LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

Check for updates

EMPIRICAL STUDY

Practice Makes Perfect, but How Much Is Necessary? The Role of Relearning in Second Language Grammar Acquisition

Jonathan Serfaty and Raquel Serrano
University of Barcelona

Abstract: This study investigated how much practice is necessary for learners to attain durable second language (L2) grammar knowledge. Using digital flashcards, 119 participants practiced translating 12 sentences into an artificial language, followed by feedback, until they had typed all sentences correctly. Participants repeated this activity in one, two, three, or four relearning sessions on consecutive days. After a 14-day delay, all groups scored highly on a receptive test. However, scores on a productive test were substantially higher for groups with three or four relearning sessions. Accuracy tended to peak on the 3rd day of training. An analysis by individual training performance re-

---

CRediT author statement – **Jonathan Serfaty**: conceptualization; methodology (lead); data curation; formal analysis (lead); writing – original draft preparation; writing – review & editing. **Raquel Serrano**: conceptualization; methodology (supporting); formal analysis (supporting); writing – original draft preparation; writing – review & editing; supervision.

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

The project was funded by a *Language Learning* dissertation grant. The results reported here are based on the first author's doctoral dissertation submitted to Universitat de Barcelona in 2022. We would like to thank the anonymous *Language Learning* reviewers for their insightful comments on the manuscript, and Editors Judit Kormos and Pavel Trofimovich for their support throughout the review, revision, and production. We would also like to acknowledge the Spanish Ministry of Science and Innovation, Project PID2019-110536GB-I00 for additional support.

Correspondence concerning this article should be addressed to Jonathan Serfaty, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585 Barcelona, Barcelona, Catalonia 08007, Spain. Email: jonny.serfaty@gmail.com

The handling editor for this manuscript was Judit Kormos.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Language Learning* published by Wiley Periodicals LLC on behalf of Language Learning Research Club, University of Michigan.
DOI: 10.1111/lang.12585

vealed that participants attained durable productive knowledge if they completed two sessions without errors, regardless of how many sessions they had performed in total. The findings provide a timeframe for processes described in skill retention theory (Kim et al., 2013) and suggest a performance benchmark to indicate when learners have gained procedural L2 grammar knowledge.

Keywords digital flashcards; grammar practice; relearning; skill acquisition theory; skill retention 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 (https://quizlet.com) have been widely used, but little 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 have been found to help to bridge this gap (DeKeyser, 2010; Lyster, 2004; 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.

Although these studies differed in several ways, one salient difference was the quantity of practice. Flashcards involve criterion learning, meaning that a session ends only after every item has been answered correctly. A repeated session on a different day can be classed as a relearning session (RS; Rawson & Dunlosky, 2011) since the material has already been learned. The high scores from Serfaty and Serrano's (2020) study came after three RSs per structure, whereas the low scores from Serfaty and Serrano's (2022) study came from only one RS.

In several studies, cognitive psychologists compared learning vocabulary pairs (Bahrick, 1979; Pyc & Rawson, 2007, 2009, 2011; Rawson et al., 2018; Vaughn, Dunlosky, et al., 2016; Vaughn, Rawson, et al., 2013) and key-term definitions (Rawson & Dunlosky, 2011, 2013) to criterion after different amounts of retrieval practice. 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, especially for paired associates (Rawson et al., 2018). These studies involved learning idiosyncratic chunks of information, known as declarative knowledge, using posttest items

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

identical to the training, whereas 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, 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 retention properties (Li & DeKeyser, 2019). Moreover, the skill of applying L2 grammar rules is tested using novel items not found in training. Due to these variables, conclusions from the extant literature on the successive relearning of declarative knowledge cannot be directly applied to L2 grammar practice. For instance, for the partially procedural skill of solving mathematical problems, the advantage of distributing practice over multiple RSs compared with overlearning in one session was much smaller than the advantage of doing so for vocabulary learning (Rawson et al., 2020).

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 stage and then gradually curves off until no further improvement is observable. According to skill retention theory (Kim et al., 2013), this plateau of performance metrics signifies that knowledge has been fully proceduralized, and the resulting procedural knowledge is less prone to decay over time than is declarative knowledge. Consequently, for L2 grammar practice, there should be a point at which durable knowledge is attained, and this should be predictable from training measures. However, to the best of our knowledge, no study has yet investigated how much practice is necessary to achieve durable knowledge or whether training performance indeed predicts the durability of knowledge. This, therefore, constitutes an important gap in the literature that needs to be filled not only for theoretical but also for pedagogical purposes. Determining how much L2 grammar practice is enough would provide a useful guide for teachers and learners in allocating study time efficiently.

In our study, we compared groups that performed in 1, 2, 3, or 4 RSs on consecutive days after an initial training session (TS). First, we assessed to what extent these groups retained their productive and receptive knowledge after 14-days without exposure to the target language. Second, we analyzed training performance to find out how many sessions were needed before the participants' accuracy no longer improved. Finally, we explored whether individuals' accuracy during training could predict their accuracy in posttests,

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

which would be useful for learners who wish to know when they have attained durable knowledge. In investigating these three questions, we aimed to deepen our understanding of the learning processes for L2 grammar within the framework of skill acquisition theory, as well as to provide pedagogically relevant recommendations regarding the optimal allocation of time for L2 grammar practice.

Background Literature

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

Flashcards have also been used for L2 grammar learning (Serfaty & Serrano, 2020, 2022), targeting learners' ability to produce a full sentence in the L2 accurately without any parts of the sentence provided for them. Grammar flashcards present learners with a meaning cue (e.g., a sentence in the L1) and require them to type a full sentence in the L2. Each flashcard exemplifies a grammatical pattern. For example, by producing *I eat rice*, *You eat rice*, *He eats rice*, and *She eats rice*, learners can induce that the *–s* is added for *He* and *She*, but not for *I* and *You*. A more complex example could be a set of sentences using the conditional structure for unreal situations, such as *If I had seen him, I would have told him*, to learn the pattern "If + subject + had + past participle + object, subject + would + have + past participle + object." This approach to grammar learning differs from purely 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. Feedback from an incorrect response provides both positive and negative evidence of target forms, allowing learners to compare their attempt to the target response and alter their hypothesis (Zalbidea, 2021). Incorrectly typed responses are then repeated until learners can produce each sentence as

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

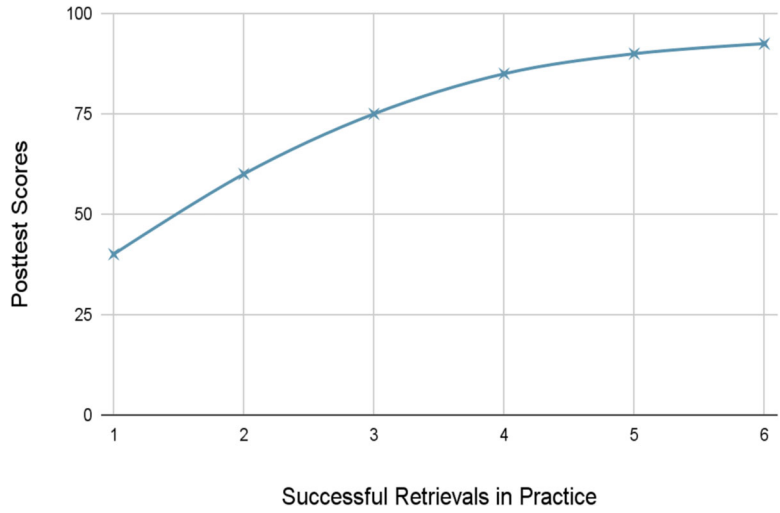


Figure 1 Illustration of subadditive effects from overlearning.

required. Repeatedly practicing the same target structure in different iterations facilitates the proceduralization of the target forms (Lyster & Sato, 2013).

Quantity of Practice From Paired-Associate Research

Cognitive psychologists have compared different quantities of paired-associate retrieval practice for translations and definitions of target words. One focus has been on overlearning, defined as the continued practice of already-learned items within the same session (e.g., Rohrer et al., 2005). It has been shown that for every additional correct retrieval required during a practice session, posttest scores become higher with a subadditive effect (e.g., Pyc & Rawson, 2009; Vaughn et al., 2013). That is, the gains from additional practice progressively diminish toward asymptote (see Figure 1).

Another research focus has been on successive relearning, with more RSs facilitating higher gains (Bahrick, 1979; Bahrick et al., 1993). In studies using incremental increases in the number of sessions (Rawson & Dunlosky, 2011, 2013; Vaughn et al., 2016), the same subadditive effect from overlearning has been found for relearning. However, relearning has seemed to be far more effective than overlearning (Rawson et al., 2018). For instance, 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 RSs (1–4 sessions). The mean score for items learned 4 times within a

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

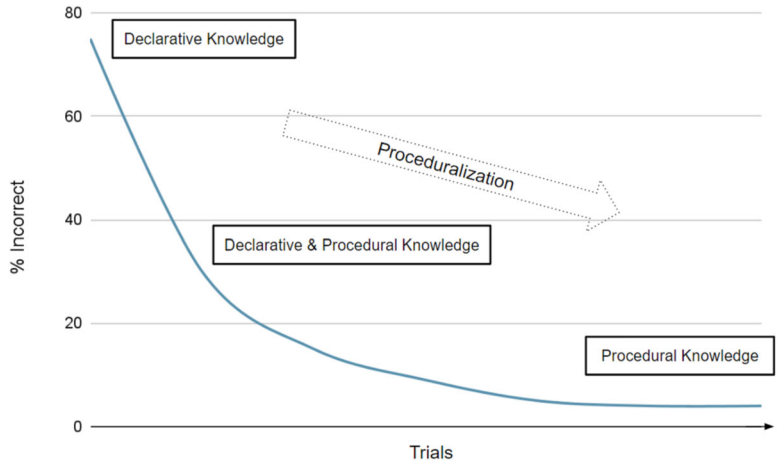


Figure 2 Theoretical improvement in accuracy and type of knowledge attained according to skill retention theory.

single session was 28%, but when the same number of successful retrievals were distributed over 4 weeks, the mean score was 74%.

Skill Acquisition Theory

The studies mentioned in the previous section only examined declarative knowledge. Procedural knowledge, however, 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 is prone to decay (Pavlik & Anderson, 2005), whereas procedural knowledge develops more slowly and is much more durable (Kim et al., 2013; Li & DeKeyser, 2017; Ullman, 2020).

Skill acquisition theory, adapted to instructed L2 learning by DeKeyser (2017, 2020), assumes that learners begin with declarative knowledge that is then proceduralized and automatized through practice. This process is characterized by the gradual reduction in errors and response times (RTs) in learners' retrieving knowledge or performing a skill. Kim et al. (2013)'s skill retention theory describes how the trajectory of progress in accuracy and RTs for performing a skill may correspond to three stages of learning from skill acquisition theory, stages that they referred to as (a) declarative, (b) declarative + procedural, and (c) procedural (see Figure 2). This curve tends to be found for

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

group means across time but may not necessarily capture individual trajectories (Ferman et al., 2009).

At Stage 1, knowledge is declarative, analogous to learners being able to explain a grammar rule. Through application, procedural knowledge begins to develop relatively quickly, moving into Stage 2. This transition is reflected in a steep reduction in errors and RTs. During Stage 2, learners still rely on their initial declarative knowledge. Learners' training performance improves more gradually as they strengthen their procedural knowledge and rely on declarative knowledge to a lesser extent. Finally, learners reach Stage 3, evidenced by a leveling off in training performance. In Stage 3, learners no longer need declarative knowledge to perform the skill, and they can instead rely on procedural knowledge which is more easily accessible and more durable. In the case of L2 grammar, learners would be able to produce a grammatical sentence even if they have forgotten the rules. Kim et al. (2013) suggested that relearning is beneficial during Stages 1 and 2, but not after Stage 3. Focusing on L2 grammar learning, DeKeyser (2017, p. 96) 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)." Applying this to grammar flashcards, we predicted that flashcard practice should be useful until reaching Stage 3, after which point, no further improvement in accuracy would be found. To the best of our knowledge, no previous study has empirically tested how different amounts of relearning affect the acquisition and retention of L2 grammar.

The Present Study

The present study compared groups practicing an artificial language with flashcards on 2, 3, 4, or 5 consecutive days. The language included four vocabulary items as well as subject and object pronouns to be inflected according to rules. In each session, the participants translated 12 English sentences into the target language and received feedback, with items repeating in a cycle until typed perfectly. All sentences were answered correctly once during every session. Our first aim pertained to the determination of the number of RSs necessary to attain durable knowledge. According to skill retention theory, procedural knowledge is more durable than declarative knowledge. Therefore, knowledge should become durable when fully proceduralized (Stage 3). Moreover, further learning within Stage 3 should provide no benefit on accuracy measures. On the basis of previous studies using grammar flashcards (Serfaty & Serrano, 2020, 2022), it seemed likely to us that the number of RSs required for

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

participants to achieve durable knowledge would be somewhere between one and four. We therefore hypothesized that knowledge would enter Stage 3 and become durable within one to four RSs, with no further improvement from relearning past this point.

Another aim of our study was to establish how many RSs are needed before accuracy no longer improves during training, which according to skill retention theory is evidence of fully proceduralized 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 the participants' scores as a function of their final attainment during training, rather than the number of RSs that they had performed, in order to explore whether it was possible to predict, during training, when an individual learner has attained durable knowledge. Analyzing groups on the basis of the number of RSs that learners perform can provide a general idea as to the effects of practice, but some structures and some learners may require different amounts of practice (Ferman et al., 2009). We hypothesized that the participants would achieve the highest posttest scores if their training performance plateaued, which Kim et al. (2013) suggested is a sign of fully proceduralized knowledge. Our research questions were:

1. How many RSs are needed to achieve durable L2 grammar knowledge as shown by a 14-day posttest?
2. After how many RSs does accuracy no longer improve during training?
3. Can individuals' accuracy during training predict when they have acquired durable L2 grammar knowledge?

Method

Pilot

We conducted a pilot study to ensure that the artificial language would not be too easy or too difficult for our participants. Using Quizlet, 30 volunteers from eight countries aged 19–63 years, 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. The participants practiced the set for either 2, 3, or 4 days. They each then performed cued-recall tests 1, 5, and 28 days after training. The results revealed highly variable abilities for learning the rules of the language, with times on task during the first session ranging from 5 min to

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

over 1 hour, corresponding somewhat to age. Subsequent posttest scores ranging from 0% to 100% were heavily dependent on first-session performance. As a result, we decided to include a guided rule-learning phase before the main task and to recruit only participants aged 18–30 years. Further, scores of 100% at the 1-day posttest consistently predicted identical scores at the 5-day and 30-day posttests, so we decided to use a single posttest with a long enough interval for forgetting to occur. Finally, since Quizlet could not track trials and times, we chose to use Gorilla (https://gorilla.sc) instead. This platform allows researchers to design online experiments and provides detailed data, such as responses and reaction times, on every trial that participants perform.

Participants

We recruited participants via Prolific (https://www.prolific.com) and paid them £10/hour. We set a filter to recruit only participants with fluency in English—because instructions and cues were in English—aged 18–30 years, with an education level of high school diploma or higher. The recruitment description heavily emphasized that this study would be difficult and 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 participants unmotivated to engage in language learning or for those with time constraints.

The participants came from 32 different countries with different L1s. We later excluded three participants on the basis of their performance (see Data Preparation section below). The final number of participants was 119, with 29 in the group with one RS and 30 in the remaining three groups. Participant attributes, including age, gender, L1, and language learning background, had no discernible impact on results (see Appendix S1 in the Supporting Information online for details).

Target Language

The target language was NamiChip, a miniature language that the first author developed for this study. We chose an artificial language in order to control for previous knowledge or practice outside of the treatment and designed the language to be easily learnable. The target language included eight pronouns, one noun (*dog*), two uninflected verbs (*have* and *like*) and one number (*two*). The target rules related primarily to the pronouns (formed by adding the correct suffix to the letter "K"), case marking (add an "s" prefix if the noun or pronoun is a subject) and number marking (double the last letter for plurals), following an SVO structure. Since all the participants were proficient in English, which

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

codes pronouns by person, case, and number, we expected these distinctions to be conceptually simple. Table 1 presents all the words in NamiChip. For each sentence, the participants needed to apply all the different target rules by choosing the proper subject/object and singular/plural marking, as well as the right form of the pronoun. We combined the words into 24 different sentences, 12 for training and 12 for posttests (see Appendix S2 in the Supporting Information online). We used novel sentences for testing to ensure that the participants had learned the rules governing the sentences rather than memorizing whole chunks.

Training Procedure

We developed and conducted the study through the online platform Gorilla. The 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 the participants could not be monitored, we hoped that their agreement to these conditions would encourage them to follow the experiment faithfully. Any nonadherence to the experiment would have been equally as likely in all four conditions.

In the guided-learning phase of the TS, the 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 the 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 the participants' 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 in the Supporting Information online displays all screens from this phase.

In the second phase of the TS, the participants saw an English sentence and were asked to provide the translation in NamiChip. If their translation was

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

incorrect, the participants saw the target sentence alongside their own response, and the flashcard was sent to the back of the cycle. If their translation was correct, they saw positive feedback, and the flashcard was removed (see Figure 3). This phase simulated commercial flashcard software, such as Quizlet, within Gorilla. Subsequent RSs simply repeated this second phase using the same 12 sentences each time.

Training Performance

The goal of the TS was for the participants to acquire and retain accurate production of the target language. In the learning phase, we operationalized accuracy as the number of trials per session. Since all sessions had a minimum of 12 trials, corresponding to 12 sentences typed perfectly, any trials above this number indicated a response with at least one error. A reduction in trials was therefore an improvement in a participant's ability to produce fully accurate sentences.

As well as recording 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). Although we initially considered this, we identified some confounding variables in the analysis of RTs in our study. For typed responses, especially of multiword 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 the participants' improved ability to type new words quickly. These variables were likely to overshadow changes in processing speed, which is measured on the scale of milliseconds. Additionally, we imposed no time pressure in all learning sessions. We encouraged the participants to finish quickly but to check their responses before submitting them, 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 presented 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 our study was accuracy development, we analyzed only the number of trials per session as our measure of performance.

Tests

We used two unannounced posttests to measure the participants' retention of knowledge 14 days after training. The first test took the same format as

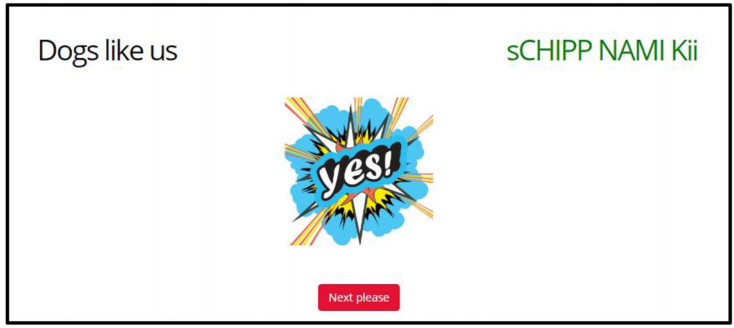
14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Translation screen with cue



Negative feedback screen



Positive feedback screen

Figure 3 Examples of screens that the participants encountered while learning NamiChip.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the training but without feedback. The 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 the participants had to translate them into English. Although both tests involved actively producing language (i.e., they tested participants' recall and not just their recognition of the target forms), the first test required productive knowledge (translating from the L1 to the L2) whereas the second test required only receptive knowledge of the target language (translating from the L2 to the L1). We have therefore referred to these tests as productive and receptive tests, respectively. Vocabulary research has consistently found receptive recall ability to be higher than productive recall ability (Laufer & Goldstein, 2004; Nakata, 2016; Nakata & Webb, 2016; Webb, 2009), and so, although productive scores were the primary interest of our study, we expected the subsequent receptive test to capture weaker knowledge traces that we might not observe in the productive test. To the best of our 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. We therefore included this comparison as an initial exploration of this issue despite its not being a main focus of the study. Cronbach's alpha was .906 for the productive test and .823 for the receptive test.

Upon completing both tests, the 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 how they felt about the experiment. This allowed us to gain some insight into the participants' experiences. We also included a space for the participants to specify any rules of the language that they remembered. We rationalized that even if the participants' knowledge had been proceduralized during training, posttest responses after a 14-day gap might involve some retrieval of declarative knowledge or that declarative knowledge could be reverse-engineered upon reflection. We included this question in order to facilitate the interpretation of our data.

Experimental Design

The TS and RSs took place on consecutive days. We chose a short interval of 1 day 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 that Prolific would allow before automatically paying participants. High retention from digital flashcard training after 14 days has previously been

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

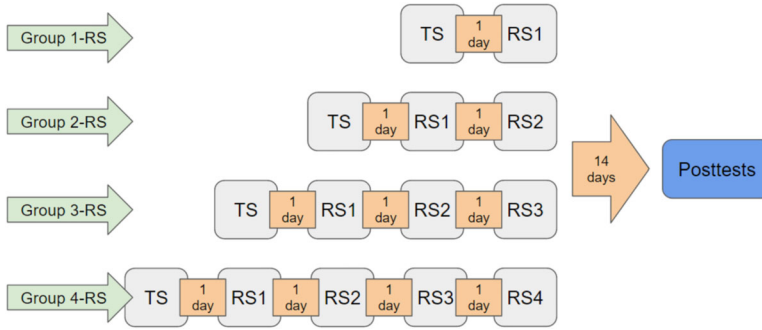


Figure 4 Experimental design. TS = training session. RS = relearning session.

shown to predict similar retention after 18 weeks (Serfaty & Serrano, 2020). The four conditions differed only in the number of RSs performed (i.e., one, two, three, or four sessions). Figure 4 shows the experimental design.

Data Preparation

Gorilla automatically scored the posttests. We manually checked incorrect responses for answers that were correct despite not exactly matching the target response. For example, we marked responses correct if there was an error in the number of spaces or if a mistake could be identified as a typographical error ($k = 10$) according to the following criteria: (a) The word did not match anything else in NamiChip, (b) every other instance of that word was typed correctly, and (c) the letter in the typographical error was adjacent to the target letter on the keyboard. Of the 430 responses that Gorilla scored as incorrect in the productive test, 51 errors were related purely to lexis, and we ignored them in order to focus on grammatical errors. Sentences were therefore marked as correct if pronoun constructions, subject/object case, pluralization, and word order were perfect, for example, where a participant wrote "chap" instead of "chip" in every instance of *dog*, but applied the grammar rules perfectly to this word. All remaining incorrect responses were related to the grammar rules.

We excluded two participants for their not having attempted to translate the cues. We removed one additional participant who had required considerably more trials in training (TS = 72, RS1 = 74) than had the participant with the next highest number of trials (TS = 53, RS1 = 36) and who had no correct answers in the posttest.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Finally, we checked the comments from the participants. We observed that the majority of the participants were able to list the rules of the language as visualized in Appendix S4 in the Supporting Information online. These observations supported the validity of the data because the participants would not be reasonably expected to remember these rules without having participated faithfully in the experiment. Moreover, the comments regarding their feelings revealed some emotional investment from both high and low scoring participants. These comments can be found online in Appendix S5 in the Supporting Information online. We checked the 5-point Likert survey on perceptions for any conspicuous differences between groups that could have affected the results. All the groups rated enjoyment and effectiveness very positively at 4–5, but ease was rated slightly lower at 3–4. Descriptive results of this survey are presented in Appendix S6 in the Supporting Information online.

Data Analysis

We performed all statistical analyses using IBM SPSS Statistics (Version 27). To analyze posttest scores, we performed generalized linear mixed models with binary outcomes in order to locate statistical differences between groups. This type of model does not require any additional assumptions to be tested for these types of variables. The fixed effects were group (number of RSs performed, 1 to 4 sessions), test (productive or receptive), and the Group × Test interaction. We included participant and item as random intercepts. To determine the effect size, we used the odds ratio, which indicates the added likelihood of a correct score from the condition with the higher mean. For example, if the odds ratio is 2, a correct response is twice as likely.

We analyzed the number of error trials through a growth curve model using a Poisson distribution. The data met assumptions for normality of residuals, homogeneity of variances, and overdispersion. We used random intercepts for participants and random slopes for progress over time. We set session as a numerical predictor. Initially, we also included group to check for unexpected differences between groups, but we removed group when we found no effect for it. The effect size was the incidence rate ratio, which is the ratio of change in the average number of trials between sessions. For example, if the incidence rate ratio was 0.5, then the mean number of error trials halved in the later session. All significance tests were two-tailed, and we interpreted a p value of .05 or lower as statistically significant, making Bonferroni corrections within models. The datasets can be accessed through the Open Science Framework (https://osf.io/v7pz3). Appendix S7 in the Supporting Information online

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

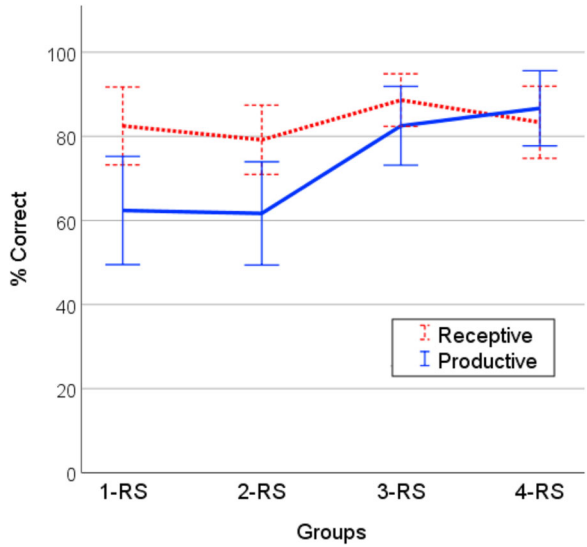


Figure 5 Mean productive and receptive scores for the four relearning session (RS) groups. Error bars = 95% confidence interval.

contains every model's syntax as well as all associated means, effects, and 95% confidence intervals.

Results

Research Question 1: How Many Learning Sessions Are Needed to Achieve Durable L2 Grammar Knowledge?

Figure 5 displays the mean scores by group for productive and receptive tests. The groups are labeled according to the number of RSs that they performed. For the productive test, mean scores for Groups 1-RS and 2-RS were almost identical at 62.36% (*SD* = 33.82, 95% CI = [50.05, 74.67]) and 61.67% (*SD* = 32.87, 95% CI = [49.90, 73.43]), but mean scores for Groups 3-RS and 4-RS were substantially higher at 81.39% (*SD* = 25.77, 95% CI = [73.53, 91.48]) and 82.50% (*SD* = 25.08, 95% CI = [78.10, 95.23]). Mean receptive scores did not show this variation between groups. From Group 1-RS to Group 4-RS, mean scores were 82.47% (*SD* = 24.33, 95% CI = [73.62, 91.33]), 79.17% (*SD* = 22.08, 95% CI = [71.27, 87.07]), 86.67% (*SD* = 23.93, 95% CI = [82.62, 94.60]), and 83.33% (*SD* = 22.95, 95% CI = [75.12, 91.55]) respectively.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2 Groups means and pairwise contrasts: Productive and receptive scores

			Group 2-RS		Group 3-RS		Group 4-RS	
Group	*EMM*	*SE*	*p*	*OR*	*p*	*OR*	*p*	*OR*
Productive scores								
1-RS	.624	.055	.928	1.030	.015	2.846	.003	3.924
2-RS	.617	.054			.015	2.931	.003	4.041
3-RS	.825	.042					.924	1.379
4-RS	.867	.041						
Receptive scores								
1-RS	.825	.043	1.000	1.238	1.000	1.654	1.000	1.063
2-RS	.792	.045			.631	2.047	1.000	1.316
3-RS	.886	.035					1.000	1.556
4-RS	.833	.041						

Note. RS = relearning session; *EMM* = estimated marginal means; *OR* = odds ratio.

The statistical model produced significant main effects for group, $F(3, 2848) = 2.90$, $p = .034$, test, $F(1, 2848) = 39.85$, $p < .001$, and the Group $\times$ Test interaction, $F(3, 2848) = 11.47$, $p < .001$. Table 2 shows the estimated marginal means and group pairwise contrasts. For productive scores, there were no statistically significant differences between the first two groups (1-RS vs. 2-RS) or between the second two groups (3-RS vs. 4-RS). It appeared that RS2 did not improve the odds of a correct posttest response, but those odds were around three times higher from RS3 and four times higher from RS4. For receptive scores, there were no statistically significant differences between groups, suggesting that RS1 was sufficient.

Research Question 2: After How Many Relearning Sessions Does Accuracy No Longer Improve During Training?

In order to answer Research Question 2, we collapsed the experimental groups and analyzed the data by session. 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 6 shows the learning curves of each group. Group 3-RS showed more variance than the other groups during the TS, but by RS1 all the groups exhibited similar performance. The learning curve appeared steeper from the TS to RS1 and then gradually flattened, in line with previous research.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3 Descriptive statistics for number of trials by session

Session	*n*	*M*	*SD*	Min	Max
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

Note. TS = training session; RS = relearning session.

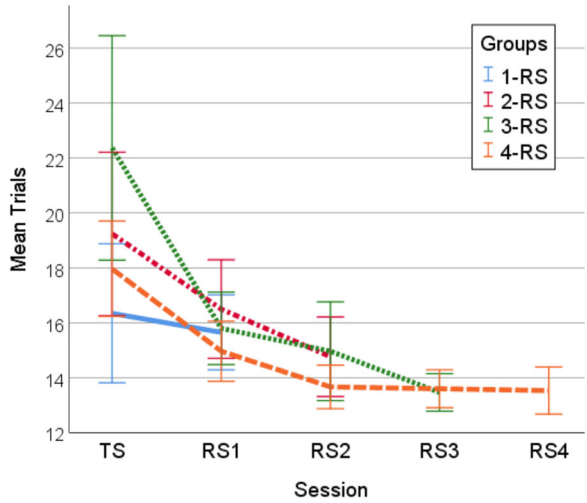


Figure 6 Mean number of trials by training session (TS) and relearning session (RS) separated by group. Error bars = 95% confidence interval.

The growth curve model (see 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 error trials. There was no significant difference in the number of error trials from RS3 to RS4. In sum, the degree of improvement roughly halved with each new session, with no further improvement at the final session.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4 Pairwise contrasts of error trials per session

Session	*EMM*	*SE*	Contrast	Contrast estimate	*IRR*	*p*
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	.009
			RS4	−0.679	0.638	.072
RS3	1.242	0.187	RS4	−0.044	0.965	.864
RS4	1.198	0.268	–	–	–	–

Note. EMM = estimated marginal mean; *IRR* = incidence rate ratio; TS = training session; RS = relearning session.

Research Question 3: Can Individuals' Accuracy During Training Predict When They Have Acquired Durable L2 Grammar Knowledge?

In order to answer Research Question 3, we first looked at the training performance of groups that did, on average, achieve high posttest scores—Groups 3-RS and 4-RS—and observed that these groups performed a mean of 14 trials in their final sessions. The mean number of trials had already reached this level by RS2 and remained above 13.5 for RS3 and RS4. Unexpectedly, this plateau was not at the true minimum of 12 trials. Upon inspecting the data, we 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 who had demonstrated perfect accuracy in a previous session later made one or two errors. We reasoned that typing mistakes could be responsible for this discrepancy. When inspecting the responses from these participants, we identified examples of accidental typographical errors. For example, one participant wrote "mee" instead of "kee" (the "m" and "k" keys are adjacent) in only one item. DeKeyser (1997) also observed that some of his participants made one or two accidental typing errors during their vocabulary test. If the 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, we have referred to any session completed within 14 trials as a minimum-trials session (MTS).

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 5 Minimum-trials session (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

Note. RS = relearning session.

Table 6 Productive and receptive scores (%) by minimum-trials session (MTS) groups

	Productive scores			Receptive scores		
MTS	*M*	*SD*	95% CI	*M*	*SD*	95% CI
0	45.83	30.10	[33.79, 57.87]	70.83	25.77	[60.52, 81.14]
1	62.12	36.71	[46.78, 77.46]	78.79	23.95	[68.78, 88.80]
2	83.33	21.61	[74.50, 92.17]	89.86	16.08	[83.28, 96.43]
3	89.29	16.48	[80.65, 97.92]	92.26	6.91	[88.64, 95.88]
4	94.64	13.72	[87.49, 101.83]	91.07	18.33	[81.47, 100.68]
5	93.75	12.50	[81.50, 106.00]	89.58	10.49	[79.31, 99.86]

Note. CI = confidence interval.

Since posttest scores peaked after RS3, but training accuracy peaked at RS2, we postulated that one MTS was not enough and that learners must perform at least two MTSs in order to acquire durable knowledge. To test this, we recoded the participants by the number of MTSs that they had achieved in succession. Some participants ($n = 18$) completed one MTS but then needed more than 14 trials in a subsequent session. Since these participants could not be classified clearly, we excluded them from this analysis. We grouped the remaining participants ($n = 101$) as follows: 0-MTS ($n = 24$); 1-MTS ($n = 22$); 2-MTS ($n = 23$); 3-MTS ($n = 14$); 4-MTS ($n = 14$); 5-MTS ($n = 4$). Table 5 shows the breakdown of MTS groups according to their membership in the original RS groups. Table 6 and Figure 7 display the descriptive statistics of posttest scores for MTS groups.

Since Group 5-MTS had only four members, we collapsed it into Group 4-MTS for the statistical analysis ($n = 18$); productive: $M = 94.44$ ($SD = 13.10$, 95% CI = [83.84, 105.05]); receptive: $M = 90.74$ ($SD = 16.64$, 95% CI = [77.32, 104.16]). There were significant main effects for MTS, $F(1, 2414) =$

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

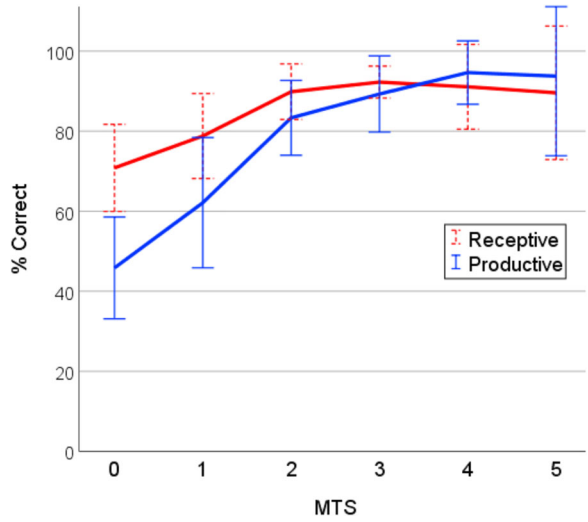


Figure 7 Mean productive and receptive scores by minimum trail session (MTS) groups. Error bars = 95% confidence interval.

8.173, $p < .001$, test, $F(4, 2414) = 14.452$, $p < .001$, and the MTS × Test interaction, $F(4, 2414) = 4.908$, $p = .001$. For productive scores, there was no statistically significant difference between Groups 0-MTS and 1-MTS, but these groups differed significantly from all the other groups. These effect sizes were substantially higher than for the analysis by number of RSs. For instance, the *OR* for the difference between Groups 0-MTS and 4-MTS was 20.09, whereas the highest *OR* in the original analysis was 4.04. No other differences in productive scores were statistically significant. For receptive scores, differences between Group 0-MTS and other groups approached significance. Table 7 displays all pairwise contrasts.

Discussion

Our study 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 2, 3, 4, or 5 consecutive days, with all items recycled until answered correctly within each session. The 1st day was the TS and subsequent days were RSs. The participants performed translation tests on novel sentences 14 days after the treatment, first from the English cue to the target language (productive knowledge) and then from the target language to English (receptive knowledge).

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 7 Pairwise contrasts by minimum-trials session (MTS) groups: Productive and receptive scores

			1-MTS		2-MTS		3-MTS		4-MTS	
MTS	*EMM*	*SE*	*p*	*OR*	*p*	*OR*	*p*	*OR*	*p*	*OR*
Productive scores										
0	.458	.056	.184	1.938	< .001	5.909	< .001	9.848	< .001	20.091
1	.621	.057			.023	3.049	.014	5.081	.001	10.366
2	.833	.043					.673	1.667	.184	3.400
3	.893	.046							.673	2.040
4	.944	.030								
Receptive scores										
0	.708	.051	1.000	1.529	.045	3.647	.069	4.910	.059	4.035
1	.788	.048			.428	2.385	.428	3.210	.428	2.638
2	.899	.035					1.000	1.346	1.000	1.106
3	.923	.039							1.000	1.217
4	.907	.038								

Note. EMM = estimated marginal mean; *OR* = odds ratio.

Number of Learning Sessions and Durable L2 Grammar Knowledge

To answer Research Question 1, we compared the participants' posttest scores after different numbers of RSs. After one or two RSs, productive scores were very similar (62%), but after three or four RSs, scores were considerably higher (81–83%) with no significant differences within the first or second pair of groups. These means represent a wide range of scores for Groups 1-RS and 2-RS, whereas Groups 3-RS and 4-RS made very few errors. Therefore, RS3 substantially raised the odds of participants learning and retaining the target language, with no statistically significant improvement from RS4.

We found different results for receptive scores; they were similarly high for all groups, meaning that receptive knowledge was durable after only one RS. This is similar to Pili-Moss et al.'s (2020) study in which 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 has also been 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.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The findings for productive knowledge support our speculation that the different results from previous grammar flashcard studies were related to the number of sessions. Participants in Serfaty and Serrano's (2020) study practiced target structures on 3 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. At a 14-day posttest, mean scores were still over 80%. Likewise, the participants in the present experiment also scored above 80% after three RSs at a 14-day posttest. In comparison, participants in Serfaty and Serrano's (2022) study practiced target structures with only one RS, resulting in mean posttest scores of 50% at a 7-day posttest. Similarly, in the present experiment, those who studied with one RS scored around 62% 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.

Number of Relearning Sessions and Training Performance Plateau

Research Question 2 aimed to quantify the rate of progression to Kim et al. (2013)'s Stage 3, theorized to occur when training performance plateaus. The number of trials dropped from the TS until RS2, suggesting an improvement in accuracy. From RS2 to RS3, the difference was statistically significant but by less than a single trial. Therefore, on average, two RSs were needed to reach the point at which accuracy no longer meaningfully improved. This overall pattern is comparable 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). All of these studies, including ours, differed in terms of training, target material, testing, and time on task, yet all 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 on the applicability of artificial language studies to authentic L2 learning. However, Sato and McDonough did not systematically prompt target forms, and the instructor provided no feedback during the tasks. Had these conditions been met, results might have been similar to the artificial language studies.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Training Accuracy Predicting Durable L2 Grammar Knowledge

After observing that high productive posttest scores came after RS3, one session after trials reached their minimum, we postulated that learners must complete two MTSs in order to gain durable knowledge. A new analysis revealed that high scores (83% for productive and 90% for receptive knowledge) were achieved if the participants had performed a minimum of two MTSs, regardless of how many sessions they had 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 indicated that the participants were 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, but some still could not perform an MTS by their fifth session. Similarly, Zerr et al. (2018) found that learners requiring fewer trials to reach criterion for memorizing word pairs remembered more in posttests. Individual differences clearly play a large role in learning outcomes, and any recommended amount of practice should be based on achievement rather than absolute time on task.

Theoretical Implications

We can interpret the results obtained in this study through skill retention theory (Kim et al., 2013). According to this framework, in Stage 1 learners have only declarative knowledge, in Stage 2, they develop early procedural knowledge through practice but are still dependent on declarative knowledge, and, in Stage 3, further practice results in procedural knowledge that learners can use independently of declarative knowledge. Whereas declarative knowledge is prone to decay, independent procedural knowledge is thought to be more durable. The results of our study indicate that many of the participants reached Kim et al.'s (2013) Stage 3 after RS3, and that reaching this stage in training did predict later retention. Importantly, no further gains were evident after the participants continued to learn past this stage, which supports claims that learning past Stage 3 is less useful (DeKeyser, 2017; Kim et al., 2013).

However, comments from the 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 skill retention theory 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 nontimed and noncommunicative task. On the other hand, all the participants scored highly in the

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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, the participants were reminded of previous knowledge and were therefore able to translate the sentences to English and recount the rules in the debriefing phase. Three participants wrote comments to this effect, which can be seen in Appendix S8 in the Supporting Information online.

Our findings have further implications for the literature on successive relearning. In particular, Rawson et al. (2020) examined the relearning of solving mathematical problems. In Rawson et al.'s study, they defined the criterion as a single correct answer to any problem that represented the target rule. This meant that learners ended practice after one correct response and were therefore not allowed to proceduralize their knowledge through continued correct practice. Rawson et al. reported that relearning potency, which is the advantage of distributing successful retrievals over multiple days, was less impressive for this type of knowledge compared to verbal learning. However, as our study has shown, procedural knowledge must be practiced correctly many times in order to gain durability. It could be that solving mathematical problems is similar to grammar learning and that correctly solving multiple iterations of the same target problem during each session would result in greater relearning potency.

Limitations and Future Directions

Future research could replicate our experiment with a longer delay before testing in order to test the durability of knowledge. If learners do fully proceduralize their knowledge after two MTSs, then this knowledge should still be accessible after a longer retention interval. Moreover, if learners failing to reach this threshold are still relying on declarative knowledge, which would be in the process of decay, scores would be expected to drop further at a longer retention interval. 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. In contrast, longer intersession intervals may promote better retention of the declarative knowledge of rules (Bird, 2010; Nakata et al., 2023; for review, see Rogers, 2023). If so, learners might still have full access to their declarative knowledge at the posttest and thus would not need to reach Stage 3 in order to score highly.

Another variable to investigate could be the mode of output. As we 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

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 fluency. Finally, the data imply that the participants were adhering well to instructions, 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.

Conclusion

In our study, we investigated the effects of relearning on grammar flashcard training. Testing the participants after 14 days, we found that productive knowledge was much higher after three RSs (four sessions in total). The participants who reached the highest levels of productive knowledge were those who had 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 grammar structures, which could be useful for preparing students for communicative language practice. After using digital flashcards to practice grammar structures, 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 without any errors on two different days, the teacher could assign the student 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 algorithm could be designed to automatically assign or reassign activities to students on the basis of their performance.

It remains to be seen whether the number of error-free sessions needed to acquire durable knowledge would be the same for different types of L2 structures, for L2 vocabulary, or even for other subject domains. Results may also differ between experimental conditions among adults, as in our study, and results from school students in authentic learning environments. It is however

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 retaining 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.

Final revised version accepted 10 March 2023

Open Research Badges



This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data are available at https://osf.io/v7pz3.

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

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

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

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

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*, *14*(1), 22–37. 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), 399–432. 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. *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.), *The Routledge handbook of vocabulary studies* (pp. 304–319). Routledge. https://doi.org/10.4324/9780429291586-20

Nakata, T., Suzuki, Y., & He, X. (2023). Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing? *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12553

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.

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

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Pili-Moss, D., Brill-Schuetz, K. A., Faretta-Stutenberg, M., & Morgan-Short, K. (2020). Contributions of declarative and procedural memory to accuracy and automatization during second language practice. *Bilingualism: Language and Cognition*, *23*(3), 639–651. https://doi.org/10.1017/S1366728919000543

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*(8), 1917–1927. https://doi.org/10.3758/BF03192925

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. (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., & Janes, J. L. (2020). All good things must come to an end: A potential boundary condition on the potency of successive relearning. *Educational Psychology Review*, *32*(3), 851–871. https://doi.org/10.1007/s10648-020-09528-y

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. (2023). Spacing effects in task repetition research. *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12526

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

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

Serfaty, J., & Serrano, R. (2020). Examining the potential of digital flashcards to facilitate independent grammar learning. *System*, *94*, Article 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

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

Zerr, C. L., Berg, J. J., Nelson, S. M., Fishell, A. K., Savalia, N. K., & McDermott, K. B. (2018). Learning efficiency: Identifying individual differences in learning rate and retention in healthy adults. *Psychological Science*, *29*(9), 1436–1450. https://doi.org/10.1177/0956797618772540

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary
Appendix S1. Demographic Data.
Appendix S2. Items.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Appendix S3. All Screens from Training.
Appendix S4. Coded Comments by Participants.
Appendix S5. Participant Comments Arranged by Productive Score.
Appendix S6. Perceived Enjoyment, Ease, and Effectiveness of the Treatment by Group.
Appendix S7. All Models' Syntax, Means, and Effects.
Appendix S8. Posttest Comments About Declarative Knowledge.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12585 by Readcube (Labtiva Inc.), Wiley Online Library on [28/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License