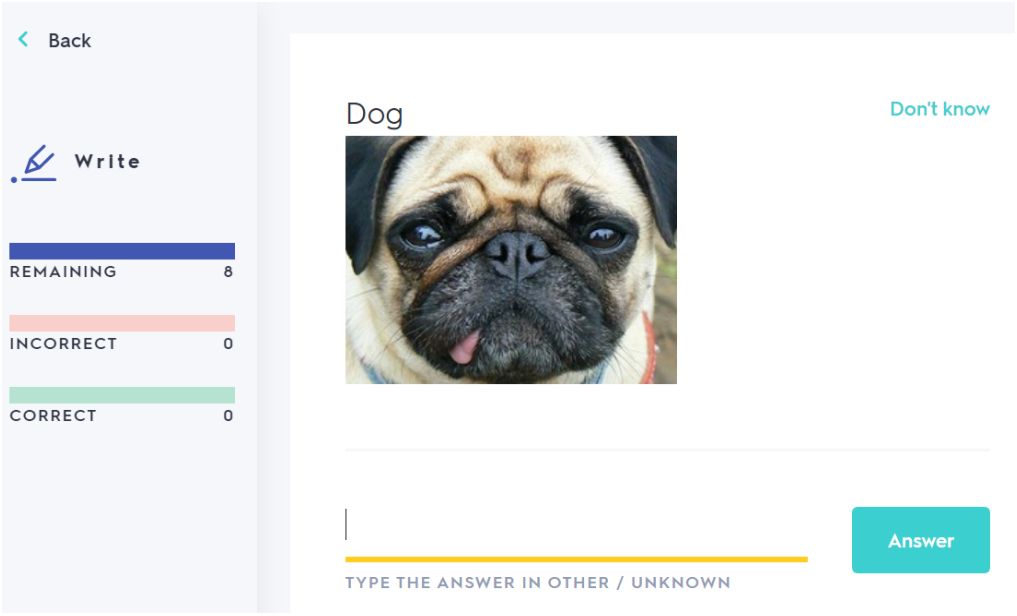
Zung, I., Imundo, M. N., & Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? *Memory*, *30*(8), 1–19. https://doi.org/10.1080/09658211.2022.2058553

Appendix A – Target Items

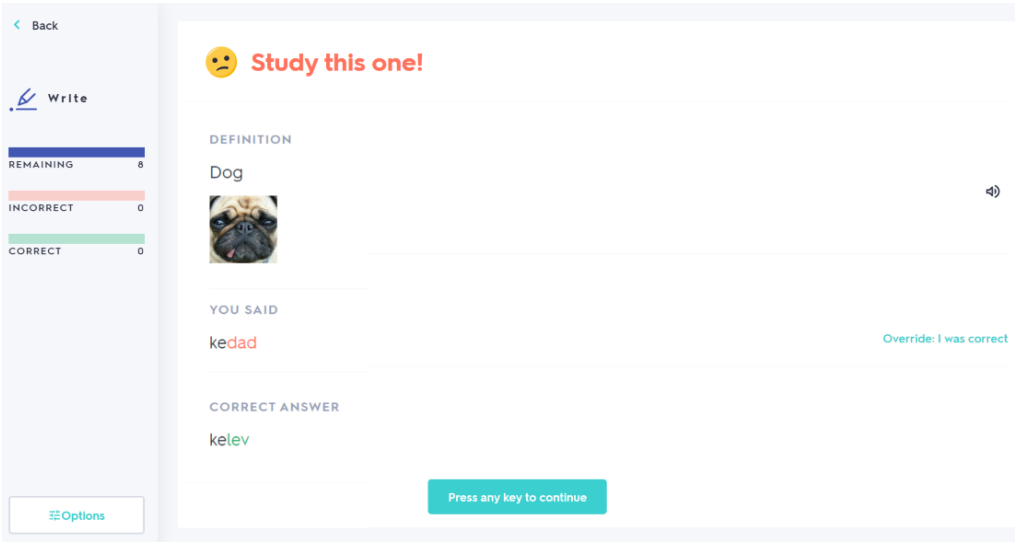
Target	English	Image	Target	English	Image
Kelev	Dog		Lehem	Bread	
Hatul	Cat		Halav	Milk	
Namer	Tiger		Mayim	Water	
Keves	Sheep		Tapuz	Orange	
Arnav	Rabbit		Gezer	Carrot	
Tanin	Crocodile		Marak	Soup	
Nesher	Eagle		Basar	Meat	
Karish	Shark		Glida	Ice Cream	

Appendix B – Quizlet Screenshots

Cue:



Feedback after an incorrect response:



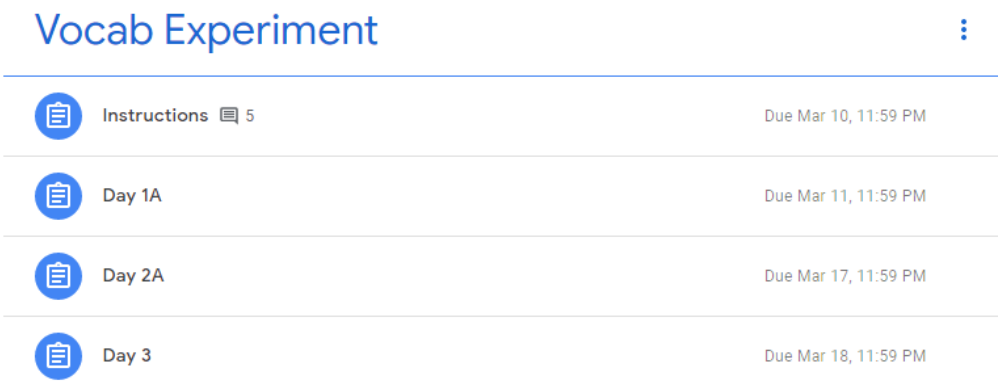
Appendix C – Mean age and training time of experimental groups

Posttest RI	Animals first	Food first	Total
RI-7	***n* = 24** Age: *M* = 13.46 (2.13) Time: *M* = 25.95 (15.67)	***n* = 24** Age: *M* = 13.17 (1.86) Time: *M* = 21.23 (7.98)	48
RI-28	***n* = 26** Age: *M* = 13.23 (1.58) Time: *M* = 22.75 (11.63)	***n* = 22** Age: *M* = 13.95 (1.96) Time: *M* = 25.05 (10.25)	48
Total	50	46	97

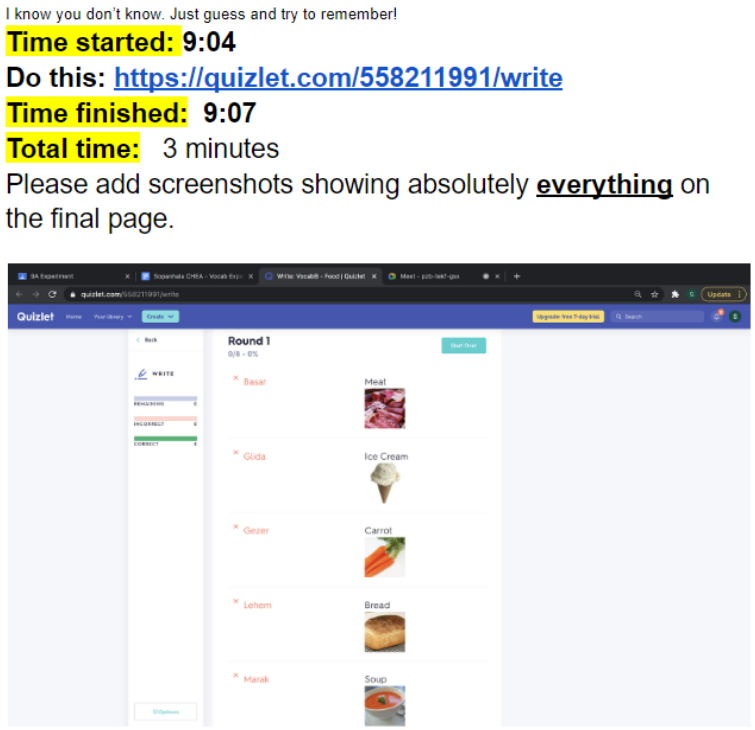
Comparison between RI-7 & RI-28	t	sig
Age	0.650	.518
Time on Task	0.337	.737

Appendix D – Google Classroom Screenshots

Assignments appeared in the participant's classroom at the specified time. Each assignment contained a Google Doc with the Quizlet link matching that participant's experimental condition.



The Google Doc had space to record their times and provided the link to the Quizlet set. They were required to add screenshots of the final page, which showed their progress in each round. Google Classroom also tracks the time that the doc was opened and submitted.



Chapter 5:

Practice makes perfect, but how much is necessary? The role of relearning in L2 grammar acquisition.

Resubmitted as:

Serfaty, J., & Serrano, R. (resubmitted). Practice makes perfect, but how much is necessary? The role of relearning in L2 grammar acquisition. *Language Learning.*

Reviewer 1: Major Revisions

Reviewer 2: Accept

Reviewer 3: Minor Revisions

Practice makes perfect, but how much is necessary?

The role of relearning in L2 grammar acquisition.

Abstract

This paper investigated how much practice is necessary to attain robust L2 grammar knowledge. Using digital flashcards, 119 participants learned an artificial language and practiced translating sentences from English with feedback. Participants performed one, two, three, or four relearning sessions on consecutive days. The number of trials needed to complete each session was recorded. At a 14-day posttest, groups with three or four relearning sessions achieved similarly high scores on productive tests, with significantly lower scores for the other groups. Receptive scores were high for all groups. Accuracy tended to peak on the third day. An analysis by individual training performance revealed that durable knowledge was attained after completing two sessions with minimum trials, regardless of how many sessions were performed. The findings provide a timeframe for processes described in Skill Acquisition Theory (DeKeyser, 2020) and demonstrate the amount of repeated output practice needed to proceduralize L2 grammar knowledge.

Keywords

flashcards; grammar; practice; relearning; skill acquisition theory

Introduction

With the recent global surge in online learning, self-paced second language (L2) practice is more prevalent than ever. For vocabulary, digital flashcard

apps like Quizlet are widely used, but less is known about using such apps for practicing the formulation of grammatical sentences. Grammatical accuracy often lags behind the development of vocabulary and fluency, and form-focused activities help to bridge this gap (Lyster, 2004; DeKeyser, 2010; Swain, 1988). Serfaty and Serrano (2020) demonstrated that flashcard apps could also be used for grammar practice, reporting gains of 82% that were largely retained after 18 weeks. However, a follow-up study using a similar methodology (Serfaty & Serrano, 2022) reported scores averaging 50% at the 7-day posttest and 25% after 28 days.

While these studies differed in several ways, one salient difference was the quantity of practice. Note that flashcards involve criterion learning, meaning that a session ends only after every item has been answered correctly. A repetition of this session on a different day can be classed as a relearning session (Bahrick, 1979; Rawson & Dunlosky, 2011), since the material has already been learned once. The high scores from Serfaty and Serrano (2020) came after three relearning sessions per structure, whereas the low scores from Serfaty and Serrano (2022) came from only one relearning session.

Cognitive psychologists have compared the retention of vocabulary pairs (Bahrick, 1979; Pyc & Rawson, 2007, 2009, 2011; Rawson et al., 2018; Vaughn et al., 2016; Vaughn et al., 2013) and key-term definitions (Rawson & Dunlosky, 2011, 2012, 2013; Rawson et al., 2013) after different amounts of retrieval practice to criterion. These studies found that for each additional successful retrieval in the learning phase, retention increased with diminishing effects. Importantly, the effects of relearning on different days

far exceeded the effects of overlearning within one session (Rawson et al., 2018). These studies concerned only the learning of idiosyncratic chunks of information, known as declarative knowledge. L2 grammar is a skill that involves combining linguistic parts according to rules. These rules might first be acquired as declarative knowledge, but according to Skill Acquisition Theory (SAT) the repeated application of these rules leads to the development of procedural knowledge, which is "knowledge that can only be performed" (DeKeyser, 2017, p. 17). Declarative and procedural L2 knowledge have been shown to have different properties in terms of learning rate and retention (Li & DeKeyser, 2019), which prevents conclusions from the extant relearning literature from applying directly to L2 grammar practice.

Studies of L2 grammar practice (e.g., DeKeyser, 1997; Ferman et al., 2009; Suzuki, 2017) have investigated the rate of improvement during the training phase and documented a subadditive pattern following the power law of practice found in other domains of skill acquisition (Newell & Rosenbloom, 1981). Accordingly, performance in accuracy and speed improves steeply in the initial stages and then gradually curves off until no further improvement is observable. It has been theorized that procedural knowledge is durable when these performance metrics level off (Kim et al., 2013). Consequently, for L2 grammar practice, there should be a point at which long-term knowledge is attained, and this should be predictable from training measures. However, to the best of the authors' knowledge, no study has yet investigated how much practice is necessary to achieve this or whether training performance indeed predicts later knowledge. This therefore

constitutes an important gap in the literature which needs to be filled not only for theoretical but also for pedagogical purposes. Determining how much L2 grammar practice is enough would provide a very useful guide for teachers and learners in efficiently allocating study time.

In the present paper, we compared groups that performed one, two, three, or four relearning sessions after an initial training session. First, we examined the groups' knowledge after 14 days without exposure to the target language. This timeframe will be deemed "long-term" in this paper, considering that it is a significant period of time to remember the rules and words of an artificial language, especially considering that participants were not told that the final session would test their knowledge. Second, we analyzed training performance to find out how many sessions are needed before accuracy no longer improves. Finally, we explored whether accuracy during training could predict accuracy in posttests, which would be especially useful for learners who wish to know when they have completed enough practice to attain long-lasting knowledge. In investigating these three questions, we aim to deepen our understanding of the learning processes for L2 grammar within the framework of SAT, as well as to provide pedagogically relevant recommendations regarding the optimal allocation of time for L2 grammar practice.

Literature Review

Digital Flashcards for Grammar

Digital flashcards are applications that prompt users to retrieve target information (e.g., L2 vocabulary item) from a cue (e.g., first language [L1]

translation). Non-target responses are met with feedback, showing the actual target, and are recycled until all items have received a correct response. Numerous studies have found flashcards to be a popular and effective method of learning, especially for L2 vocabulary (see Nakata, 2020 for a review).

Flashcards have also been used for L2 grammar learning (Serfaty & Serrano, 2020, 2022), meaning the ability to produce a full sentence in the L2 accurately without any parts of the sentence provided. Learners are presented with a meaning cue (e.g., a sentence in the L1) and type a full sentence in the L2. Each flashcard exemplifies a grammatical pattern. As a simple example, by producing *I eat rice, You eat rice, He eats rice,* and *She eats rice*, it can be induced that the *-s* is added for *He* and *She*, but not for *I* and *You*. A more complex example could be the formulation of different sentences using the third conditional structure (e.g., *If I had seen him, I would have told him*) to learn the pattern "If + subj + had + PP + obj, subj + would + have + PP + obj". This approach to grammar learning differs from more form-focused activities, such as conjugating a given verb or correcting an error, in that a full L2 sentence must be generated from meaning without support. Thus, although flashcards are designed with a target in mind, all elements of the sentence must be learned and produced perfectly during training.

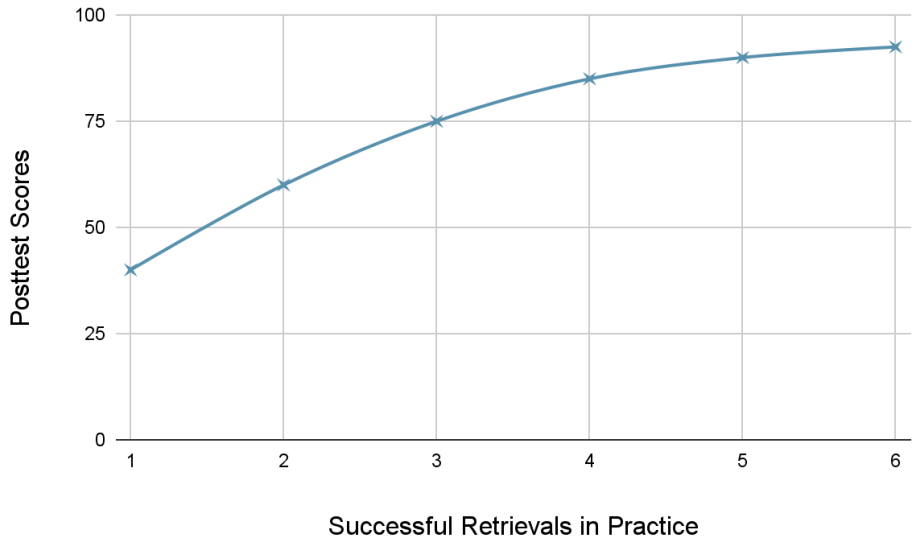
Flashcards facilitate highly controlled output practice, theorized to serve several functions (Swain, 1995). The cues prompt learners to notice any gaps in their knowledge and engage in hypothesis testing by typing their attempted L2 sentence. This is then met with explicitly corrective feedback that allows a comparison between interlanguage forms and target forms

(Zalbidea, 2021). Repeatedly practicing the same target structure allows for the proceduralization of accurate forms (Lyster & Sato, 2013).

Quantity of Practice from Paired-Associate Research

Cognitive psychologists have compared different quantities of learning for translations and definitions of target words. One focus of paired-associate research has been on overlearning, defined as the immediate continuation of learning, within the same session, of already-learned items (e.g., Rohrer et al., 2005). When additional correct retrievals are required before an item drops from a session, a subadditive effect is found (e.g., Pyc & Rawson, 2009; Vaughn et al., 2013). That is, posttest scores are higher, with progressively diminishing gains towards asymptote (Fig. 1).

Figure 1
Illustration of subadditive effects from overlearning



Another focus has been on relearning, with more relearning sessions facilitating higher gains (Bahrick, 1979; Bahrick et al., 1993). In studies using incremental increases in the number of sessions (Rawson & Dunlosky,

2011, 2012; Vaughn et al., 2016), the same subadditive effect from overlearning is found for relearning. However, relearning seems to be far more effective than overlearning (Rawson et al., 2018). Vaughn et al. (2016) compared retained knowledge from conditions with varying levels of overlearning in the initial session (1-7 correct retrievals) and varying amounts of relearning sessions (1-4). The mean score for items learned four times within a single session was 28%, but when the same number of retrievals were distributed over four weeks, the mean score was 74%.

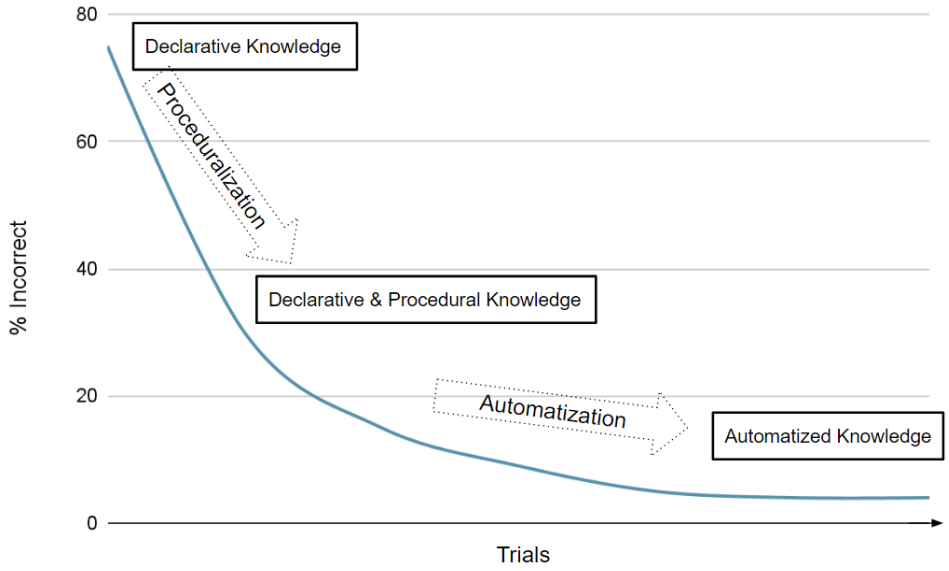
Skill Acquisition Theory

The studies mentioned in the previous section only examined declarative knowledge. Procedural knowledge, on the other hand, is the ability to perform a sequence, such as applying grammar rules in a sentence, and is developed through repeated practice. Declarative knowledge is acquired quickly but prone to decay, whereas procedural knowledge develops more slowly and is much more durable (Kim et al., 2013; Li & DeKeyser, 2017; Ullman, 2020).

SAT, adapted to instructed L2 learning by DeKeyser (2017, 2020), assumes that learners begin with declarative knowledge, which is proceduralized through practice. This knowledge then undergoes automatization, characterized by the gradual reduction in errors and response times (RTs) in retrieving knowledge or performing a skill. Kim et al. (2013)'s Skill Retention Theory (SRT) describes how the trajectory of progress in accuracy and RTs for performing a skill may correspond to three stages of learning from SAT, which they referred to as (1) declarative, (2) declarative +

procedural, and (3) procedural, shown in Figure 2. We have chosen to label the third stage as *automatized knowledge* because although declarative knowledge is no longer needed, it is also not necessarily lost.

Figure 2
Theoretical improvement in accuracy and type of knowledge attained based on Skill Retention Theory



At Stage 1, knowledge is declarative, analogous to being able to explain a grammar rule. Through application, early procedural knowledge is developed relatively quickly, moving into Stage 2. This transition, known as proceduralization, is reflected in a steep reduction in errors or RTs. During Stage 2, the learner still relies on the initial declarative knowledge. Training performance improves more gradually as procedural knowledge is slowly automatized through practice. As procedural knowledge becomes more dominant, declarative knowledge is relied upon to a lesser extent. Finally, the learner reaches Stage 3, evidenced by a leveling off in training performance. Here, declarative knowledge is no longer needed to produce a grammatical sentence. Even if rules are forgotten, learners are able to use grammar

structures appropriately. Kim et al. (2013) suggested that relearning is beneficial during Stages 1 and 2, but not after Stage 3. Applying this idea to L2 grammar learning, DeKeyser (2017) claimed that "intensive practice of known structures is only useful if it takes learners from the proceduralization stage (where declarative and procedural knowledge are used) to the automatization stage (where knowledge is completely procedural already)." (p. 96). However, to the best of our knowledge, no previous study has examined how different degrees of relearning, or intensive practice, affect the acquisition of L2 grammar knowledge in relation to the different stages reached during the training phase.

The Present Study

The present study compared groups practicing an artificial language with flashcards on two, three, four, or five consecutive days. The language included three fixed vocabulary items as well as subject and object pronouns to be constructed according to rules. All items were answered correctly during every session. If the language was still known after 14 days of disuse, we considered this to constitute robust knowledge. Our first question pertained to the number of relearning sessions necessary to attain this level of knowledge. SRT posits that knowledge is only durable after reaching Stage 3 and that no further learning past this stage is necessary. Based on previous studies using grammar flashcards (Serfaty & Serrano, 2020, 2022), it seems likely that the number of relearning sessions required to achieve robust knowledge is somewhere between one and four. We therefore hypothesized that a "threshold" for durable knowledge would be passed within one to four

relearning sessions, with no further improvement from relearning past this point.

Another aim of the present study was to establish how many relearning sessions are needed before the learner no longer sees improvement in accuracy during training, which SRT predicts to be indicative of robust knowledge. Previous studies (DeKeyser, 1997; Ferman et al., 2009; Pili-Moss et al., 2020; Suzuki, 2017) have shown a steep drop in errors after the first session and a flattening of the curve at around the fourth session, which might be an indication of what to expect from flashcard training.

Next, we examined individuals' scores as a function of their final attainment during training, rather than the number of relearning sessions they performed, in order to explore whether it was possible to predict, during training, when a learner has attained long-lasting knowledge. Groups based on the number of relearning sessions provide a general idea as to the effects of practice, but some structures and some individuals may require different amounts of practice (Ferman et al., 2009). We hypothesized that participants would achieve the highest posttest scores if their training performance plateaued, which Kim et al. (2013) suggest is a sign of durable knowledge. Our research questions (RQs) are as follows:

(RQ1) How many relearning sessions are needed to achieve durable L2 grammar knowledge, as shown by a 14-day posttest?

(RQ2) After how many relearning sessions does accuracy no longer improve during training?

(RQ3) Can an individual's accuracy during training predict when they have acquired robust L2 grammar knowledge?

Methods

Pilot

A pilot study was conducted to ensure that the artificial language was not too easy or too difficult. Using Quizlet, 30 volunteers from 8 countries aged 19-63, recruited through social media, studied the artificial language through inductive learning. They saw a sentence in English, attempted to type the translation in the artificial language, and studied the feedback in order to learn the rules. Participants practiced the set on either two, three, or four days. They each then performed cued-recall tests one, five, and 28 days after training. Results revealed highly variable abilities to learn the rules of the language, with times on task during the first session ranging from five minutes to over an hour, corresponding somewhat to age. Subsequent posttest scores, ranging from 0% to 100%, were heavily dependent on first-session performance. As a result, it was decided to include a guided rule-learning phase before the main task and to recruit only participants aged 18-30. Further, scores of 100% at the 1-day posttest consistently predicted identical scores at the 5-day and 30-day posttests, so it was decided to use a single posttest with a long enough interval for forgetting to occur. Finally, since Quizlet could not track trials and times, Gorilla (www.gorilla.sc) was chosen instead.

Participants

Participants were recruited via Prolific (www.prolific.co) and paid £10/hour. A filter was set to recruit only participants with fluency in English

(because instructions and cues were in English), aged 18-30, with a high school diploma or higher. The recruitment description heavily emphasized that this study would be difficult, involve multiple days, and that missing any day would result in no compensation. The intent was to attract only committed participants. Prolific offers a multitude of simple and well-compensated surveys that would be more suitable for any participant with time constraints or who was unmotivated to engage in language learning.

Participants came from 32 different countries with different L1s. Three participants were later excluded based on their performance (see Data Preparation). The final number of participants was 119, with 29 in the group with one relearning session and 30 in the remaining groups. Participant attributes, including age, gender, L1, and language learning background, had no discernible impact on results (see Appendix S1 in the online supporting materials for details).

Target Language

The target language was NamiChip, a miniature language developed for this paper by the first author. An artificial language was chosen in order to control for previous knowledge or practice outside of the treatment and was designed to be easily learnable. The target rules related to the marking of pronouns and nouns, with distinctions for case (subject/object), person (1st, 2nd, 3rd), and number (singular/plural) in an SVO structure. Since all participants were proficient in English, which also codes pronouns by these distinctions, this was expected to be conceptually simple. Vocabulary

included personal pronouns, one noun (*dog*), two verbs (*have* and *like*), and one number (*two*), as seen in Table 1. Pronouns were formed by adding the correct suffix to the letter "K", all subjects started with a small "s" (though capitalization was not tested), and plurals were marked by doubling the final letter. Words were combined in 24 different sentences, 12 for training and 12 for posttests (see Appendix S2), all designed to practice the same targets. Novel sentences were used for testing to ensure that participants had learned the rules governing the sentences, rather than memorizing whole chunks.

Table 1
All words in NamiChip

Subjects		Objects	
sKI	I	KI	Me
sKU	You	KU	You
sKII	We	KII	Us
sKEE	They	KEE	Them
sCHIP/sCHIPP	Dog/Dogs	CHIP/CHIPP	Dog/Dogs
Additional words: NAMI = like ; TEN = have ; BI = two			

Training Procedure

The study was developed and conducted through the online platform Gorilla. Participants first agreed to the use of their data, to not take any notes, to finish quickly while still checking for accuracy, and to set a reminder for each subsequent session. Although participants could not be monitored, it was hoped that these agreements would encourage them to follow the

experiment faithfully. While this is no guarantee of what participants did, any non-adherence to the experiment would be equally likely in all conditions and in any case, a supervised classroom setting would present the same issue. In the guided-learning phase, participants first learned the words and rules of the language by answering questions. For example, one screen showed that "CHIP" means *dog* and "CHIPP" means *dogs* and asked participants to select the rule for plurals from the options: (1) Add -s (2) Double the last letter (3) Write the whole word twice (4) This language has no plurals. Another section involved typing individual words from the language, similar to using vocabulary flashcards. Incorrect answers prompted feedback and additional chances until all the vocabulary and rules had been learned. Appendix S3 displays all screens from this phase.

In the second phase of the training session, participants saw an English sentence and were asked to provide the translation in NamiChip (Fig. 3). If incorrect, they saw the target sentence alongside their own response (Fig. 4), and the flashcard was sent to the back of the cycle. If correct, they saw positive feedback (Fig. 5) and the flashcard was removed. This phase simulated commercial flashcard software, such as Quizlet, within Gorilla. Subsequent relearning sessions simply repeated this second phase using the same 12 sentences each time.

Figure 3
Typing the translation of a cue

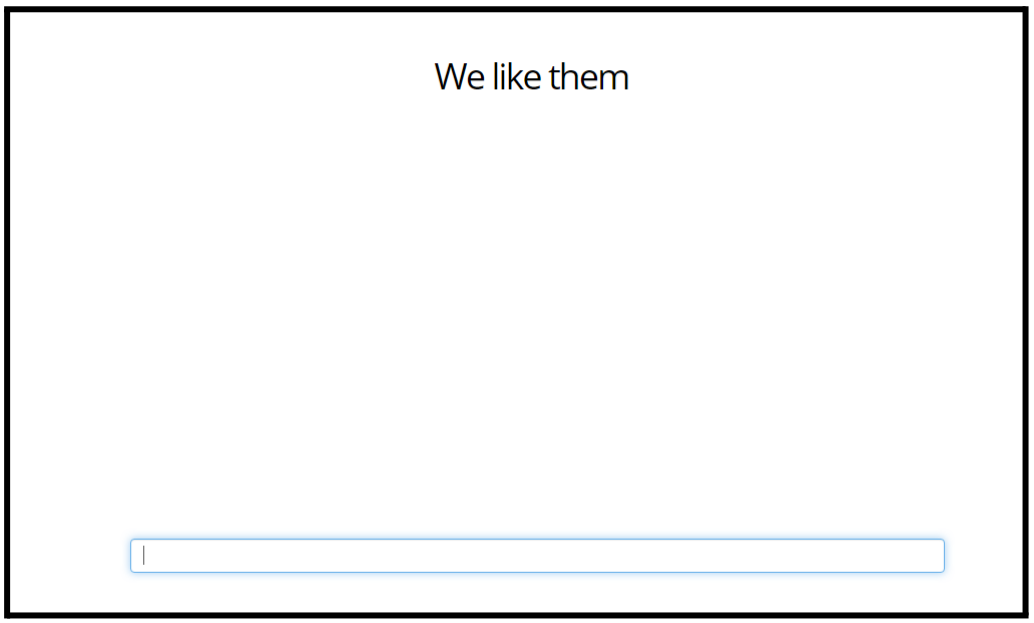


Figure 4
Negative feedback

Figure 5
Positive Feedback

Training Performance

The main goal of the training was for participants to acquire and retain accurate production of the target language. In training, accuracy was operationalized as the number of trials per session. Since all sessions had a minimum of 12 trials, corresponding to 12 correctly typed items, any trials above this number indicated incorrect responses. A reduction in trials is therefore an improvement in accuracy.

As well as the number of trials per session, previous grammar studies have recorded RTs (DeKeyser, 1997; Ferman et al., 2009; Pili-Moss et al., 2020; Sato & McDonough, 2019; Suzuki, 2017; Suzuki & DeKeyser, 2017). The present study did not analyze this metric because some confounding variables were identified. For typed responses, especially of multi-word items, speed would be influenced by a participant's typing skills, their

keyboard, and their connection speed, rather than being a pure measure of internal processing speed. Any improvement in RTs could not be separated from participants' improved ability to type new words quickly. These factors are likely to overshadow changes in processing speed, which is measured on the scale of milliseconds. Additionally, no time pressure was imposed in the training. Participants were encouraged to finish quickly but to check responses before submitting, resulting in more time for more diligent participants. Moreover, if an item was answered correctly on its second attempt, the RT might be lower than if it had been answered correctly on the first attempt because it had been recently practiced. This presents a dilemma as to whether to analyze RTs for only correct responses or for all responses. Due to these issues, and because the focus of the present study was accuracy development, the number of trials per session was the only measure of performance analyzed.

Instruments

Participants knew that there would be a final session, but they were not informed that this would be a test. The first posttest took the same format as the training, but without feedback. Participants saw English cues and were asked to provide a translation in the target language. In the second test, the same target sentences became cues and participants had to translate them into English. Although both tests involved actively producing language (i.e., they tested participants' recall and not just recognition of the target forms), the first test required productive knowledge of the target language (translating from the L1 to the L2) whereas the second test only required receptive

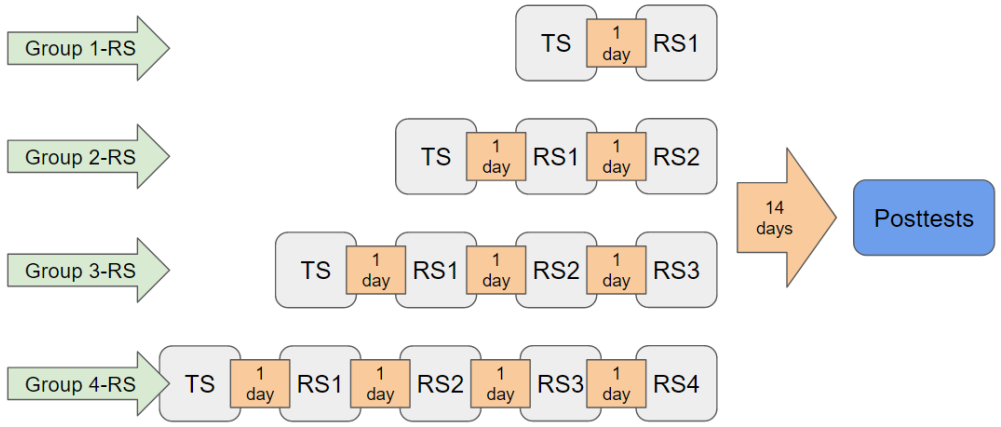
knowledge of the target language (translating from the L2 to the L1). We will therefore refer to these tests as productive and receptive tests respectively. Vocabulary research has consistently found receptive recall to be higher than productive recall (Laufer & Goldstein, 2004; Nakata, 2016; Nakata & Webb, 2016; Webb, 2009), and so while productive scores are the primary interest of the present paper, the subsequent receptive test was expected to capture weaker knowledge traces which might not be observable from the productive test. To the best of the authors' knowledge, no previous study has compared productive and receptive grammar knowledge, operationalized as L1 to L2 versus L2 to L1 translation, after different amounts of practice. It was therefore included as an initial exploration of this issue, despite not being a main focus of the paper.

Upon completing both tests, participants entered a debriefing phase where they were asked to rate the training for enjoyment, ease, and perceived effectiveness on a 5-point Likert scale, followed by a space for them to write about their feelings. This allowed us to gain some insight into the participants' experiences. We also included a space for them to specify any rules of the language that they remembered. It was rationalized that even if knowledge was proceduralised during training, posttest responses after a 14-day gap may involve some retrieval of declarative knowledge, or that declarative knowledge could be reverse-engineered upon reflection. This question was included in order to facilitate the interpretation of our data.

Experimental Design

The training session (TS) and relearning sessions (RSs) took place on consecutive days. A short interval of one day was chosen in order to minimize participant attrition while still allowing enough time for the consolidation of knowledge between sessions (Ferman et al., 2009). The retention interval between training and testing was 14 days, the longest possible interval before Prolific would automatically pay participants, who might not then complete the posttest. The four conditions differed only in the number of RSs performed (1, 2, 3, or 4). Figure 6 shows the experimental design.

Figure 6:
Experimental design.



Note: TS = Training Session. RS = Relearning Session.

Data Preparation

Posttests were scored automatically in Gorilla. Incorrect responses were manually checked for answers that were correct despite not exactly matching the target response. For example, if there was an error in the

number of spaces or, in one case, where the English response had the spelling error “doggs”.

Two participants were excluded for not having attempted to translate the cues. One additional participant was removed for having required considerably more trials in the first two sessions (TS = 72, RS1 = 74) than the participant with the next highest number of trials (TS = 53, RS1 = 36), with no correct answers in the posttest.

Finally, the comments from participants were checked. It was observed that the majority of participants were able to list the rules of the language extensively. These observations support the validity of the data because participants would not be reasonably expected to remember these rules without having participated faithfully in the experiment. Moreover, the comments about feelings revealed some emotional investment from both high and low scoring participants. These comments can be found online in Appendix S4. The 5-point Likert survey on perceptions was checked for any ostentatious differences between groups, which could have affected results. All groups rated enjoyment and effectiveness very positively at 4-5, while ease was rated slightly lower at 3-4 for all groups. Descriptive results of this survey are presented in Appendix A.

Analysis

All statistical analyses were performed using SPSS 27 (IBM, 2020). To analyze posttest scores, Generalized Linear Mixed Models (GLMMs) with binary outcomes were performed in order to locate statistical differences between groups. This type of model does not require any assumptions to be

met. Participant and Item were included as random effects, with Group and Test as the predictors. The effect size is the odds ratio (*OR*), which indicates the added likelihood of a correct score from the condition with the higher mean. For example, if *OR* = 2, a correct response is twice as likely.

The number of error trials was analyzed through a Growth Curve Model using a poisson distribution. The data met assumptions for normality of residuals, homogeneity of variances, and overdispersion. Random intercepts were used for participants and random slopes were used for progress over time. Session was set as a numerical predictor. Initially, Group was also included to check for unexpected differences between groups, but removed when no effect was found. The effect size is the incidence rate ratio (*IRR*), which is the ratio of change in the average number of trials between sessions. For example, if the *IRR* is 0.5, then the mean number of error trials halved in the later session.

All significance tests were two-tailed and a *p* value of .05 or lower will be interpreted as significant, with Bonferroni corrections within models. The datasets can be found online along with Appendix S5, which contains every model's syntax as well as all the means, effects and associated 95% confidence intervals.

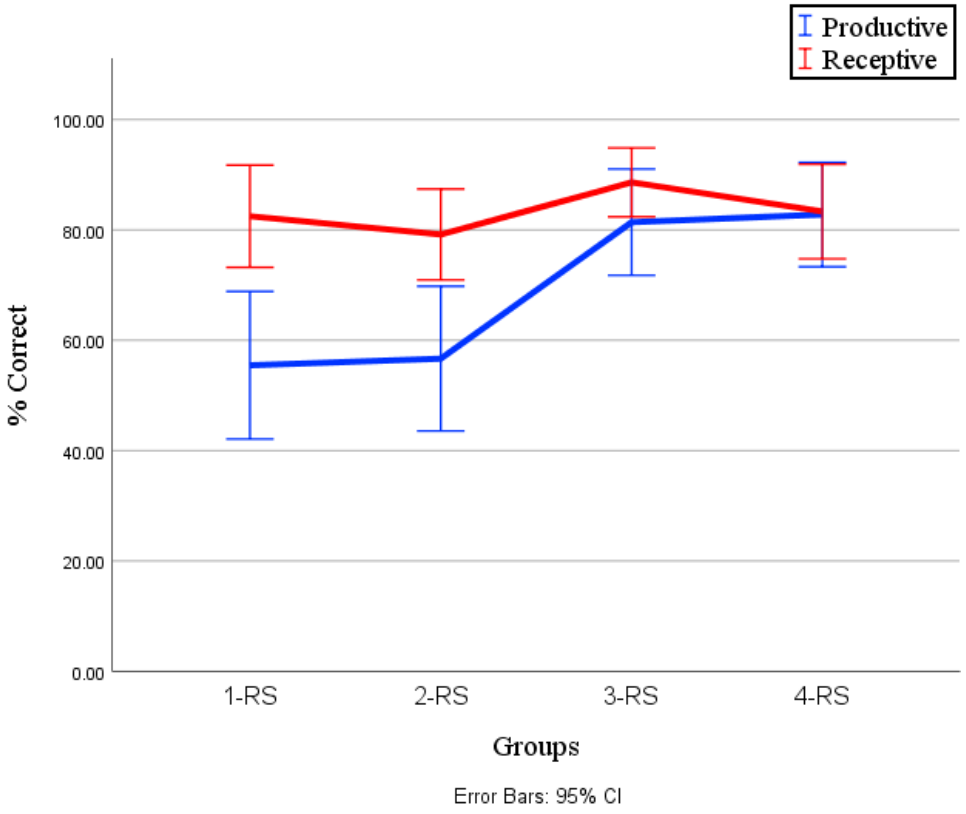
Results

RQ1: How many learning sessions are needed to achieve durable L2 grammar knowledge?

Figure 7 displays the mean scores by group for productive and receptive tests. The groups are labeled according to the number of RSs they

performed. For the productive test, Groups 1-RS and 2-RS scored 55.46% (35.16) and 57.67% (35.11), while Groups 3-RS and 4-RS obtained a much higher 81.39% (25.77) and 82.78% (25.33). Receptive scores did not show this variation between groups. From Group 1-RS to Group 4-RS, scores were 82.47% (24.33), 79.17% (22.08), 88.61% (16.74), and 83.33% (22.95) respectively.

Figure 7
Productive and receptive scores



The statistical model produced significant main effects for Group (F[3,2848] = 3.343, p = .018), Test (F[1,2848] = 80.409, p < .001) and their interaction (F[3,2848] = 3.343, p < .001). Table 2 shows the estimated marginal means (EMMs) and pairwise contrasts. For productive scores, there were no statistically significant differences within the first two groups (1-RS

vs 2-RS) or within the second two groups (3-RS vs 4-RS). It appears that RS2 did not improve the odds of a correct posttest response, but those odds were around 3.5 times higher after RS3. No further improvement in odds was obtained from RS4. For receptive scores, there were no statistically significant differences between groups, suggesting that RS1 was sufficient.

Table 2
Groups means and comparisons for productive and receptive scores

Productive Scores			Pairwise Contrasts		
Group	**EMM**	**SE**	**Group 2-RS**	**Group 3-RS**	**Group 4-RS**
Group 1-RS	.555	.063	*p* = 1.000 *OR* = 1.050	*p* = .002 *OR* = 3.512	*p* = .001 *OR* = 3.860
Group 2-RS	.567	.062		*p* = .001 *OR* = 3.344	*p* = .002 *OR* = 3.676
Group 3-RS	.814	.049			*p* = 1.000 *OR* = 1.099
Group 4-RS	.828	.047			
Receptive Scores			Pairwise Contrasts		
Group	**EMM**	**SE**	**Group 2-RS**	**Group 3-RS**	**Group 4-RS**
Group 1-RS	.825	.042	*p* = 1.000 *OR* = 1.238	*p* = 1.000 *OR* = 1.654	*p* = 1.000 *OR* = 1.063
Group 2-RS	.792	.044		*p* = .603 *OR* = 2.042	*p* = 1.000 *OR* = 1.316
Group 3-RS	.886	.035			*p* = 1.000 *OR* = 1.556
Group 4-RS	.833	.041			

RQ2: After how many relearning sessions does accuracy no longer improve during training?

RQ2 asked how many RSs are needed before accuracy reaches asymptotic performance. Note that only the TS and RS1 included all participants. RS2 included Groups 2-RS, 3-RS, and 4-RS, RS3 included Groups 3-RS and 4-RS, and RS4 included only Group 4-RS.

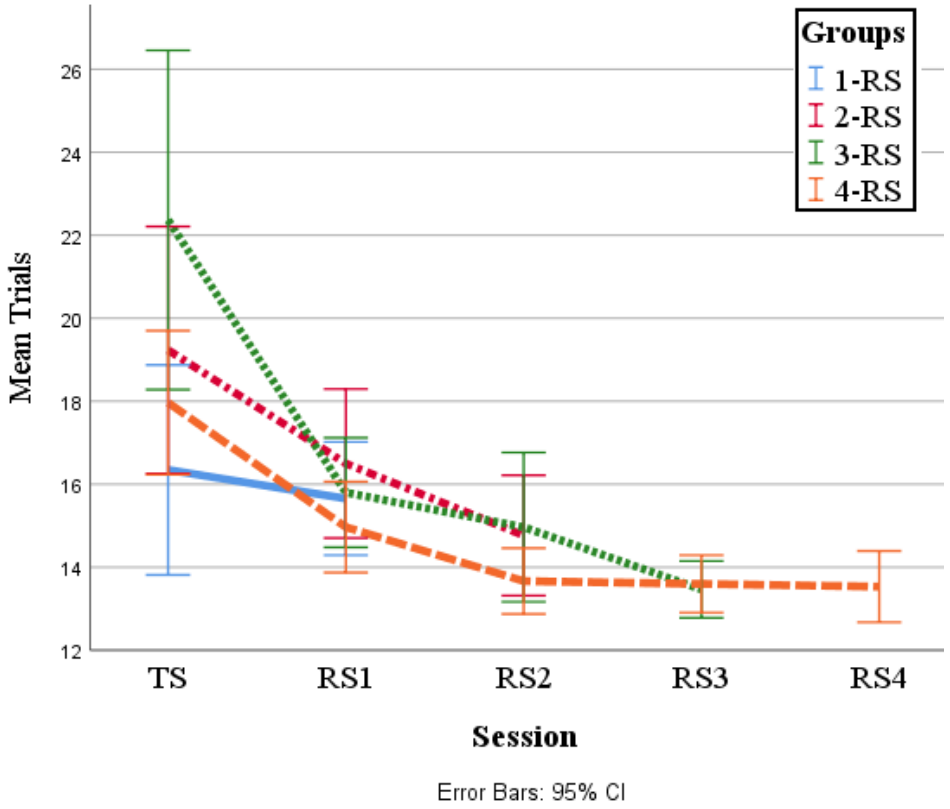
Table 3 shows the number of trials required to complete each session. Figure 8 shows the learning curves of each group. Group 3-RS showed more variance than other groups during the TS, but by RS1 all groups exhibit similar performance. The learning curve appears steeper from the TS to RS1 and then gradually flattens, in line with previous research.

Table 3
Trials by Session (minimum 12)

Session	N	Mean Trials	Std. Deviation	Minimum	Maximum
TS	119	19.00	8.11	12	53
RS1	119	15.73	3.77	12	36
RS2	90	14.47	3.78	12	37
RS3	60	13.53	1.83	12	20
RS4	30	13.53	2.30	12	22

Figure 8

Trials by Session separated by group



The Growth Curve Model (Table 4) analyzed the reduction in error trials across sessions. The main effect of Session was significant ($F[4,413] = 32.268$, $p < .001$), showing a statistically significant decrease in trials from TS to RS1, from RS1 to RS2, and from RS2 to RS3. The extent of this decrease diminished as sessions progressed, from approximately -2, to -1, to -0.5 trials. There was no significant difference in trials from RS3 to RS4. In sum, the extent of improvement roughly halved with each new session, with no further improvement at the final session.

Table 4
Pairwise comparisons of error trials per session

Session	*EMM*	*SE*	Contrast	Contrast estimate	*IRR*	Sig
TS	5.100	0.444	RS1	-2.188	0.571	<.001
			RS2	-3.223	0.368	<.001
			RS3	-3.858	0.244	<.001
			RS4	-3.902	0.235	<.001
RS1	2.912	0.247	RS2	-1.034	0.645	<.001
			RS3	-1.670	0.427	<.001
			RS4	-1.714	0.411	<.001
RS2	1.877	0.195	RS3	-0.635	0.662	0.009
			RS4	-0.679	0.638	0.072
RS3	1.242	0.187	RS4	-0.044	0.965	0.864
RS4	1.198	0.268				

RQ3: Can an individual's accuracy during training predict when they have acquired robust L2 grammar knowledge?

RQ3 asked whether an individual's accuracy during training could predict the acquisition of durable knowledge. To answer this question, we examined the number of trials an individual needed to complete each session and compared this with their posttest results.

The mean number of trials reached 14 at RS2 and remained at 14 ($M >$ 13.5) for RS3 and RS4. Unexpectedly, this plateau was not at the true minimum of 12 trials. Upon inspecting the data, it was found that 15 of the 85 participants who had finished their final session with 13 or 14 trials had previously completed at least one session in 12 trials, meaning that even participants that had demonstrated perfect accuracy in a previous session were making one or two errors. We reasoned that typing mistakes could be responsible for this discrepancy. If participants had been using commercial software, they would have had the option to override these mistakes, but under experimental conditions this was not possible. We therefore defined the minimum number of trials for this experiment as 14 trials or fewer. For ease of exposition, any session completed within 14 trials will be referred to as a minimum-trials session (MTS).

The average number of trials by RS2 was already very close to 14 (M = 14.47), but it was only after RS3 that posttest scores were high. We therefore hypothesized that one MTS was not enough and that learners must perform at least two MTSs in order to acquire robust knowledge. In order to test this, we re-coded participants by the number of MTSs they achieved in succession. Some participants (n = 18) completed one MTS but then needed

more than 14 trials in a subsequent RS. This may have been due to guessing correctly by chance in an earlier session or a lack of attention in a later session. Since they could not be classified clearly, they were excluded from this analysis. The remaining participants (n = 101) were grouped as follows: MTS-0: n = 24; MTS-1: n = 22; MTS-2: n = 23; MTS-3: n = 14; MTS-4: n = 14; MTS-5: n = 4. The breakdown of MTS groups in terms of membership of the original RS groups can be seen in Table 5. Descriptive statistics of posttest scores for MTS groups are displayed in Table 6 and Figure 9.

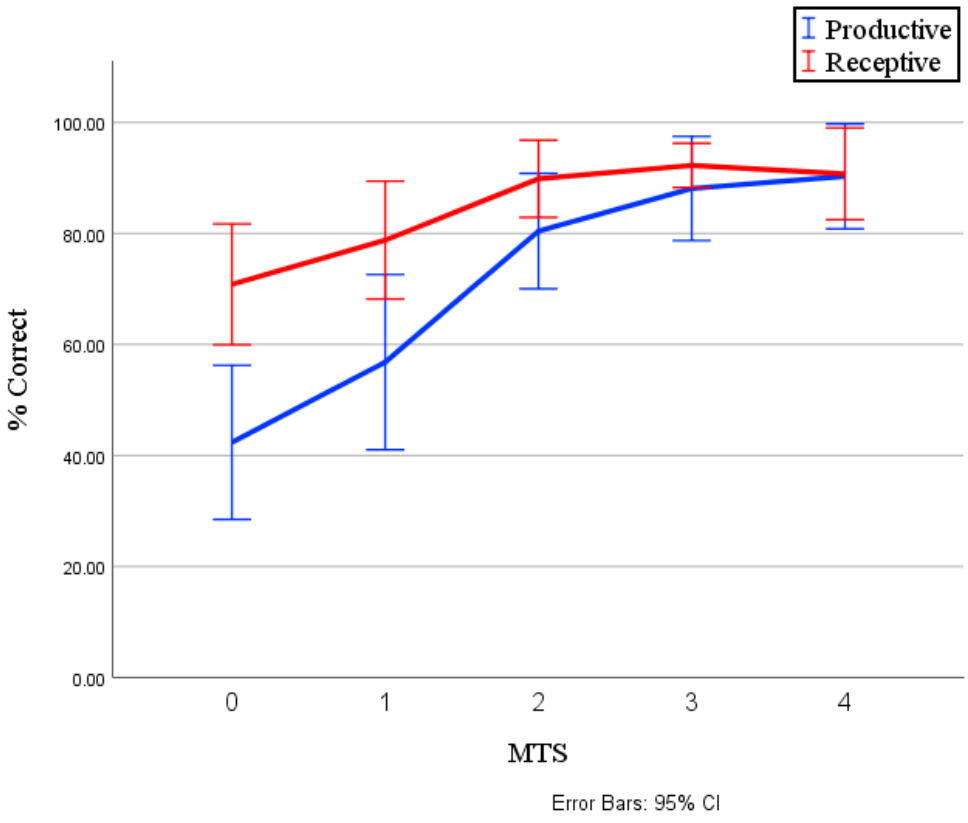
Table 5
MTS groups by original group membership

MTS	Group 1-RS	Group 2-RS	Group 3-RS	Group 4-RS	Total
0	14	6	2	2	24
1	6	10	6	0	22
2	9	5	8	1	23
3	0	3	6	5	14
4	0	0	5	9	14
5	0	0	0	4	4

Table 6
Productive and receptive scores (%) by MTS groups

MTS	Productive Scores	SD	Receptive Scores	SD
0	42.36	32.87	70.83	25.77
1	56.82	35.60	78.79	23.95
2	80.43	24.05	89.86	16.08
3	88.10	16.25	92.26	6.91
4	89.29	20.78	91.07	18.33
5	93.75	12.50	89.58	10.49

Figure 9
Productive and receptive scores by MTS groups



Since MTS-5 had only four members, it was collapsed into MTS-4 for the statistical analysis (n = 18; productive: M = 90.28, SD = 19.01; receptive: M = 90.74; SD = 16.64). There were significant main effects for MTS (F[1,2414] = 39.883, p < .001), Test (F[4,2414] = 8.325, p < .001) and their interaction (F[4,2414] = 3.621, p < .001). For productive scores, statistically significant differences were found between MTS-0 and all other groups, and between MTS-1 and all other groups. It should be noted that these effect sizes are substantially higher than for the analysis by number of RSs. For instance, the *OR* for the difference between MTS-0 and MTS-4 is 12.635, whereas the highest *OR* in the original analysis was 3.860. No other differences in productive scores were statistically significant. For receptive scores, differences between MTS-0 and other sessions approached significance. Table 7 displays all pairwise contrasts.

Discussion

The present paper investigated the effects of relearning L2 grammar through digital flashcards. Groups practiced formulating the same set of sentences in an artificial language with English cues on two, three, four, or five consecutive days, with all items recycled until answered correctly within each session. The first day was the training session (TS) and subsequent days were relearning sessions (RSs). Translation tests on novel sentences were performed 14 days after the treatment, firstly from the English cue to the target language (productive knowledge) and then from the target language to English (receptive knowledge). We will now summarize the findings from each RQ before discussing their implications.

Table 7
Pairwise comparisons for productive and receptive scores by MTS groups

Productive Scores			Pairwise Contrasts			
MTS	**EMM**	**SE**	**MTS-1**	**MTS-2**	**MTS-3**	**MTS-4**
0	.424	.055	$p = .296$ $OR = 0.559$	$p < .001$ $OR = 5.594$	$p < .001$ $OR = 10.069$	$p < .001$ $OR = 12.635$
1	.568	.058		$p = .011$ $OR = 3.124$	$p = .004$ $OR = 5.624$	$p = .001$ $OR = 7.057$
2	.804	.045			$p = .543$ $OR = 1.800$	$p = .354$ $OR = 2.259$
3	.881	.047				$p = .717$ $OR = 1.255$
4	.903	.038				
Receptive Scores			Pairwise Contrasts			
MTS	**EMM**	**SE**	**MTS-1**	**MTS-2**	**MTS-3**	**MTS-4**
0	.708	.051	$p = 1.000$ $OR = 1.529$	$p = .041$ $OR = 3.647$	$p = .064$ $OR = 4.910$	$p = .054$ $OR = 4.035$
1	.788	.048		$p = .410$ $OR = 2.385$	$p = .410$ $OR = 3.210$	$p = .410$ $OR = 2.638$
2	.899	.034			$p = 1.000$ $OR = 1.346$	$p = 1.000$ $OR = 1.106$
3	.923	.039				$p = 1.000$ $OR = 1.217$
4	.907	.037				

RQ1: How many learning sessions are needed to achieve durable L2 grammar knowledge?

Our first RQ asked how many RSs are needed to gain durable grammar knowledge. After RS1 and RS2, productive scores were comparable (55% and 57%), but after RS3 and RS4, they were considerably

higher (81% and 83%) with no significant differences within these pairs. It should be noted that these means represent a wide range of scores for the first two groups, whereas very few errors were made among the latter two groups. Therefore, RS3 substantially raised the odds of participants learning and retaining the target language, with no further improvement from RS4.

Different results were found for receptive scores, which were similarly high for all groups, meaning that receptive knowledge was durable after only one RS. This is similar to Pili-Moss et al. (2020), whose training data for participants learning an artificial language showed that receptive knowledge peaked after two sessions whereas productive knowledge improved until the final fourth session. This pattern is also mirrored in vocabulary studies that have tested both productive and receptive vocabulary knowledge (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; Nakata, 2016; Nakata & Webb, 2016; Webb, 2009), in that receptive knowledge tends to develop faster and is more durable (de Bot, 1996).

The findings support our speculation that the different results from previous grammar flashcard studies were related to the number of sessions. Participants in Serfaty and Serrano (2020) practiced target structures on three days, followed by a 1-day posttest with mean scores over 80% and a third of participants achieving 100%. The high accuracy in this posttest probably reinforced their knowledge (Karpicke & Roediger, 2008), making a total of three RSs. Tested again after 14 days, scores were still over 80%. Likewise, participants in the present experiment scored above 80% after three RSs at a 14-day posttest. In comparison, participants in Serfaty and Serrano (2022) practiced target structures with one RS, and posttest scores only reached 50%

at a 7-day posttest. Similarly, in the present experiment, those who studied with one RS scored around 50% at their posttest. Considering that these three experiments differed in many aspects, including the use of an artificial or natural language, the age and background of participants, experimental or classroom settings, and the complexity of target structures, these similarities in scores are salient.

RQ2: After how many relearning sessions does accuracy no longer improve during training?

RQ2 asked how many RSs would be needed for improvements in accuracy to plateau at the training stage. The number of trials dropped from the TS until RS2. From RS2 to RS3, the difference was statistically significant but less than a single trial. Therefore, on average, two RSs were needed to reach the minimum number of trials.

The overall pattern is similar to previous studies of artificial language learning that have reported a reduction in errors following a similar learning curve (DeKeyser, 1997; Ferman et al., 2009; Pili-Moss et al., 2020; Suzuki, 2017). These studies differed to each other and to the present study in terms of training, target material, testing and time on task, yet all of these studies have reported a plateau in error rates at around the third or fourth session. In contrast, participants from Sato and McDonough (2019) did not reach asymptotic accuracy after a fifth session of interactive task practice within an authentic L2 classroom. This could place doubt onto the applicability of artificial language studies to authentic L2 learning. However, in the latter study target forms were not prompted systematically and no feedback was

provided by the instructor during the tasks. Had these conditions been met, results may have been similar to the artificial language studies.

RQ3: Can an individual's accuracy during training predict when a learner has acquired robust L2 grammar knowledge?

RQ3 asked whether an individual's accuracy during training could be used to predict long-term L2 grammar knowledge. After observing that high productive posttest scores came after RS3, one session after trials reached their minimum, a new hypothesis was formed that learners must complete two minimum-trials sessions (MTS) in order to gain robust knowledge. A new analysis revealed that high scores (80% for productive and 89% for receptive) were achieved if a minimum of two MTSs were performed, regardless of how many sessions were performed overall, and with no further improvement from additional MTSs.

From these results, it is clear that individuals require different amounts of practice. An MTS indicates that participants are practicing known items and for many of our participants this occurred for the first time during RS2. However, some participants achieved an MTS in their first session while some still could not perform an MTS by their fifth session. Consequently, anyone attempting to gain robust knowledge would be advised to practice until they have achieved a certain number of MTSs, rather than for any predetermined number of sessions.

Theoretical Implications

The results obtained in this study can be interpreted through Skill Retention Theory (SRT; Kim et al., 2013). According to this framework, in Stage 1 learners only have declarative knowledge, in Stage 2 procedural knowledge is developed through practice, but still dependent on declarative knowledge, and in Stage 3 further practice results in procedural knowledge that can be used independently of declarative knowledge. Since declarative knowledge is prone to decay, only independent procedural knowledge can be durable. The results of the present study indicate that many participants reached Kim et al. (2013)'s Stage 3 after RS3, and that reaching this stage in training did predict later retention. Importantly, no further gains were evident after continuing to learn past this stage, confirming claims that learning past Stage 3 is not useful (DeKeyser, 2017; Kim et al., 2013).

However, comments from participants made after the posttest revealed considerable declarative knowledge, and so it is not certain whether procedural knowledge had become independent of declarative knowledge, as SRT would predict. Interestingly, many of the lower-scoring participants also exhibited a lot of declarative knowledge of rules after the posttests. One interpretation could be that declarative knowledge, however well-retained, is not sufficient for accurately producing the L2, even in a non-timed and non-communicative task. However, all participants scored highly in the receptive test, implying that declarative knowledge could be useful for comprehension. Another interpretation could be that declarative knowledge was not necessarily retrievable before the posttest, but upon seeing the L2 forms as cues in the receptive test, participants were reminded of previous

knowledge and were therefore able to translate the sentences and recount the rules in the debriefing phase. Three participants wrote comments to this effect, which can be seen in Appendix B.

In sum, the present study has confirmed that a learner's level of knowledge is predictable at the training stage and that a threshold of learning must be crossed to gain durable knowledge. Crucially, these results have shown that this threshold is reached after performing two practice sessions without errors.

Limitations and Future Directions

Future research could test our interpretation of results by replicating the experiment with a longer delay before testing. If participants have truly achieved durable knowledge after two MTSs, then this knowledge should still be accessible after a longer retention interval. Moreover, if learners failing to reach this threshold have less durable knowledge, scores after a longer delay would be expected to drop below 50%. Similarly, this design could be replicated with longer intersession intervals. If the declarative knowledge of rules is somewhat forgotten between sessions, more trials could be expected in the first RS and possibly later. This could affect the number of sessions required to reach Stage 3. On the other hand, longer intersession intervals may promote better retention of the declarative knowledge of rules (Bird, 2010), resulting in better posttest scores from fewer sessions.

Another variable to investigate could be the mode of output. As noted, our MTS was not a truly error-free session because some typing mistakes

caused interference. A study with oral output, rather than typing, would eliminate this methodological issue while also testing whether the results from this study generalize to oral output practice. This change would also enable an exploration of automatization by measuring reaction times and utterance speed.

Finally, the data imply that participants were adhering to instructions well, and their comments expressed both commitment and enjoyment of the process, but a replication of this experiment under supervised conditions would be desirable to confirm the internal validity of the findings. Still more interesting would be to conduct the experiment in a less-controlled authentic L2 classroom, with a wider range of abilities, motivations, and exposure to the target language, in order to investigate whether the findings are generalizable to authentic learning conditions.

Conclusions and Pedagogical Recommendations

The present paper investigated the effects of relearning on grammar flashcard training. Tested after 14 days, it was found that productive knowledge was much higher after three RSs (four sessions in total). Participants that reached the highest levels of productive knowledge were those who completed at least two sessions with minimum trials, and this turned out to be a much better predictor of individual success than the number of RSs. To our knowledge, this is the first study to compare retention of L2 grammar after different amounts of practice.

Though we cannot be certain about the type of knowledge gained from such practice, it is evident that digital flashcard practice leads to high

accuracy, especially for relatively simple and regular structures, which could be useful to prepare students for communicative language practice. In doing so, it is more likely that learners will automatize accurate target forms while focusing on communication goals. To this end, teachers could assign flashcard sets to students on different days and monitor their progress. When one student has completed a set twice without any errors, on different days, the teacher could assign them a new set without interrupting the training schedule of their classmates. These sets could be part of classroom learning or assigned as homework. Using insights from this study and future research, an eTutor could be designed to automatically assign or reassign activities to students based on their performance.

It remains to be seen whether the number of error-less sessions needed to acquire durable knowledge would be the same for different types of L2 structures, L2 vocabulary, or even for other subject domains. It is however clear that sessions dedicated to relearning should be factored into curricula. If knowledge and skills are only practiced once or twice, many students may go through their education achieving high scores in tests without remembering what they have learned. It may be advisable for teachers and curriculum writers to reduce the amount of content being taught in order to allow enough practice of the most fundamental content for students to attain useful, long-lasting knowledge.

References

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*(3), 296–308. https://doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*(5), 316–321. https://doi.org/10.1111/J.1467-9280.1993.TB00571.X

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

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

De Bot, K. (1996). The psycholinguistics of the Output Hypothesis. *Language Learning*, *46*(3), 529–555. https://doi.org/10.1111/J.1467-1770.1996.TB01246.X

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*(2), 195–221. https://doi.org/10.1017/S0272263197002040

Dekeyser, R.M. (2010). Practice for Second Language Learning: Don't Throw out the Baby with the Bathwater. *International Journal of English Studies*, *10*(1), 155–165. https://doi.org/10.6018/IJES/2010/1/114021

DeKeyser, R. M. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge Handbook of Instructed Second Language Acquisition* (1st ed., pp. 15–32). Routledge. https://doi.org/10.4324/9781315676968-2/

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

Ferman, S., Olshtain, E., Schechtman, E., & Karni, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics*, *22*(4), 384–412. https://doi.org/10.1016/j.jneuroling.2008.12.002

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Han, Z. (2012). Fossilization. In C. Chappelle (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford, UK: Wiley Blackwell. https://doi.org/10.1002/9781405198431.wbeal0436

Jean, G., & Simard, D. (2011). Grammar teaching and learning in L2: Necessary, but boring? *Foreign Language Annals*, *44*(3), 467–494. https://doi.org/10.1111/j.1944-9720.2011.01143.x

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

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

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

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

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, *26*(03). https://doi.org/10.1017/S0272263104263021

Lyster, R., & Sato, M. (2013). Skill Acquisition Theory and the role of practice in L2 development. In M. del Pilar García Mayo, M. Juncal Gutiérrez Mangado, & M. Martínez-Adrián (Eds.), *Contemporary Approaches to Second Language Acquisition* (pp. 71–92). John Benjamins. https://doi.org/10.1075/AALS.9.07CH4

Nakata, T. (2016). Effects of retrieval formats on second language vocabulary learning. I*RAL - International Review of Applied Linguistics in Language Teaching*, *54*(3), 257–289. https://doi.org/10.1515/iral-2015-0022

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

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? *Studies in Second Language Acquisition*, *38*(3), 523–552. https://doi.org/10.1017/S0272263115000236

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of Skill Acquisition and the Law of Practice. In J. R. Anderson (Ed.), *Cognitive Skills and Their Acquisition* (1st ed., pp. 12–66). Psychology Press.

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*(8), 1917–1927.

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

Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test-restudy practice: Implications for student learning. *Applied Cognitive Psychology*, *25*(1), 87–95. https://doi.org/10.1002/acp.1646

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology*: General, *140*(3), 283–302. https://doi.org/10.1037/A0023956

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review*, *24*(3), 419–435. https://doi.org/10.1007/s10648-012-9203-1

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*, *142*(4), 1113–1129. https://doi.org/10.1037/A0030498

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

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology: Applied*, *24*(1), 57–71. https://doi.org/10.1037/xap0000146

Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology*, *19*(3), 361–374. https://doi.org/10.1002/ACP.1083

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

Serfaty, J., & Serrano, R. (2022). Lag effects in grammar learning: A desirable difficulties perspective. *Applied Psycholinguistics*, *43*(3), 513–550. https://doi.org/10.1017/S0142716421000631

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

Swain, M. (1988). Manipulating and complementing content teaching to maximize second language learning. *TESL Canada Journal*, *6*(1), 68. https://doi.org/10.18806/tesl.v6i1.542

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford University Press.

Ullman, M. T. (2020). The Declarative/Procedural Model: A neurobiologically motivated theory of first and second language. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition.* (3rd ed., pp. 128–161). Routledge.

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 & Cognition*, *44*(6), 897–909. https://doi.org/10.3758/s13421-016-0606-y

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review*, *20*(6), 1239–1245. https://doi.org/10.3758/s13423-013-0434-z

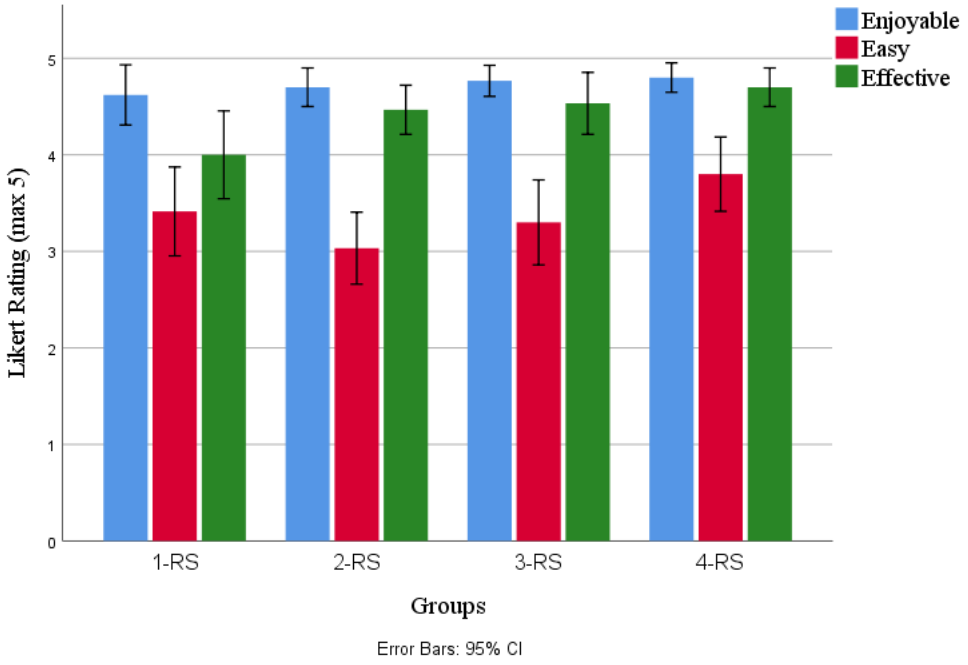
Webb, S. A. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *The Canadian Modern Language Review*, *65*(3), 441–470. https://doi.org/10.3138/cmlr.65.3.441

Zalbidea, J. (2021). On the scope of output in SLA: Task modality, salience, L2 grammar noticing, and development. *Studies in Second Language Acquisition*, *43*(1), 50–82. https://doi.org/10.1017/S0272263120000261

Appendix A

Perceived enjoyment, ease, and effectiveness of the treatment by group (1-5)

Group	N	Enjoyment	Ease	Effectiveness
1-RS	29	4.62 (0.82)	3.41 (1.21)	4.00 (1.20)
2-RS	30	4.70 (0.54)	3.03 (1.00)	4.47 (0.68)
3-RS	30	4.77 (0.43)	3.30 (1.18)	4.53 (0.86)
4-RS	30	4.80 (0.41)	3.80 (1.03)	4.70 (0.54)



Appendix B

Comments after the posttest indicating that declarative knowledge was regained during the receptive test

ID	Group	Productive Score	Receptive Score	Comment
609911e00dd4843e39cb88ac	1-RS	0%	17%	“My head went blank and I forgot most of the words, until they appeared on screen.”
6175804ea3428148dea5659b	1-RS	0%	42%	“Without seeing the conlang, I struggled to remember much. I could remember pluralization was formed with double final letters. I knew subjects were marked differently but couldn't recall how. Once I had a refresher translating from the conlang to English I recalled most of it, though...I think.”
60ff663e49cef209ba7f758f	2-RS	17%	50%	“I feel like I did okay, " I remembered more after translating from the made up language to English. Seeing the words reminded me.”

Chapter 6: Conclusion

I will now present a short summary of the four studies featured in this thesis, followed by a discussion of the implications of this work, as a whole, on theory, research methods, and pedagogy. The final section will detail some suggested experiments to further expand our understanding of issues raised by this work, namely the transferability of practice, the differential retention characteristics of productive and receptive knowledge, and the possible interaction between the distribution and quantity of relearning sessions. This section includes supplementary and anecdotal data that was not included in the publications.

6.1 Summary of studies

Study 1 (Serfaty & Serrano, 2020) examined accuracy development and retention through output practice using digital flashcards. The participants were from an undereducated and resource-poor background and required English in order to enter further education or find employment in the city. They had no access to formal English education and studied together with error-filled textbooks and intermittent short-term volunteers in an outdoor setting. Before the treatment, the 31 participants aged 9-17 completed a pretest of to-be-trained items. As a simple illustration of the type of progress that was made, some answers from that test are provided in Table 1.

TABLE 1
A selection of pretest answers from Study 1

Target response	**Sample Pretest Answers**		
He is playing volleyball	He Playing verybul	He playing volleyball.	He playing volleyball.
There are girls in my house.	There is a girl stay in my house.	In my house has One gril.	My house have a girl
Are the girls eating?	dose the girl is eating?	Girls eating	what gir eating
Is there a girl in the shop?	where are the girl in shop	in the shop have one girl?	What have a girl in the shop?

When speaking to peers in English, these types of errors might not impede communication and there is in fact an argument that a dialect of Cambodian English is emerging (Moore & Bounchan, 2010). However, it is clear that this level of accuracy, as compared to standard recognized forms of English, is not sufficient to meet any of these participants' education or employment goals, and that communication with foreigners would be far easier if their language conformed more closely to the English spoken by people outside of Cambodia.

In light of these goals, eight types of simple sentences were practiced using the Cram.com app, with five exemplars per structure. Participants saw the sentence in their L1 Khmer and typed the English translation. The flashcard software showed them feedback and repeated any incorrect responses in the next cycle. Each structure was repeated the following day and

on one additional day. The number of structures per day increased so that they were practicing four structures simultaneously by the end of the training. After two weeks, participants completed a posttest of trained items (2 per structure) and equivalent but untrained items (2 per structure). A third of participants scored 100% in both tests, with the rest also scoring highly. A second test after 14 days revealed no statistically significant changes in results. A final 18-week posttest was performed, with half the students completing a review session the week before. With the review, gains were equal to the immediate posttest, and without the review there were only a couple of additional errors. A control group in the same setting obtained 0% gains from pretest to posttest without the flashcard training. The important findings were that learned knowledge transferred to untrained items and that gains were durable after a substantial delay. Moreover, the intervention was successful without any associated linguistic instruction or guidance.

Study 2 (Serfaty & Serrano, 2022) aimed to shed light on the possible distribution effects for this type of grammar practice. It also extended the research into how flashcards can be used by including high-proficiency participants, very advanced grammar features, and by using L2 scenarios as cues instead of L1 translations. Two structures were studied by 117 secondary school students, aged 10-18, in an English-medium international school. The structures were complex to meet the needs of the students who were following a British curriculum. Each structure was studied via Quizlet on two days: one structure with a 1-week interval and the other with a 1-day interval. Participants were tested on their retention of these structures after either one week or one month. Globally, no difference in condition was found. However,

for individual students, one condition was better than the other. For students with lower English proficiency and who required more time to learn and relearn these structures, the shorter interval was better, whereas the reverse was true for higher proficiency and faster students. The important finding of this paper was that results met the predictions of the Desirable Difficulty Framework as specified by Suzuki et al. (2019) in that the more difficult condition was only desirable when other factors of difficulty were low.

Study 3 (Serfaty & Serrano, in review) extended the above findings by repeating the former experiment with vocabulary items. Of the 96 students analyzed in this study, 77 took part in both experiments. The aim was to elucidate the differences in lag effects between grammar and vocabulary practice when controlling for the setting, participants, and exact methodology. The two categories of vocabulary were Hebrew words in order to control for prior knowledge and other exposure to the targets. The experiment was part of a wider project involving memory and aptitude tests. Results showed that the 1-week interval was always slightly better than the 1-day interval for vocabulary learning, with a very small effect. From a psycholinguistic perspective, this is an interesting contrast with the grammar experiment. However, the difference in productive scores between the two conditions was small. A more meaningful difference was seen in the receptive test, according to which the longer interval facilitated better retention at the 4-week delayed posttest. It was speculated that receptive knowledge was better retained between sessions and was therefore reinforced during the relearning session. In contrast, productive knowledge may have been lost between sessions and thus was encoded at the relearning session as if it was the first encounter. The

important findings of this paper were that a longer intersession lag led to higher scores in a classroom setting, that receptive and productive knowledge were differently affected by lag, and that lag effects were different for vocabulary and grammar.

Study 4 (Serfaty & Serrano, resubmitted) addressed the question of how much grammar practice is actually necessary. Study 1 (Serfaty & Serrano, 2020) included three practice sessions plus an immediate posttest with ceiling scores, which acted as a fourth practice session. Knowledge was well retained from this schedule. However, in Study 2 (Serfaty & Serrano, 2022), scores were very low after only two sessions. There were many methodological differences between these studies, such as the difficulty of the structures, the likelihood of using these structures, and while the first study included extremely motivated learners attending the sessions through independent choice, the second study was conducted in a privileged classroom among already-proficient English users who may have been less intrinsically motivated to learn. Thus, the issue of quantity was not the only factor that could explain the difference in outcomes.

Using an online design, adult participants from across the globe completed two, three, four, or five sessions of learning an artificial language in the style of commercial digital flashcards. The main difficulty in the language was in encoding subject and object pronouns. Training data showed a familiar trend for the error rate, according to which accuracy improved sharply in the initial stages, followed by a more gradual improvement until finally performance was stable. Analyzed by the number of sessions completed, a threshold of high retention was found after the fourth session. After this point,

no further improvement was detected. However, when re-analysed according to how many sessions the participants performed with the minimum number of trials, without considering the total number of sessions, results were predicted with larger effects. Any participant with two minimum trial sessions (MTSs) received high scores, similar to those found in Study 1 (Serfaty & Serrano, 2020), with no further improvement after more than two MTSs. Participants with only one MTS achieved intermediate scores while those that never performed an MTS achieved lower scores, similar to those found in Study 2 (Serfaty & Serrano, 2022). The important implication of this finding is that high retention could be predicted at the training stage. Different tasks may require a different number of sessions, but if a learner knows the indicators of long-term knowledge, they could be sure to get enough practice. If this hypothesis is correct, schools could employ individualized learning schedules rather than fixed schedules. Instead of having some students pass and some students fail, all students could pass eventually. The only difference would be that the more able students could cover more material in a shorter time.

6.2 Implications of the current work

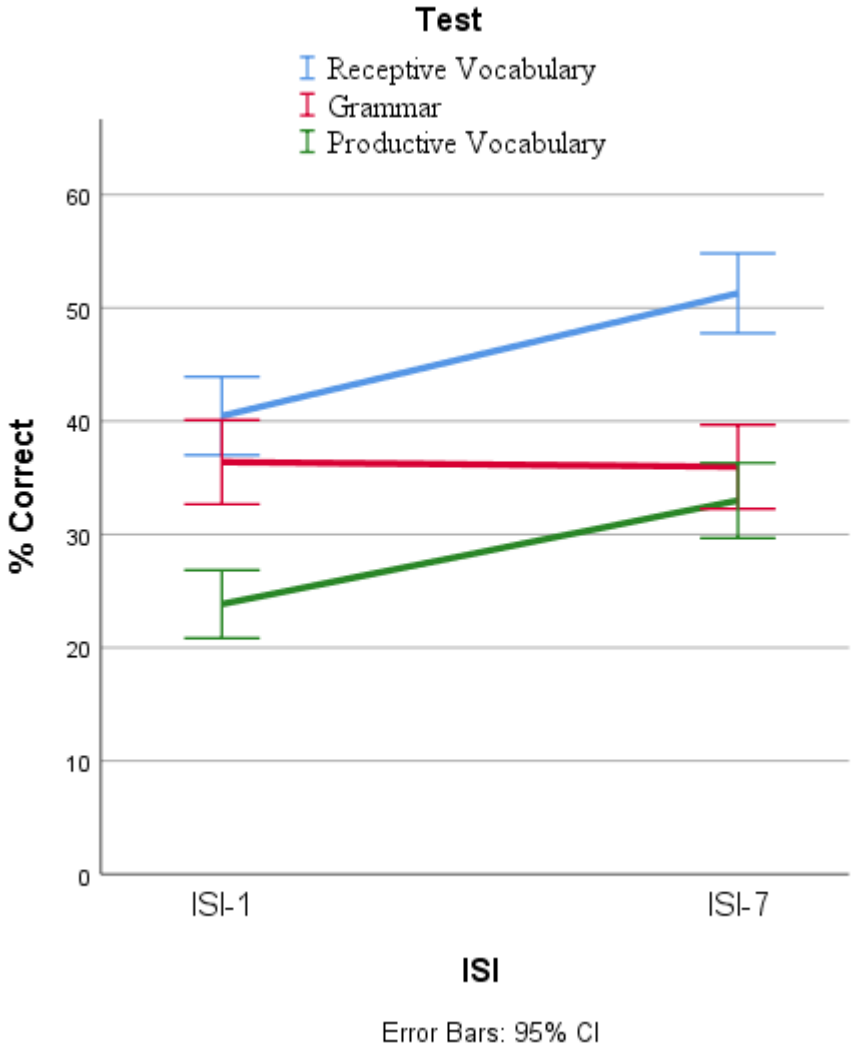
6.2.1 IMPLICATIONS FOR THEORY

6.2.1.1 Grammar and vocabulary

Some points are worth highlighting in terms of implications for theories of L2 learning and practice. The first is a confirmation that lag effects are different for grammar and vocabulary learning, as has been speculated in the literature (Li & DeKeyser, 2019; Ullman & Lovelett, 2018). The current

work reported a consistent lag effect for vocabulary (Study 3: Serfaty & Serrano, in review) but not for grammar (Study 2: Serfaty & Serrano, 2022). Having used the same number of items, intervals, and participants for both experiments, it was possible to combine these results into a single chart. Figure 1 visualizes the differences in lag effects found for grammar and vocabulary learning.

FIGURE 1
Scores on vocabulary tests (productive and receptive) and a grammar test, after two sessions of learning through Quizlet with an intersession interval (ISI) of either one day or seven days, with a maximum score of 8 per test.



One explanation mentioned in Study 3 (Serfaty & Serrano, in review) is that vocabulary is simpler than grammar and so the longer lag added desirable difficulty (Bjork, 1994) to vocabulary learning more consistently than it did for grammar. However, another explanation could be found in Skill Acquisition Theory (SAT).

According to models of skill acquisition (DeKeyser, 2020; Kim et al., 2013), declarative knowledge is usually learned first, followed by a period of proceduralization in which the learner gradually depends less on declarative knowledge and more on procedural knowledge, before finally the learner is able to rely fully on procedural knowledge and produce language without conscious effort. For instance, the word "work" would be rapidly stored in declarative memory as well as the rule of adding *-ed* for the past tense. As the learner practices this *-ed* rule repeatedly, procedural memory would encode the sequence and eventually take over from declarative memory, allowing the speaker to use the regular past tense without consciously retrieving the grammar rule.

Regarding the present data, the vocabulary experiment (Study 3: Serfaty & Serrano, in review) may have only tested declarative knowledge, whereas the grammar experiment (Study 2: Serfaty & Serrano, 2022) potentially involved procedural knowledge. Repeated productive practice of the same L2 target structure, as occurred in the initial training sessions, may have facilitated L2 proceduralization. This assumption is supported by data from Study 4 (Serfaty & Serrano, resubmitted), showing the learning curve associated with proceduralization for accuracy improvement from an experiment using the same method of learning. Decreasing response times

have also been used as evidence of this process in studies involving a similar training after two sessions (Li & DeKeyser, 2019; Suzuki, 2017; Suzuki & DeKeyser, 2017), or even within a single session (Lambert et al., 2017; Suzuki, 2021). Although the measure of time on task used in Study 2 was not as refined as response time data, there was a clear decrease in time on task in the third session, which involved previously studied structures.

Assuming that the grammar knowledge was proceduralized to some extent, our findings would be in line with claims that these types of knowledge are differently influenced by lags. Most pertinently, Li and DeKeyser (2019) separated their target skill of learning Mandarin tones into declarative knowledge (vocabulary matching) and procedural knowledge (pronunciation of novel items) and found a longer lag advantage for the former at RI-28, but an advantage to the shorter lag for the latter at both RIs. Similar grammar studies that sought to examine the proceduralization of grammar knowledge (Suzuki, 2017; Suzuki & DeKeyser, 2017) also found no advantage to a longer lag. In contrast, for studies examining only the receptive knowledge of grammar rules (Bird, 2010; Rogers, 2015), which according to the D/P model are stored in declarative memory (Ullman & Lovelett, 2018), the lag effect was found.

Vocabulary retrieval is also subject to a reduction in reaction times and errors through extensive practice, resulting in automatized declarative knowledge (Segalowitz & Segalowitz, 1993; Segalowitz et al., 1998). Skill Retention Theory (Kim et al., 2013) predicts that after knowledge reaches asymptotic performance in accuracy and speed, it is automatized and durable. One might then expect the same level of automatization, and therefore

durability, for both grammar and vocabulary, assuming the same number of sessions and using the same method of learning.

However, there is reason to doubt this assumption. Once a vocabulary item has been retrieved within one session, the item cannot be processed again as it is already accessible in working memory (Callan & Schweighofer, 2010). In order to extensively practice vocabulary items, the learner would need to have enough time between each retrieval attempt for the item to be forgotten to some extent. This would require many sessions over a long period of time. In comparison, grammar flashcards allow the learner to retrieve the same target in different iterations. The complexity and variety of each item allows for repeated effortful practice within one session. For example, in Study 2 (Serfaty & Serrano, 2022) participants studied each target in eight different iterations, while in Study 4 (Serfaty & Serrano, resubmitted), there were 12 iterations of the structure per session. The specific sentences might be different, but the target skill is in the application of rules. These rules cannot be retrieved verbatim and thus grammar learning requires cognitive effort on every trial. Therefore, grammar structures can be extensively practiced within one session whereas vocabulary items cannot.

In Study 4 (Serfaty & Serrano, resubmitted), participants on average achieved the desired threshold of learning after retrieving the target structure on four different days with 12 items per day. Therefore, these participants applied the target rules with 100% accuracy 48 different times (12 items x 4 sessions). This is remarkably similar to findings from the application of a coding language in which procedural knowledge was identified after around 50 applications (Anderson et al., 1997). This can be taken as a preliminary

indicator of the number of successful attempts needed to achieve durable knowledge. If the same quantity applies to vocabulary retrieval, then a learner would need to retrieve each item around 50 times as well. As previously stated, a true retrieval would require some time to pass before the item is not accessible from working memory. Therefore, in order to achieve automatised vocabulary knowledge, the learner would need to retrieve each item on 50 different days. The true number may be lower than this, but the point stands that vocabulary knowledge might require many more sessions than grammar to achieve the same durability.

This distinction between the depth of learning possible for grammar and vocabulary within a limited number of days could affect the type of knowledge obtained. After several sessions, vocabulary would still be stored as non-automatised declarative knowledge whereas grammar might be stored as semi-automatized procedural knowledge. It has been claimed that automatized procedural knowledge is not subject to a lag effect, since it is strengthened through repeated application, whereas declarative knowledge is aided by longer lags that induce effortful retrieval (Kim et al., 2013; Li & DeKeyser, 2019; Ullman & Lovelett, 2018).

6.2.1.2 Productive and Receptive knowledge

Another important finding from this thesis refers to the differential practice effects on productive and receptive knowledge. Productive knowledge was defined as the ability to generate the L2 from memory with a specific meaning in mind, while receptive knowledge is the ability to comprehend or recognize the L2, which can be shown by expressing this meaning in the L1 (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; McLean &

Hogg, 2013; Nakata, 2020; Webb, 2005). Articles situated within the SAT framework have often asserted that practice is skill specific, as in the following example from Kachinske (2021): "*[...] practice should be skill specific; once knowledge has been proceduralized in one skill, for instance comprehension, it becomes more difficult for that knowledge to be generalised in another skill, for example production. In other words, in order to develop receptive knowledge, learners need practice comprehending input, and in order to develop productive knowledge, learners need to practice producing language* (p31)." These claims are based on studies in which receptive practice was not as effective for the productive skill, but these studies have always shown that productive training also leads to high receptive scores, sometimes higher than from the receptive training (DeKeyser, 1997; de Jong, 2005; Webb, 2009). It is only in productive tests where a meaningful difference between directions of training is found. Cross-sectional studies involving both productive and receptive testing show that receptive knowledge usually develops before productive knowledge (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; Rodgers, 2011).

This issue of skill-specificity was indirectly addressed in the studies that measured both types of knowledge for grammar (Study 4: Serfaty & Serrano, resubmitted) and vocabulary (Study 3: Serfaty & Serrano, in review) after purely productive practice. It was found, in line with previous research, that receptive scores were higher in both studies. Receptive knowledge was also better maintained at the delayed posttest (Study 3: Serfaty & Serrano, in review). Posttests from Study 4 (Serfaty & Serrano, resubmitted) showed that receptive knowledge had already peaked after two sessions but that productive

knowledge peaked after four sessions. Therefore, all of the data suggest that practice is not skill specific, but that the weaker form of knowledge (receptive) develops before the stronger form of knowledge (productive). Moreover, Barclay and Pellicer-Sanchez (2021) showed that recall ability decayed faster than recognition ability, just as productive knowledge decayed faster than receptive knowledge in Study 3 (Serfaty & Serrano, in review), suggesting that more difficult kinds of knowledge are acquired more slowly and decay faster. In light of this body of evidence, the issue of skill-specificity should be re-interpreted in terms of desirable difficulty. If practice is easy (receptive or recognition), then only weak knowledge will develop, but if practice is desirably difficult (productive or recall), then stronger knowledge can develop.

6.2.1.3 Learner-related difficulty

Finally, the studies included in this thesis have shown that learner-related difficulty is the most important factor in determining the appropriate difficulty and quantity of practice. Drawing on Bjork's Desirable Difficulty Framework (1994) and the sources of difficulty in L2 practice proposed by Housen & Simoens (2016), Suzuki et al. (2019) theorized that the optimal practice conditions depended on a combination of linguistic difficulty, learner-related difficulty, and practice conditions. The two structures used in Study 2 (Serfaty & Serrano, 2022) may not have been different enough for a statistically significant interaction to be detected. The first structure was the future perfect progressive (e.g., *I will have been reading this article for twenty minutes by the time I take a break*) and the interrogative third conditional (e.g., *What would you have done if you hadn't started reading this article?*). The first was considered less complex because it involves fewer transformations,

more repeated chunks, and is conceptually less abstract. However, both structures involve two clauses with many opportunities for errors, and neither were expressible specifically in the participants' L1 (roughly translated: *I read this article for twenty minutes then take a break; What do if don't read this article?*). The two age groups (10-12 and 13-18) also may not have been different enough to produce an interaction with spacing, since all participants were in secondary school. However, lag interacted significantly with the participants' prior knowledge, measured by an English proficiency test, and by their ability to complete the task, measured by their total time on training.

Similarly in Study 4 (Serfaty & Serrano, resubmitted), learner-related difficulties during training proved to be a much better predictor of posttest scores than the number of sessions. While the mean number of sessions needed to cross the threshold of lasting knowledge was four, final outcomes were better predicted when participants were grouped by the number of sessions completed with a minimum number of trials.

Taken together, these studies underscore that results mostly depend on the participants, not the training conditions, and rather than trying to specify ideal amounts of time between sessions or amounts of sessions, the goal of research should be to accurately measure when knowledge has become durable for the individual learner.

6.2.2 IMPLICATIONS FOR RESEARCH METHODS

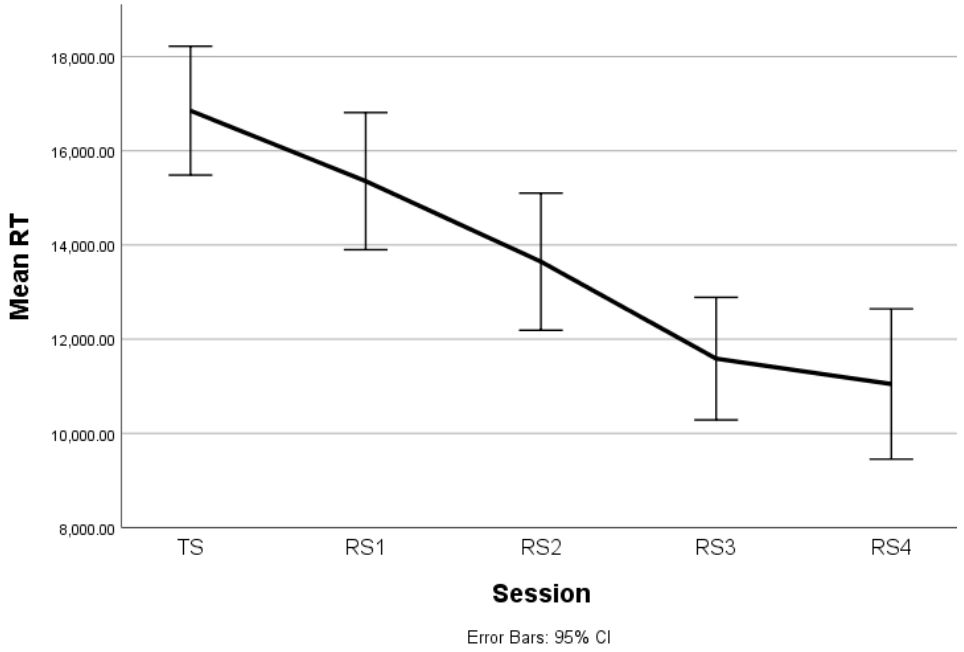
The research undertaken in this thesis has spotlighted digital flashcards as a promising component for future methodologies investigating L2 practice through quantitative variables. Firstly, from an analytics point of view, the

treatment of digital flashcards has enabled the detection of statistically significant differences between conditions, with a range of effect sizes, in areas where differences have not been found before, namely in lag effects for grammar learning (Study 2: Serfaty & Serrano, 2022) and for vocabulary learning in a classroom (Study 3: Serfaty & Serrano, in review). This could be attributed to the simple and systematic nature of the learning, preventing many extraneous variables from distorting the effects under investigation. Secondly, digital flashcards allowed for research to be conducted among atypical samples for the field of applied linguistics (Shepperd, 2022). In one case, research was conducted in a remote rural community in an environment lacking in technology, teachers, or physical infrastructure for learning (Study 1: Serfaty & Serrano, 2020), and in another case research was conducted simultaneously with 119 participants from 27 different countries (Study 4: Serfaty & Serrano, resubmitted). Thirdly, no human instruction is needed for these studies, which reduces the cost of research and the amount of time spent on data collection for researchers. For these reasons, digital flashcards have proved to be a valuable tool for this type of research.

In particular, flashcard training with an unsupervised online experimental design on Gorilla (Study 4: Serfaty & Serrano, resubmitted) produced the learning curve of accuracy associated with proceduralization. Response times, though not included in the publication due to confounds as a result of the chosen experimental design, also followed the expected learning curve, as shown in Figure 2.

FIGURE 2

Response times for items answered correctly on their first attempt for each training session of Study 4



This supports the validity of online platforms for studying the proceduralization of L2 grammar, removing significant practical barriers to performing longitudinal studies in L2 acquisition. Participants could engage in practice in a place and at a time of their convenience, just as they would with a language learning app. By recruiting through an online platform, such as Prolific, the researcher could filter participants by attributes of interest and immediately replace any participants lost through attrition. Therefore, when funding and time are limited, online research could provide valuable data.

As a recommendation for future studies, the validity of such experiments could be further improved by imposing time limits on the responses and the presentation of feedback. While this would reduce ecological validity, it would help to ensure that participants complete their responses as quickly as possible and are not spending variable amounts of time analyzing and rehearsing the feedback. Any response not completed within this time limit could be considered incorrect and the resulting data would not

contain any anomalies due to participants being interrupted or distracted during the training. For reference, the mean response time in the initial training session for incorrect responses was 17 seconds and for correct responses was 15 seconds. Therefore, a limit of 20 seconds would be conservative.

6.2.3 IMPLICATIONS FOR PEDAGOGY

Since the studies included in the present thesis used a tool for L2 practice that is ecologically valid and increasingly used in L2 classes, a number of direct pedagogical implications can be drawn.

6.2.3.1 Plugging a gap in undeveloped educational contexts

Firstly, flashcards are effective for L2 grammar learning and can be used in low-resource environments without any need for explicit instruction (Study 1: Serfaty & Serrano, 2020; Study 2: Serfaty & Serrano, 2022; Study 3: Serfaty & Serrano, in review). In cases where human instructors are not able to provide optimal teaching, such as in developing countries with underdeveloped educational systems, automated learning apps could be a good solution for acquiring and practicing accurate L2 grammar knowledge. In these settings, the ability to communicate effectively in a dominant L2, often English, can lead to life-changing employment opportunities that would otherwise not be available (Haidar, 2019; Hamid, 2016; Moor & Bounchan, 2010). This can be especially challenging with little exposure to the L2 and from a typologically distant L1 (Muñoz & Cadierno, 2021). In Study 1 (Serfaty & Serrano, 2020), the experimental group made high gains in two weeks of more than 80%, compared to 0% gains in a control group over the

same length of instruction and the same non-experimental exposure to the L2, and these gains were still present after 18 weeks. The difference was that the experimental group received accurate input, sufficient amounts of retrieval attempts, and consistent feedback, with no time limits.

The benefits of controlling input and output through technology might not be evident to many scholars. Much of the extant L2 classroom research has taken place under privileged circumstances in which the expertise of the teacher is taken for granted. In contrast, many classrooms around the world are completely non-interactive and rely exclusively on rote learning (Kim, 2005; Venkataramanan, 2016; Visal et al., 2022), while the teachers might have little to no knowledge of the subject matter. In Cambodia, as our example, the educational system promotes rote learning and conducts monthly exams for all grade levels (MoEYS, 2022). Low test scores will result in students repeating that grade level. Until recently, it was common practice for students to pay the teacher for answer keys in order to pass these exams, a problem that persists in rural public schools (Maeda, 2021). The teacher's role is therefore to read or write the target information for students to copy and be tested on, before immediately moving onto the next part of the textbook.

A positive intervention in a wealthy country is useful, but a positive intervention in a developing country can transform a student's educational achievement, cognitive abilities, and future prospects. It is far easier to install an internet connection than it is to train a teacher, who has themselves been raised in an underdeveloped educational system and may not be able to apply 21st century pedagogy (Visal et al., 2022). Many basic cognitive skills might not ever be taught, but software can be designed to elicit analytic thinking and

control the accuracy of the subject matter being learned. While an effort is being made to enhance teacher training (Pearson, 2022), educational software represents a promising short-term solution for these students.

6.2.3.2 Individualized learning schedules

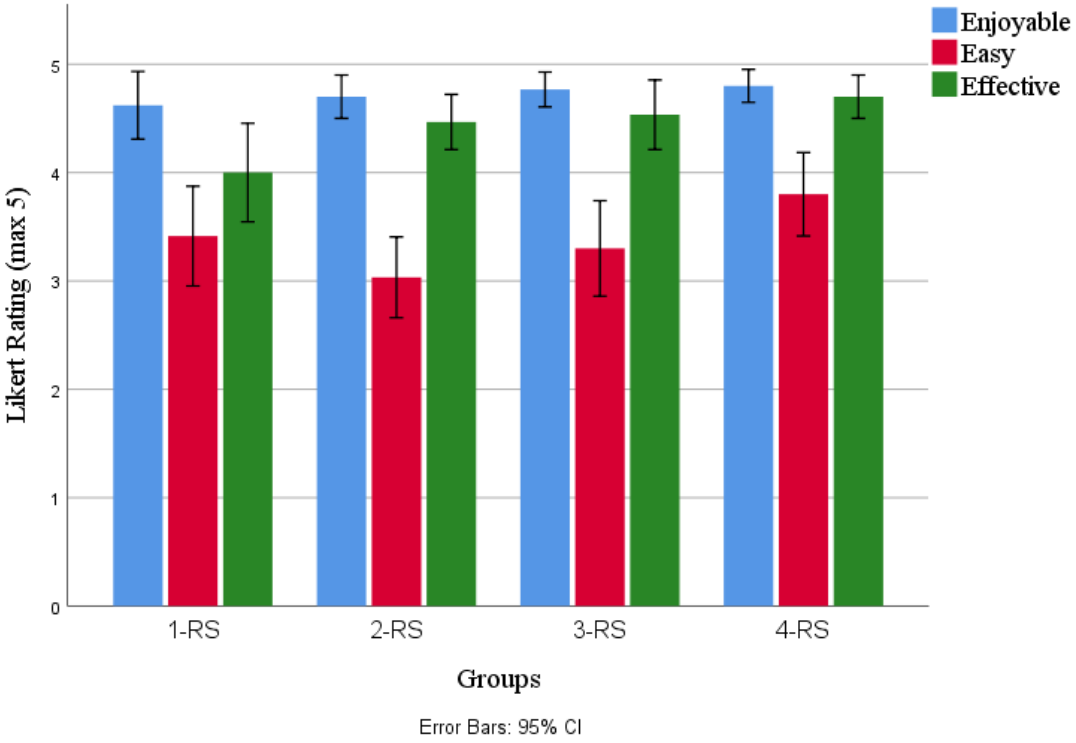
On a more general scale, this thesis has shown the importance of individualized learning. In a traditional classroom, there will always be slower students that fall behind quite quickly and can then never catch up. Study 2 (Serfaty & Serrano, 2022) showed that learners required very different amounts of time to complete an identical task, from a few minutes to an hour. Study 4 (Serfaty & Serrano, resubmitted), demonstrated that adults, presented with an identical task, required between 12 and 53 trials, without including anomalies. Some participants could complete an error-free session immediately while others still could not do so on their fifth session. Using individualized tasks in classrooms would ensure that every student gets the amount of time and practice they need on every point and allow students to learn at their own pace.

6.2.3.3 Gamification of learning

Finally, practicing grammar through digital flashcards is perceived positively by learners. A survey of participants from Study 4 (Serfaty & Serrano, resubmitted) asked about their level of enjoyment, perceived ease, and perceived effectiveness of the method, as shown in Figure 3. Whether participants repeated practice once, twice, three, or four times, the method was perceived as enjoyable, effective, and suitably challenging when rated on a 5-point scale.

FIGURE 3

Perceptions of enjoyment, ease and effectiveness of learning a language through digital flashcards among participants from Study 4



Comments from the participants also revealed a significant amount of emotional investment in the study, on both ends of the spectrum. All comments can be found in Appendix S4 in Chapter 5. A selection of comments are presented in Table 2**,** chosen to exhibit the range of emotions expressed.

Anecdotal evidence supports this, as a school in Cambodia is currently using this method with all secondary school students on a weekly basis. The target sentences utilize the vocabulary and subject matter of the unit they are studying and the reading materials associated with their topic are designed to include these target structures. For example, in Grade 7, the first topic is

TABLE 2
A selection of comments made after the posttest from Study 4

Productive Test Score	**Answer to "How do you feel?" after the posttests.**
100%	Pretty good! This was fun, I'm almost sad to let it go now. I hope I did okay.
100%	I feel great! I hope I got most of those right...
92%	Proud of myself
83%	a little bit stressed, it was hard to remember after some time
42%	gutted I didn't remember some of the words for the first part
33%	i feel ashamed for forgetting how "have" was translated

healthy food and one of the target structures is the present simple. The flashcards contain sentences such as "Eggs have a lot of protein" and "Cake has a lot of sugar". The flashcards are therefore designed to target not only language but to reinforce content knowledge. The activity is one of many options presented for independent online learning, yet all students choose to engage in it. Unit tests have proven the students' overwhelming success in learning the target structures, though the transferability of this knowledge to communicative language use has not yet been explicitly tested.

Many educators and scholars might assume that this type of task would be boring or arduous for young learners. This could be because the closest thing to digital flashcards that they have experienced is simple paper and pen translations or paper flashcard drills. Flashcards apps are in fact more akin to a game. Learners see immediate feedback tailored to what they wrote and see constant statistics of their progress. It was observed during the data collection process for the classroom studies (Study 2: Serfaty & Serrano, 2022; Study 3: Serfaty & Serrano, in review) that students like to announce how many items were correct in each round and compare this figure with friends. There were many excited exclamations at finally achieving a correct response as well as groans of mock despair after making an error.

Additionally, although the particular mode of form-recall used in this thesis was chosen for methodological control, Quizlet can be used as an exciting team game in which the correct answer could appear on any of the team members' screens. Similar websites like Kahoot!, Quizizz, and Baamboozle offer more gamification features, award points, and incorporate humor. For example, both Quizizz and Baamboozle use memes to show success or failure.

6.3 Future directions for research into L2 practice

Some of the future directions specified within the published papers have already been addressed within this thesis, such as the use of digital flashcards among learners with different proficiencies and backgrounds, and the optimisation of learning through distribution and frequency. There are some gaps mentioned that have not yet been filled. Study 1 (Serfaty &

Serrano, 2020) mentioned some possible research into how digital flashcard learning could affect more open and spontaneous language production. Study 3 (Serfaty & Serrano, in review) raised questions of how receptive and productive knowledge might be differently retained between sessions and how this might affect the differential lag effects of these two kinds of knowledge. Study 4 (Serfaty & Serrano, resubmitted) also mentioned further study in how productive and receptive knowledge might be acquired at different rates and be differentially retained. Additionally, it was recommended that a replication of the study is needed with a longer retention interval, since two weeks is not long enough to be useful if the goal is long-term L2 learning. Two "minimum trial sessions" (MTSs) was good for two weeks, but perhaps more practice would be needed to achieve retention after 18 weeks or 52 weeks. That paper also speculated as to how results might differ with a longer interval between study sessions, which would induce more forgetting between sessions. More forgetting could affect how quickly learners cross the threshold of learning necessary to achieve durable knowledge. In sum, the experiments have highlighted directions for future research which can be summed up as (1) transfer of gains from digital flashcard learning to creative language use, (2) productive and receptive retention properties, and (3) interactions between session distribution, testing intervals, and the number of relearning sessions necessary to achieve mastery. The remainder of this section will outline some possible experimental designs to address these three areas.

6.3.2 TRANSFER

Practicing grammar in a controlled manner would only be useful if it facilitates accuracy in more open-ended and spontaneous language activities, for example in a spoken task or in open writing. There is reason to think that such knowledge can transfer to more creative linguistic output. After the posttest of Study 2 (Serfaty & Serrano, 2022), participants were asked to write their own sentence, as part of a short conversation or a paragraph, using the target structures, in order to preliminarily explore whether participants could apply the learned rules to their own original meaning. It was observed that any student with high scores in the posttest was also able to create a perfectly grammatical and original sentence using the target structure. Some examples of these sentences are presented in Table 3. It is notable that in these examples, the language around the target structure contains errors, but the target structure itself is flawless, with the exception of capitalization.

This is a starting point in that it shows the transferability of knowledge to non-cued usage. However, this task immediately followed a test that focused on producing these structures and these structures were specifically requested. It does not show whether the structures would appear in more open-ended language production when not specifically prompted and after a reasonable delay. In order to test whether digital flashcard training would affect free language use, the following experimental design is proposed:

TABLE 3
A selection of sentences written by participants of Study 2 after the posttest

Structure A
“I will start learning abroad in 2027. I will have get my master degree in 2033 **I will have been studying abroad for 6 years by the time I get my master degree.”** - *Participant from Grade 6*
“I will have been drawing for 30 minutes by the time i get to color it.” - *Participant from Grade 8*
"You: Hello! What have you been? Them: I've been playing the piano lately. I will continue to play piano in 2020. I will attend the piano competition in 2025. You: That's a very long time. Them: yeah, **I will have been playing the piano for 5 years by the time I attend the piano competition." -** *Participant from Grade 10*
“I had a dream about my mom will have getting me a car on my birthday, but she said that I'm too young. **I will have been studying in college for 2 years by the time my mom agrees to buy me a car.**” - *Participant from Grade 11*
Structure B
"I was sick, so I didn't go anywhere ." said Jamie. "where would you have gone if you hadn't been sick?" asked Carl. "**I would have gone to Disney land if I hadn't been sick."** said Jamie.” - *Participant from Grade 6*
“A: You didn't review for the test, so you failed it. B: **Would you have failed if you had reviewed for the test?**” - *Participant from Grade 7*
“You: Hey! Them: Hello, I'm kind of sad that you didn't pick me up yesterday. You: Oh my gosh! I'm so sorry. I was busy with my work, so I forgot to pick you up. Them: **Would you have picked me up if you hadn't been busy with your work?** You: Of course! I hope you can forgive me. Them: It's okay, I forgive you.” - *Participant from Grade 10*
“Person 1: I woke up very depressed today, so I didn't went to eat. Person 2: **Where would you have gone if you hadn't woken up depressed?** Person 1: Not sure, It depends.” - *Participant from Grade 11*

Participants are asked to write an essay that elicits a specific structure. For example, to write about a past event to elicit the past tense, or a hypothetical future to elicit the conditional. Following this task, the participants engage in digital flashcard training. This would ideally be over many sessions to replicate the type of mastery found in Study 4 (Serfaty & Serrano, resubmitted). After a sufficient delay, because we are interested in durable improvements, the participants are given the essay task again, perhaps with a slightly different topic. The two essays can then be analyzed for the presence and accuracy of the target structure. An improvement score could be derived and compared to a control group that did not engage in digital flashcard training, but were exposed to the items from the flashcards through traditional pedagogy.

A similar design could target comprehension skills. Rather than writing an essay, the participants could be asked comprehension questions from a reading passage or audio extract containing the target. A different passage or extract would be used for the posttest, counterbalanced. The treatment would be the same as the one outlined above. Another experiment could target speaking skills through a task that elicits a specific structure. This might be easier to achieve with more basic structures among participants with a more limited language range, since it is difficult to elicit specific structures in communicative tasks. As an example, the structure might be the interrogative present simple for beginner learners, and the task could be to interview someone about their daily habits (*What do you do after school? What time do you eat dinner?*). As before, accuracy before and after digital flashcard

training would be used to derive a score to be compared with a control group.

The research questions might be as follows:

1) Does flashcard training lead to higher presence and higher accuracy of target structures in open-ended language production?
2) Is this effect stronger from flashcard training than from traditional teacher-led pedagogy?

6.3.3 PRODUCTIVE AND RECEPTIVE RETENTION PROPERTIES

Study 4 (Serfaty & Serrano, resubmitted) showed that receptive knowledge was acquired and retained after only two sessions. Study 3 (Serfaty & Serrano, in review) found a stronger lag effect for receptive knowledge and speculated that receptive knowledge was better retained between sessions. To test these claims explicitly, an experiment could include a test at the start of each training session. This would require the use of research-focused software such as Gorilla, in order to manipulate the procedure and record the responses. Whether for vocabulary or for grammar, the same experimental design could apply:

Participants study flashcards during the first session and a different category in a different session, in order to create two different intersession intervals (ISIs). In the relearning session, involving both categories, they begin with a productive test, followed by a receptive test, followed by training for any items not remembered correctly in the productive test. Only these incorrect items would need to be part of the training because in the previous

design, a correct response at the beginning of the session causes an item to drop-out.

The first metric of interest would be the accuracy on both tests, which would show definitively whether some words were remembered receptively but not productively, as hypothesized. This metric would then be compared between the two ISI conditions, to check whether receptive knowledge was equally well-retained from the different intervals.

The second metric could be response times. It would be expected that productive tests induced longer response times on correct answers after a longer lag, but the key insight would come from words answered incorrectly in the productive test but correctly in the receptive test. If these words induce longer response times from the 7-day ISI compared to the 1-day ISI, this would confirm the hypothesis that successful receptive retrieval is more effortful after a longer lag, even when productive knowledge has decayed. This would explain why the lag effect was stronger in receptive knowledge in Study 3 (Serfaty & Serrano, in review).

A third metric could be a comparison of the incorrect responses on the productive test to the target responses. By assigning point values to correct letters (1 point) and correct letters in their correct positions (2 points), assuming that all items are of equal length, a similarity score could be computed. If ISI-1 items have a higher similarity score than ISI-7 items, this would confirm that ISI-1 words were closer to being remembered, even if they were not fully retrievable.

The final stage of this experiment would be a posttest. Statistical models could use each of these metrics as a predictor variable with posttest

scores as response variables. A comparison of the effects would indicate which of these metrics is best for predicting lag effects and retention, which would go on to inform theoretical accounts of L2 learning and memory.

The research questions could be as follows:

1) Is receptive knowledge better retained than productive knowledge after ISI-1 and ISI-7?
2) Is successful receptive retrieval more effortful after a longer ISI?
3) Are unsuccessfully retained items closer to being remembered after a shorter ISI?
4) Which of these factors best predicts retention of receptive and productive knowledge at posttest?

6.3.4 INTERACTION BETWEEN SESSION DISTRIBUTION AND QUANTITY FOR ATTAINING MASTERY

Much of the work in Skill Acquisition Theory has concentrated on the process of acquiring a skill, but not on the rate of acquisition or on how retention depends on the mastery achieved during training. Kim et al. (2013)'s Skill Retention Theory states that retention is longer after a skill has been automatized. Study 4 (Serfaty & Serrano, resubmitted) focused on how long this process takes, using ISIs of one day. A replication is needed using different ISIs and different retention intervals (RIs). The experimental design could otherwise remain identical. A possible design is proposed in Table 4.

TABLE 4
Conditions in the proposed experimental design for researching the interaction between ISI, RI, and the number of minimum-trial sessions required to attain mastery

Groups	Number of sessions	ISI	RI
1-4	2,3,4,5	1	14
5-8	2,3,4,5	7	14
9-12	2,3,4,5	1	28
13-16	2,3,4,5	7	28

By comparing results between these 16 groups, the following research questions could be addressed:

(1) Is the rate of improvement in accuracy different when studying at ISI-1 or ISI-7?

(2) Is the number of minimum-trial sessions needed to achieve durable knowledge different for ISI-1 and ISI-7

   (a) when tested at RI-14?

   (b) when tested at RI-28?

(3) Which combination of ISI and number of sessions leads to mastery in the fewest trials?

On the one hand, ISI-7 could lead to more forgetting, thus slowing down progress and requiring more sessions before asymptotic accuracy is achieved. On the other hand, ISI-7 might lead to better retention of declarative knowledge (the rules) due to the lag effect. The insights gained from such a

large-scale study would enable a theory of mastery and retention to be developed within the SAT framework. Such a theory would have wide-reaching implications for language learning and education in general. Of course, the same design could be replicated for vocabulary learning, fluency practice, or even in other domains of education and training.

6.4 Closing Remarks

This thesis has presented four studies of L2 practice using digital flashcards. The studies have collectively addressed issues regarding retention of learned knowledge, optimisation of practice, and differences between types of L2 knowledge. Thanks to the use of digital flashcards as the tool, these studies are easily replicable under different conditions of interest and also allow the findings to be applied by real language learners. A road map of possible follow-up research has been described. It is hoped that findings from this research have a positive impact for all L2 learners, but especially for those that depend on language learning in order to secure a bright future.

References

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 932–945. https://doi.org/10.1037/0278-7393.23.4.932

Barclay, S., & Pellicer-Sánchez, A. (2021). Exploring the learning burden and decay of foreign language vocabulary knowledge. *ITL - International Journal of Applied Linguistics*, *172*(2), 259–289. https://doi.org/10.1075/itl.20011.bar

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.

Callan, D. E., & Schweighofer, N. (2009). Neural correlates of the spacing effect in explicit verbal semantic encoding support the deficient-processing theory. *Human Brain Mapping*, *31*(4), 645-659. https://doi.org/10.1002/hbm.20894

De Jong, N. (2005). Can second language grammar be learned through listening?: An experimental study. *Studies in Second Language Acquisition*, *27*(02). https://doi.org/10.1017/S0272263105050114

DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*(2), 195–221. https://doi.org/10.1017/S0272263197002040

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

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Haidar, S. (2019). The role of English in developing countries. *English Today*, *35*(3), 42–48. https://doi.org/10.1017/S0266078418000469

Hamid, M. O. (2016). The linguistic market for English in Bangladesh. *Current Issues in Language Planning*, *17*(1), 36–55. https://doi.org/10.1080/14664208.2016.1105909

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

K, V. (2016). Learning by rote prevalent in top schools too. Retrieved from https://www.thehindu.com/opinion/op-ed/learning-by-roteprevalent-in-top-schools-too/article2707183.ece

Kachinske, I. (2021). Skill Acquisition Theory and the role of rule and example learning. *Journal of Contemporary Philology*, Ss Cyril and

Methodius University, B Koneski Faculty of Philology, *4*(2), 25–41. https://doi.org/10.37834/JCP2142025k

Kim, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science, 14*(1), 22–37. https://doi.org/10.1080/1464536X.2011.573008

Kim, K. H. (2005). Learning from each other: Creativity in East Asian and American education. *Creativity Research Journal*, *17*(4), 337–347. https://doi.org/10.1207/s15326934crj1704_5

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196. https://doi.org/10.1017/S0272263116000085

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

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

Maeda, M. (2021). Exam cheating among Cambodian students: When, how, and why it happens. *Compare: A Journal of Comparative and International Education*, *51*(3), 337–355. https://doi.org/10.1080/03057925.2019.1613344

McLean, S., Hogg, N., & Rush, T. W. (2013). Vocabulary learning through an online computerized flashcard site. *The JALT CALL Journal*, *9*(1), 79–98. https://doi.org/10.29140/JALTCALL.V9N1.149

MoEYS. (2022). K-12 student learning assessment framework. Retrieved from http://www.moeys.gov.kh/index.php/en/ig/3164.html#.YysFonZByUl

Moore, S. H., & Bounchan, S. (2010). English in Cambodia: Changes and challenges. *World Englishes*, *29*(1), 114–126. https://doi.org/10.1111/j.1467-971X.2009.01628.x

Muñoz, C., & Cadierno, T. (2021). How do differences in exposure affect English language learning? A comparison of teenagers in two learning environments. *Studies in Second Language Learning and Teaching*, *11*(2), 185–21 https://doi.org/10.14746/ssllt.2021.11.2

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

Pearson, J. (2022). Capacity development in Cambodia: The challenge of changing an educational culture. In V. McNamara & M. Hayden (Eds.), *Education in Cambodia. Education in the Asia-Pacific Region: Issues, Concerns and Prospects*, vol 64. (pp. 175–194). Singapore: Springer. https://doi.org/10.1007/978-981-16-8213-1_10

Rodgers, D. M. (2011). The automatization of verbal morphology in instructed second language acquisition. *IRAL - International Review of Applied Linguistics in Language Teaching*, *49*(4). https://doi.org/10.1515/iral.2011.016

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

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*(3), 369–385. https://doi.org/10.1017/S0142716400010845

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics*, *19*(1), 53–67. https://doi.org/10.1017/S0142716400010572

Serfaty, J., & Serrano, R. (2020). Examining the potential of digital flashcards to facilitate independent grammar learning. *System*, *94*, 10234 https://doi.org/10.1016/j.system.2020.102342

Serfaty, J., & Serrano, R. (2022). Lag effects in grammar learning: A desirable difficulties perspective. *Applied Psycholinguistics*, *43*(3), 513–550. https://doi.org/10.1017/S0142716421000631

Serfaty, J., & Serrano, R. (in review). The optimal scheduling of Quizlet sessions for L2 vocabulary learning.

Serfaty, J., & Serrano, R. (resubmitted). Practice makes perfect, but how much is necessary? The role of relearning in L2 grammar acquisition.

Shepperd, L. (2022). Including underrepresented language learners in SLA research: A case study and considerations for internet-based methods. *Research Methods in Applied Linguistics*, *1*(3), 100031. https://doi.org/10.1016/j.rmal.202100031

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. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*, *71*(2), 285–325. https://doi.org/10.1111/lang.12433

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

Suzuki, Y., Nakata, T., & DeKeyser, R. (2019). The Desirable Difficulty Framework as a Theoretical Foundation for Optimizing and Researching Second Language Practice. *The Modern Language Journal*, *103*(3), 713–720. https://doi.org/10.1111/modl.12585

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

Visal, S., Oeurn, C. C., & Chhinh, S. (2022). The teaching profession in Cambodia: Progress to date and ongoing needs. *Indian Journal of Science and Technology*, *14*(12), 115–13 https://doi.org/10.1007/978-981-16-8213-1_7

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–5 https://doi.org/10.1017/S0272263105050023

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*(3), 360–376. https://doi.org/10.1177/0033688209343854