UNIVERSITAT DE BARCELONA

Digital Flashcards for English Grammar: A Pilot Study in Rural Cambodia

Master's thesis submitted by

Jonathan Serfaty

Supervisor: Dr. Raquel Serrano

Applied Linguistics and Second Language Acquisition in Multilingual Contexts

Department of English Studies and Modern Languages and Literatures

Universitat de Barcelona

Academic Year 2018-2019

Acknowledgements

This paper would not have been possible without the help and cooperation of many people. First of all, my supervisor Dr. Raquel Serrano who approved this topic and allowed me to pursue my research in Cambodia, helping me along the way over many a Skype call. Without this guidance, I would have been lost. Secondly, the volunteers who bravely gave up their smartphones for use in this study, and helped me keep track of each participant's progress in the training schedule. I simply could not have carried out this research without this help. Finally, of course, the wonderful students of Green Village School who welcomed the training and tests with inspiring enthusiasm. I have never met teenagers with more motivation to succeed. Their drive to change their lives through hard work and perseverance are a lesson to us all.

Table of Contents

Digital Flashcards for English Grammar: A pilot study in rural Cambodia

Abstract

Digital flashcards are widely used and studied for vocabulary memorisation, but there has been no previous research into using this tool for grammar learning. This study aims to address this gap by examining whether full-sentence flashcard training could cause learners to notice grammatical patterns in their output and apply these inferred rules to novel sentences. The participants were school-aged students in rural Cambodia, where English proficiency is highly valued but difficult to obtain. In a pre-test / post-test design, students spent eight days typing translations from their L1 Khmer to English using the smartphone app Cram.com Flashcards, with each item repeating in a cycle until answered without errors. Post-tests of trained and untrained items took place one day, two weeks, and eighteen weeks after treatment. Results showed high relative gains for all students (M = 81%) and minimal losses at the final post-test. Equal results between trained and untrained items demonstrated that participants had indeed inferred grammar rules from the training, and a refresher session for one group fully mitigated losses. The findings are discussed in terms of the facilitating effect of output on form acquisition, and it is recommended that further research into digital flashcards for grammar is conducted under different conditions, to better understand which factors influence gains. It is further recommended that apps be used in environments where trained teachers and other resources are unavailable.

1. Introduction

With the rise of portable online devices, the possibilities for independent language learning have increased exponentially. Countless mobile applications designed to teach languages are now available, as well as limitless access to authentic language samples through online news, TV streaming, or YouTube. One of the simplest examples of these tools is digital flashcards. Users select or create custom lists of items which they want to learn, and then test themselves by trying to recall the item while viewing its translation. Online flashcards can be accessed on any connected device, allowing learners to keep track of new words they encounter on the go, and practice them at any moment. While other learning apps are only available in the most popular languages, flashcards cater for any language with typing capabilities, however obscure they might seem to the companies creating them. This

opens up the opportunity to learn languages in places whose native language is not usually represented, and who may not have the option of travelling abroad or even meeting foreigners. In other words, these new technologies democratize the ability to learn a foreign language.

Second Language Acquisition (SLA) research into flashcard training has, until now, solely focused on vocabulary learning. Findings have been largely positive (e.g. Dizon, 2016; Andarab, 2017; Sanosi, 2018), but vocabulary is only one component of language. Especially for learners with typologically distant first and second languages, understanding how to put words together to form meaning in a more grammatically nuanced language can be a real challenge.

This pilot study will explore the possibility of using flashcards to improve grammatical accuracy by using exemplary sentences of grammatical rules as flashcard items. Throughout this paper, the term "grammar" will denote native-like morphosyntax, without implying any deeper metalinguistic knowledge. Subjects will be students in a rural village in Cambodia, whose need to learn English is great, but who lack the resources to do so.

2. Literature Review

This section will first review previous research into digital flashcards for language learning, which has thus far focused on memorising vocabulary, followed by evidence from SLA theories that justify an attempt to adapt this tool for grammar learning.

2.1. CALL for Grammar

Computer assisted language learning (CALL) can refer to any use of technology for language learning, including translation devices, instructional websites, or smartphone applications. Advantages of CALL (summarized by Pokrivcakova, 2014; Obodoeze, 2018) include individualised pacing, reliable input, and self-tracking for autonomous learning. For low-resource environments in particular, CALL can be used where classroom time, materials, and trained teachers are lacking (Bikowski, 2018).

Research comparing computer instruction to human instruction have generally favoured the CALL groups. For example, McEnery, Baker and Wilson (1995) compared two approaches for teaching parts of speech: the traditional human-taught method and CyberTutor, a software which provided undergraduate English Language Learners (ELLs) instant feedback on annotations. In the post-test, CALL subjects scored 89.34% for accuracy compared with only 13.64% for the human-taught group. Likewise, Nutta (1998) compared human-instruction to computer-instruction among English as a Second Language (ESL) learners, where the software ELLIS provided lessons similar to the traditional teacher with video explanations and practice activities on selected structures. Open-ended writing tasks showed a significant advantage to the CALL group, with no differences on Cloze and multiple choice tests. In the two-week delayed post-test, the computer-led group actually increased their score, with speculation that the two-week gap allowed subjects to practice structures communicatively. Similar results have been found among English learners of other backgrounds. For example, Mohamad (2009) compared online grammar instruction and teacher-led instruction for Malaysian ESL students. The online group performed better in tests and also produced fewer errors in their essay writing. Abu Naba'h (2012) investigated teaching the passive voice with software to Jordanian school pupils learning English, with a significant advantage to the CALL group, and Abuseileek and Rabab'ah (2009) found similar results for instruction of verb tenses with EFL learners in Saudi Arabia.

More recently, software has been created to provide a unique language learning experience, rather than mimic a human teacher. Cerezo, Caras, and Leow (2016) compared 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 (83%, 91.3% vs. 63.2%, 60.2%) and far higher retention on the two-week post-tests (72.6%, 81.6% vs. 38.2%, 39.7%). They concluded that CALL could replace teacher-led instruction and create more class time for communicative activities. 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.

All these studies showed computer instruction to be at least as good as human instruction. However, they all relied on software designed to teach specific rules to learners of specific native languages (L1s) and target languages (L2s), using computers. This is not generalizable 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, customizable flashcards are accessible to anyone with internet access, in any written language, and are already widely used for language learning.

2.2. Digital Flashcards for Vocabulary

Flashcard training comprises two stages (The Two-Stage Framework, described in Kornell & Vaughn, 2016). In the retrieval stage, the learner sees a stimulus cue and attempts to produce the paired-associate. In the feedback stage, the target item is presented. The process varies in the type of output required from the user, the strictness of what is accepted as correct, the addition of audio, images, or hints, and the criterion for how many times an item must be answered correctly before disappearing from the pack (Nakata, 2011). Digital flashcards are similar to paper flashcards, but add more interaction, gaming elements, audio input, and personalised statistics that aid the learning process.

No previous studies using digital flashcards for grammar were found, but flashcards for vocabulary have been researched in depth. For instance, Carrier and Pashler (1992) found that recalling an English word from an Eskimo cue strengthened conceptual associations more than seeing both words simultaneously. Kang (2010) found a similar advantage for retrieval practice over restudy in learning Chinese logographs from English cues. Kang, Gollan and Pashler (2013) compared retrieval practice with imitation for learning Hebrew vocabulary. In the retrieval condition, subjects saw an image and attempted to produce the Hebrew word, while in the imitation condition they heard the target word while viewing the

picture, and repeated it. The retrieval condition outperformed imitation in both receptive (selecting the target picture) and productive (saying the target word) measures. More recently, the flashcard software Quizlet has been the focus of a range of vocabulary studies. In one example, Ashcroft, Cvitkovic and Praver (2018) compared low, intermediate, and high proficiency Japanese learners of English using Quizlet flashcards and paper flashcards to study 120 words from an academic word list. Both digital and paper flashcards were effective, but digital significantly outperformed paper among low-proficiency learners.

Retrieval Effort Hypothesis. Explanations for the benefits of retrieval followed by feedback, as in digital flashcards, 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 imitation drills is that retrieval demands more cognitive effort than repeating or reciting (Roediger & Karpicke, 2006). Pyc and Rawson's (2009) Retrieval Effort Hypothesis (REH) applied the principles of Bjork's framework to flashcard training. It posits that "difficult but successful retrievals are better for memory than easier successful retrievals". Pyc and Rawson tested the hypothesis by controlling two factors, which they called the Criterion and the Interstimulus Interval (ISI). The Criterion is the number of times an item must be correctly produced before being removed, so a criterion of two would mean that an answer must be produced correctly two separate times. The ISI is the number of items appearing between each attempt at a given target item. For example, an ISI of 5 would mean the user will see Item A, then 5 other items before another attempt at Item A. Their assumptions were that a lower criterion increases difficulty because the subject has had less opportunity to answer correctly before the post-test, and a higher ISI increases difficulty by increasing the time between having seen a target item and having to retrieve it. They produced evidence for these assumptions by measuring latency in answers, with longer latencies implying greater effort. Results showed that greater effort led to poorer performance during training but better results in the post-test, as predicted by Bjork's (1994) Desirable Difficulties Framework.

The Testing Effect. The REH focuses on successful but difficult retrievals, without addressing the effects of unsuccessful retrievals. In a review on retrieval effects, Kornell and Vaughn (2016) describe how even unsuccessful retrieval attempts are beneficial. In fact, the more confidently an incorrect response is given, the more effective the subsequent feedback.

This is known as the testing effect. The testing stage causes the learner to pay more attention to the feedback stage (see Roediger & Karpicke, 2006b; Kornell, 2009). This effect can be magnified by adding more testing stages. Izawa (1970) demonstrated this "Test-Potentiation" effect by comparing 5 conditions in which learners had 25 trials for each item. In the ST condition, trials alternated between Test and Study (Feedback), with each condition adding more Tests until STTTTT (five test trials between each study trial). Results showed that the effect of feedback in the study trial increased for every test trial that preceded it.

Drop-out Schedules. As mentioned above, a criterion of one is best for creating high retrieval effort. It is also best for efficiency. Pyc and Rawson (2007), compared drop-out schedules, whereby items drop out from the training set when answered correctly, to "conventional" schedules where items are shown equally until all have been answered correctly. They found that both yielded similar results, but with significantly fewer trials for the drop-out schedule. Retrieving well-known items added little to the learning process, in line with Bjork and Bjork's (1992, 2011) New Theory of Disuse which predicts greater learning for less well-known items. On the other hand, a higher criterion can lead to better long-term retention. Nelson, Leonesio, Shimamura, Landwehr, and Narens (1982) saw better retention when subjects were required to produce two correct responses before an item dropped out, and greater retention still when the criterion was four. This has been confirmed in subsequent studies (see Pyc & Rawson, 2009; Rawson & Dunlosky, 2011). In sum, a criterion of one is best for efficiency, but a higher criterion allows for more repetition, which is better for retention.

2.3. Theory Supporting Digital Flashcards for Grammar

The above research relates to using flashcards to commit single items to memory, which is perfect for vocabulary, but grammatical patterns can be applied to limitless combinations of vocabulary. Flashcards involve producing a target item from memory followed by feedback, much like a low-level language learner consciously attempting to produce accurate output. In fact, when *retrieval* is relabelled as *output*, then a feature of Swain's (1993, 1995, 1998) Output Hypothesis becomes highly reminiscent of the testing effect described above. According to the hypothesis, one of the main functions of output is to allow learners to notice the gap in their knowledge when their output is met with feedback.

Accordingly, flashcards for grammar learning can be framed as an attempt to cause the noticing of feedback by requiring output. The following section briefly describes the concepts of output and noticing within SLA, and outlines previous attempts to provide empirical evidence supporting output for noticing.

The Output Hypothesis was based on observations of French immersion schools in Canada. According to Krashen's (1985) Input Hypothesis, the students should have developed highly proficient French after 7 years of high quality input, and yet their morphosyntax was still inadequate in speech and writing. Swain put this down to a lack of opportunities to produce the language. As Swain (1995) explains, when producing the target language, learners may encounter a gap in their knowledge and consciously recognize what they need to learn. This gap can become evident when a learner realises they do not know how to communicate a message accurately, or by receiving negative feedback from an interlocutor, perhaps asking for clarification or misunderstanding them. Evidence from an empirical study into a learner's thought process showed that 40% of the times that learners encountered such a gap in output, they reflected on syntax and morphology (Swain & Lapkin, 1994, summarized in Swain, 1995). This process is commonly referred to as *noticing*.

According to Schmidt (2010), summarising the Noticing Hypothesis, "input does not become intake for language learning unless it is noticed, that is, consciously registered". Since the formulation of the hypothesis, attempts have been made to measure the effects of noticing using crossword puzzles (Leow, 2000), grammar exercises (Mennim, 2007), feedback (Mackey, 2006), and testing (Soleimani and Najafi, 2012). These researchers reported facilitative effects of noticing, but the issue of how to measure noticing has remained problematic. For example, Soleimani and Najafi measured noticing by how many words the students underlined. Underlining can be passive and does not guarantee what exactly the students have consciously registered. Meanwhile Leow noted each time a student verbalized their noticing, and Mackey measured self-reporting, but these measures do not account for non-reported, internal noticing. Discussing these limitations, Leow (2018) pointed out that asking participants to "think-aloud" does not reveal internal cognition, especially implicit processing, and could actually cause different levels of processing to occur, obscuring the data. However, more exact methods have been found in recent research

on eye-tracking technology. In one example, Godfroid and Uggen (2013) exposed 43 novice adult learners of German to 12 stem-changing verbs in sentences with different subjects. They found a positive relationship between time spent looking at verb stems and post-test production scores of those verbs. The effect was, however, quite low - approximately 2 seconds of extra time looking produced 8.6% higher probability of correct production in the post-test. This could be explained by the passive nature of reading. Participants were not instructed to attend to verb forms, only to read the sentences for meaning. Applying the Desirable Difficulties Framework here, reading seems not to generate enough effort to trigger sufficient levels of noticing for acquisition.

If noticing is required for acquisition, and a function of output is to enhance noticing, then output should be beneficial for acquisition. A series of studies have provided empirical evidence that training through output-plus-feedback does enhance noticing. The method of triggering output varies by study. An early example would be the clarification prompts used by Nobuyoshi and Ellis (1993), who encouraged 6 adult ELLs in Tokyo to repeat their utterances with accurate use of the English past tense form. They found that learners improved their accuracy when asked to clarify their meaning, and concluded that output provided the opportunity to increase control of already comprehensible forms, rather than teaching new forms. Later, Izumi and Bigelow (2001) compared groups of ESL students under four conditions that combined the requirement for output, in this case note taking, and enhanced input, which involved relevant sections being underlined and key words being highlighted. The four conditions were O+I+, O-I+, O+I-, and O-I-. They found greater noticing and learning of relative clause forms for the O+ groups with no measurable effect from enhanced input. Further evidence comes from Khatib and Alizadeh (2012), who gave Iranian ELLs different output tasks to promote noticing of past tense forms. One group heard a dictogloss and reconstructed it, while another wrote openly, based on visual cues and prompts. Both were provided with a model text and asked to underline important words or phrases, which constituted the measure of noticing. Results indicated that the reconstruction group displayed more noticing, and both output groups outperformed the input-only group. Reproducing exact strings of language, as in the reconstruction task and in flashcard training, was more effective than open output practice, in terms of enhancing noticing. Reconstruction was also used by Donesch-Jezo (2011), who compared 45 Polish medical students' learning

of appropriate metadiscoursal language under three conditions: Group A had enhanced input with metalinguistic explanation, Group B had the same enhanced input with prior instruction on rules and error correction in tasks but no metalinguistic explanations, and Group C (the output group) had non-enhanced input, with metalinguistic explanations, and also did a dictogloss reconstruction beforehand. Groups A and C outperformed Group B, with no significant difference between them in the immediate post-test, perhaps demonstrating an advantage for metalinguistic explanations. However, Group C significantly outperformed all groups in the two-month delayed post-test, indicating that output is beneficial for retention.

Every study mentioned above involved learners of privileged backgrounds, but the school system of the present participants provides little opportunity for critical thinking practice. There is reason to believe that the effects of noticing mentioned above may be differently effective for these participants. 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 (Philps, 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). With this in mind, previous research may not be a good indicator of learning outcomes in the present sample. In fact, reviews of SLA sample demographics (Norris & Ortega, 2000; 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 there has been a recent call for SLA research for more diverse populations ("SLA for all?", 2019), with a wider variety of languages and socioeconomic backgrounds. This paper may be seen as a step towards addressing this gap.

2.4. Summary

Flashcards constitute a form of CALL which can be adapted without limit or cost for learners in low-resource environments, with underrepresented L1s. Most research has focused on memorising isolated vocabulary items and improving efficiency by manipulating the number of items, timing, and number of repetitions. When applying this tool to learning

grammatical patterns, we must look to other theories. Studies have shown that output tasks enhance noticing, which in turn lead to higher grammatical accuracy. Flashcards provide the necessary conditions of output-plus-feedback which, according to previous research, should trigger learners to notice the gap in their knowledge and facilitate the acquisition of grammatical structures. That being said, different outcomes may be expected among less privileged populations.

3. The Present Study

The primary aim of this study is to discover whether flashcards can be used to improve L2 grammatical accuracy. A secondary aim is to investigate a potential solution for language learners without access to formal education or other language learning opportunities. Flashcards have already been widely used for memorizing vocabulary items, but this study will explore their use at the sentence level to induce pattern learning. The research questions are as follows

RQ1: Can full-sentence flashcard training cause the acquisition of generalizable grammatical patterns?

RQ2: How effective will the training be on grammatical accuracy

(i) immediately after treatment?

(ii) two weeks after treatment?

(iii) eighteen weeks after treatment?

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

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. Franciosi Yagi, Tomoshige & Ye, 2016; Gilsang, 2018; Ashcroft et al., 2018). This was a pre-emptive attempt at mitigating expected losses. The refresher was a single training session for half the participants (see section 4.4.1) intended to reawaken knowledge acquired in the treatment.

4. Methodology

4.1. Context

The setting was a rural village in Cambodia, where education is a major concern. Wealthy families are able to send their children to English-speaking international schools in the capital, but the majority of citizens, living in the provinces, have only state schools operating a few hours per day with teachers who themselves have had little education. Despite the school system officially being free, students are expected to pay daily bribes in return for attending classes and getting a passing grade. Those that cannot afford the bribes must drop out of school, leaving them with few prospects. The challenge of breaking this cycle seems hopeless to many. Some dream of studying abroad on a scholarship, while others simply want to make a liveable wage in the tourism or NGO sector. Whatever the ambition, a common theme is the need to communicate in English. Tourists from around the world use English as a common language, NGOs use English in daily operations among international staff and donors, and English is also the official language of the ASEAN community, the economic partnership that unites countries in South-East Asia. Adding to this need is that the online world and all it has to offer is generally inaccessible in the Cambodian language, Khmer, despite the fact that internet access is now widespread in the provinces thanks to portable devices. Typing in Khmer is made difficult with 74 base characters, not counting the subscripts, rounded style, and diacritics. The lack of available samples and differences in typology mean that translation software, such as Google Translate, produces confusing results. In short, English proficiency is needed to access online resources, study, or earn enough money to make a change.

4.2. School

Green Village School is a grass-roots initiative to provide English education for local school-aged children. This unofficial free school created by one ambitious resident had been open for one year at the time of this study. The school welcomes short-term foreign volunteers to teach the approximately 150 students who attend every day in a makeshift outdoor classroom in addition to their regular schooling. Although problematic in many

ways, the students do use English every day to communicate with the volunteers and the result has been a high level of fluency developed in a short time. In contrast, morphosyntactic accuracy has not developed well and several factors may be responsible. Firstly, none of the volunteers have been native speakers or trained teachers. This, combined with textbooks that are full of errors, limit the accuracy of input. Secondly, little positive transfer is possible. The Khmer language, of the Austroasiatic family, has no inflections, conjugations, or cases. Tenses and questions are inferred with auxiliary words and particles, and politeness is communicated in vocabulary and pronoun choices. For example, the word for "eat" changes depending on whether the subject is older, younger, an animal, a monk, or a king, but with no inflections for person, quantity, or tense. Students have developed an interlanguage based on Khmer syntax, distant enough from the target language as to cause problems in communication.

4.3. Participants

The school has ten classes distributed into six levels of age and ability. All students from levels four to six were recruited for this study, while lower-level students were considered too young to participate. Within these classes, some students did not participate because they were not available for all of the 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, with clusters around ages 12 and 15.

4.4. Instruments

4.4.1. Treatment

Tool. The tool was a free app by *Cram.com*, which allows users to create custom flashcards. This tool was chosen for several reasons. It is easy to use and modify from any phone or computer, does not require an internet connection once synced, and supports the Khmer script. Settings allow users to choose what is accepted as correct. In this study, errors in letter case, punctuation, or spaces were ignored. This was useful for participants who were inexperienced with typing in English. Khmer has no spaces, capitalization rules, and very

little punctuation, and these added complications would have constituted an interfering factor in the study. The app's "Memorize" mode was employed, whereby the flashcards automatically drop from the cycle when answered correctly once. "Text-input" was activated, requiring written answers from participants. Through this, it is guaranteed that each item will appear correctly once in every participant's output, and that every incorrect answer will be met with feedback in the form of a textual recast. 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.

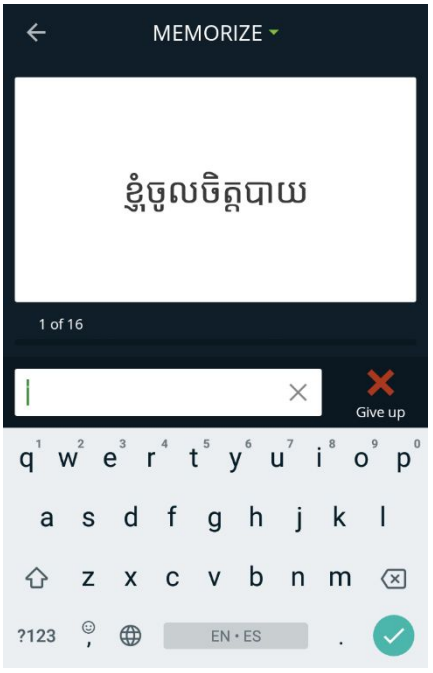


Fig. 1: Participants see the item in Khmer and must type the English translation.

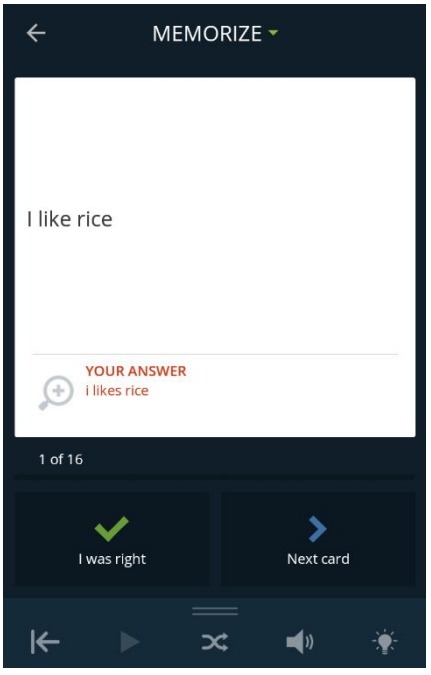


Fig. 2: Participants see their response along with the target form.

Items. The target items were full sentences, grouped into categories (although each item contained multiple grammatical features). The first four groups were declarative sentences: (1) present simple, (2) present continuous with *is*, (3) present continuous with *am/are*, and (4) *there is/are*. The remaining sets were the same items in the interrogative form. For example, the first item of group 1 was "I like rice" and the first item of group 5 was "Do you like rice?". Each group consisted of 5 items, making a total of 40 items in the treatment. The chosen items were simple sentences using vocabulary the participants already knew and used regularly, based on this researcher's experience in the context. The rationale

of using familiar vocabulary was to keep the focus of the study on grammatical form. Items provided the opportunity to practice common errors, again based on 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.

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 1 - Items of the treatment

Item Distribution. Flashcards were organized into 8 sets, and groups were distributed so that only five items, or one group, were introduced for the first time in each set. Thus, the first set contained only the first group, and the second set repeated the first group while introducing the second group. As the items became more familiar, the size of the sets increased to keep the retrieval effort high and to allow for items to be repeated on different days. This corresponds with previous research (see section 2.2) which found that repetition led to higher retention. The distribution of groups and sets is shown in Table 2.

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

Table 2 - Distribution of items across sets.

*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 was the easiest and group 7 had been seen twice within the final three sets, the assumption being that recency would compensate for fewer repetitions.

Refresher. The refresher was a single set of flashcards containing the 16 items that appear in Test A (see section 4.4.2), two items from each group. The refresher was intended to remind students of previously acquired knowledge with a single study session of test items.

4.4.2. Tests

Two tests were used (see Appendix 1). Test A 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 comprises an equivalent 16 items, using only vocabulary and grammatical structures found in the treatment, but in novel combinations. Test A is designed to test how well participants can reproduce items seen in training, while Test B is designed to test whether participants can generalize those grammatical patterns for novel items. Therefore, Test A and Test B represent trained and untrained items respectfully. Test B was added because if trained items scored highly, we would not know whether the gains were due to memorising or from learning the morphosyntactic rules. If untrained sentences score similarly to trained sentences, presuming they are significantly higher than the pre-test, then results must be down to grammar acquisition.

4.5. Procedure

The pre-test was administered on smartphones using Google forms. All 3 post-tests (immediate, 2-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 for any tests and were encouraged to check answers thoroughly before submitting, this is not expected to have affected results.

Each test was coded for when it took place and which items it included. The 'T' stands for Time of testing, followed by a '1' for the pre-test, '2' for the immediate post-test, '3' for the two-week delayed post-test and '4' for the eighteen-week delayed post-test. "A" denotes trained items and "B" denotes untrained items. For example, T2A represents trained items in the immediate post-test.

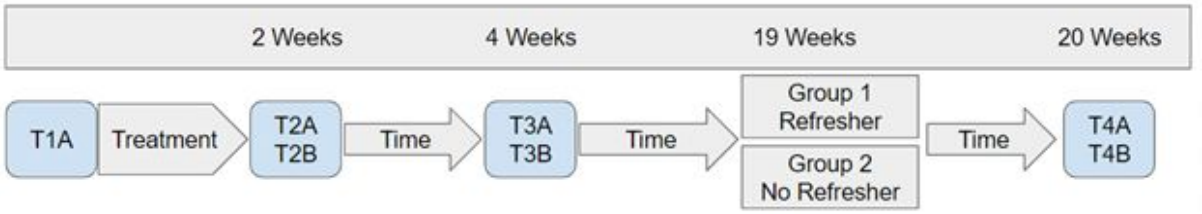
Participants first completed Test A (items to be trained) as a pre-test (α = .852) and started the treatment the following day. Participants completed each set individually on separate days, using smartphones from volunteer teachers. Due to sporadic availability, many participants missed a day and caught 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.

The day after the final treatment session, participants took Tests T2A (α = .744) and T2B (α = .744). They were not given advanced warning of delayed post-tests. The first delayed post-test (T3A: α = .733; T3B: α = .669) was given two weeks after treatment (as per Nutta, 1998; Cerezo et. al, 2016), and the final post-test (T4A: α = .734 ;T4B: α = .749) was given eighteen weeks after the immediate post-test. One week before this final test, the sample was divided into two groups[1] matched for age, gender and previous scores. Group-R

[1] The original groups were matched for all previous scores. However, due to absences on the day, members of Group-R ended up with overall higher scores and gains at T2A, and fewer members. To address this imbalance, three participants were excluded from the data for T4 to preserve the comparability of the groups.

(N = 14) took a refresher treatment, one set of the same 16 items from Test A, while Group-NR (N = 14) had no extra treatment. Figure 3 illustrates the experimental design.

Fig. 3



T = Testing Time
A = Trained Items
B = Untrained Items

Number of weeks starts from T1. For example, the two-week delayed post-test took place at the "4 Weeks" mark.

4.6. Scoring

Items were scored dichotomously, 1 point for each correct answer and 0 for incorrect answers. To be correct, answers must match the target item exactly, with the following exceptions: (1) If a base vocabulary word was spelled incorrectly but otherwise used correctly, for instance "gril" instead of "girl"; (2) If the wrong vocabulary 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 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 1 test per participant at random (14.2% of total tests), with interrater agreement of 100%.

4.7. Analysis

As results were not normally distributed (see Appendix 2) and the sample was small, analysis was conducted using non-parametric tests which test for equal distribution of median ranks and report a p-value (alpha = 0.05) for whether the null-hypothesis of equal distribution should be accepted or rejected. Test scores for trained and untrained items were compared

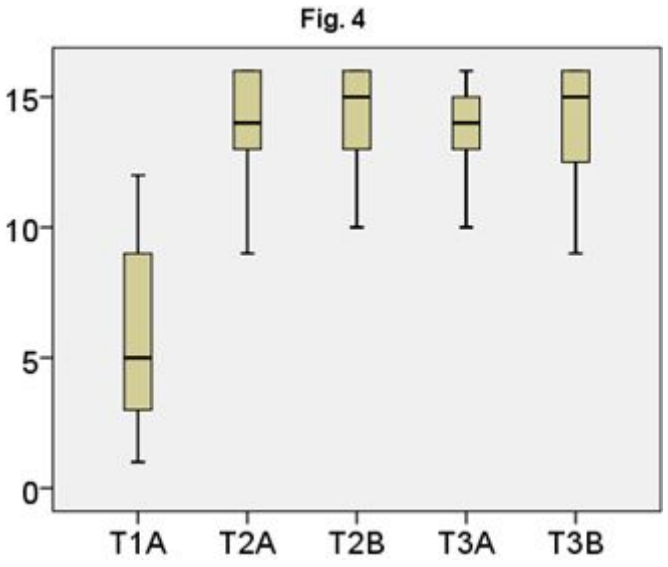
within subjects using the Related-Samples Wilcoxon Signed Rank Test for T2, T3, and T4, to check if results were due to memorisation or grammar learning. Next, Test A scores were compared between times using Independent Mann-Whitney *U* tests to establish the amount learned and retained. Relative gains 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 (as per Peters & Webb, 2018) was (learned items/(total number of items - known items)) x 100. 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. Mann-Whitney *U* tests were used to confirm equal distribution of age, gender, and previous scores between groups. The comparison between T3 and T4 scores constituted the measure of forgetting for each group.

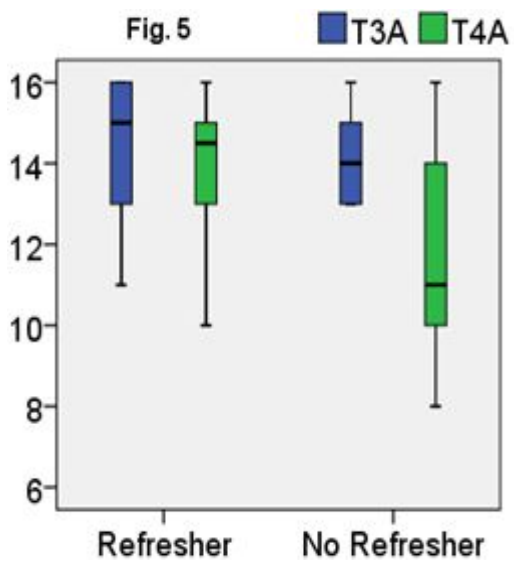
5. Results

Descriptive statistics of the results for all tests are displayed in Table 3. Figure 4 shows boxplots for T1, T2 and T3 with all participants. Figure 5 shows boxplots for T3 and T4 according to groups, with and without the refresher.

Table 3	**Mean Raw Score***	**Median Raw Scores***	**Mean Gains %**	**Median Gains %**
T1A	5.71 (3.42)	5	-	-
T2A	13.81 (2.29)	14	81.35 (18.98)	85.71
T2B	14.39 (1.96)	15	-	-
T3A	13.39 (2.55)	14	76.24 (17.57)	80
T3B	13.94 (2.14)	15	-	-
T4A	R: 14 (1.96) NR: 11.79 (2.52)	R: 14.5 NR: 11	R: 79.03 (13.32) NF: 65.19 (23.48)	R: 79.29 NF: 59.42
T4B	R: 14.43 (2.21) NR: 12.64 (2.27)	R: 15.5 NF: 12.5	-	-

*Maximum score = 16
T4 data are reported according to R = refresher; NF = no refresher.





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. Table 4 shows the results.

Table 4	**Median Rank Test A**	**Median Rank Test B**	**Z Score**	**Sig**
T2	14	15	2.441	0.015
T3	14	15	1.79	0.074
T4[2]	13	14	1.29	0.010

At T2, Test A scores (M = 13.81, SD = 2.29, Mdn = 14) were significantly lower than Test B scores (M = 14.39, SD = 1.96, Mdn = 15), Z = 1.54, 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 = 0.074. 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 = 1.29, 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

[2] All 31 participants are used in this calculation, regardless of experimental group, because the comparison is within subjects. If split for group, Group-R: Z = 1.897, p = .058; Group-NR: Z = 1.845, p = .065.

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 with "You", "She, 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 than its counterpart. When item *13* (Are the girls eating?), the biggest outlier, is removed from the data, then no significant differences are found between trained and untrained items for T2 (*Z* = .651, *p* = .515), T3 (*Z* = .775, *p* = .439), or T4 (*Z* = 1.083, *p* = .279).

5.2. Gains and Retention

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 highly significant (*Z* = 4.874, *p* < .00001). When converted to relative gains, the immediate post-test (T2A) showed mean gains of 81.35% (*SD* = 18.98, *Mdn* = 85.71%), ranging from 40% (*N* = 1) to 100% (*N* = 10).

The two-week delayed post-test (T3A) produced a mean score of 12.42 (*SD* = 2.55, *Mdn* = 14), which in relative gains from pre-test is 76.56% (*SD* =17.56, *Mdn* = 80.00). The mean score dropped by 1.35, but a Related-Samples Wilcoxon Signed Rank Test found no significant difference between T2 and T3 scores, *Z* = .898, *p* = .369.

At the eighteen-week post-test (T4A), Group-NR (*N* = 14) scored 11.79 (*SD* = 2.52, *Mdn* = 11). T3A scores for this subset (*M* = 13.64, *SD* = 1.87, *Mdn* = 14) were significantly higher than at T4, *Z* = 2.040, *p* = .041. The final overall gains at T4 were 65.19% (*SD* = 23.48, *Mdn* = 59.42).

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 statistically different, Z = .051, p = .959. This difference is salient considering that Groups -R and -NR were matched for distribution of T3A scores, Z = 1.222, p = .246. For this group, the final overall gains were 79.03% (SD =13.32, Mdn = 79.29).

An Independent-Samples Mann-Whitney U Test revealed that the difference in gains at T4A between Group-R (M = 79.03, SD = 13.32, Mdn = 79.29) and Group-NR (M = 65.19, SD = 23.48, Mdn = 59.42) did not reach statistical significance, U = 58, p = .069.

6. Discussion

This paper set out to explore whether flashcards may be used to improve grammatical accuracy. ELLs aged 9 to 17 in a low-resource, low-education context underwent 8 sessions of flashcard training in which they produced target language samples, prompted by translations from their L1 Khmer. In each session, each item had to be typed correctly once for it to drop from the set. Otherwise, participants were presented with the target response and the item returned to the cycle. Each research question will now be discussed in turn.

RQ1: Can full-sentence flashcard training cause the acquisition of generalizable grammatical patterns?

The first research question asked whether the treatment led to acquisition of grammatical patterns, as opposed to memorisation. To test this, scores from items used in training were compared to equivalent items using the same vocabulary and structures but in novel combinations. The results showed that there was 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, the trained items should have scored higher than untrained items, according to the experiment's rationale. Furthermore, if neither memorisation nor grammar learning had taken place, then post-test scores would logically be similar to pre-test scores. Given that scores on the pre-test were low (M = 5.71/16) and that post-test scores were high (T2A: M = 13.81/16, T2B: M = 14.89/16) we can confidently

conclude that grammatical patterns were learned as a result of the treatment. This is especially interesting as participants were never instructed to infer rules from the samples, nor that there would be a post-test of untrained items. No rules or explanations were given with the samples, which means that students must have either been formulating new rules about form, or cementing previously taught rules into their interlanguage. Assuming 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 (Nobuyoshi & Ellis, 1993; Izumi & Bigelow, 2001; Donesch-Jezo, 2011; Khatib & Alizadeh, 2012) and that flashcard training provides the necessary conditions for this to occur.

RQ2: How effective will the training be on grammatical accuracy
(i) immediately after treatment?
(ii) two weeks after treatment?
(iii) eighteen 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 = 81.56%), and include 10/31 participants with 100%. Gains remained high after two-weeks (M = 76.24%) and eighteen-weeks (Group-R: M = 79.03%; Group-NR: M = 65.19%). The drop in scores is statistically visible for the non-refresher group when comparing scores for the two-week delayed post-test (M = 13.64/16) and the eighteen-week delayed post-test (M = 11.79/16), though only by two items. The findings are somewhat similar to Cerezo et al.'s (2016) results from videogame instruction among beginners, which also used written translation post-tests, with gains of 83% at immediate post-test and 63.2% 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%. It seems, based on these data, that flashcard training for low-level learners may actually be more effective for grammar than for vocabulary, in terms of long term learning.

The question now arises why flashcards led to such high retention at the two-week delayed post-test, compared with Cerezo et al.'s videogame. This may be partially explained

by having repeated each item on at least two days, mostly three days, producing the same effect as having a higher criterion (see section 2.2). Retention may also have been aided by the participants practicing the newly learned forms in their daily English output, as Nutta (1998) suggested, and noticing the forms in their input. Moreover, while scores did not markedly drop, they also did not climb, inferring that retention was due to the treatment and not any formal learning that may have occurred in the intermediate time frame.

Nobuyoshi and Ellis (1993) specified two stages of acquisition: (1) internalizing new forms, and (2) increasing control over already internalized forms, concluding that output tasks were probably only useful for the latter. The high retention rate in this study certainly suggests that participants entered the second stage of acquisition. Even if a form was completely new to the students, they were forced to internalise it on its first appearance by repeatedly attempting to produce it until successful. By revisiting these items on subsequent days, students were working with items that had already passed through the first stage of acquisition. Framed in these terms, digital flashcards may accelerate the rate of acquisition through these two stages.

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. The difference between groups' relative gains, with respect to the pre-test, approached but did not meet statistical significance. This implies that overall, the two groups learned a similar amount over the course of the study. However, when comparing between T3 and T4 scores, it seems that Group-R managed to maintain their previous knowledge from T3, whereas Group-NR showed small but significant losses over time. It should be noted that the overall retention was high, meaning that the refresher only needed to be powerful enough to prevent the losses of one or two items. It remains unclear, therefore, 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.

6.1. Implications for Theory

Most previous research into the use of flashcards, linguistic or non-linguistic, has been related to memory. The present results show that untrained items improved to the same extent as trained items, which means that memorising was not responsible for these gains. Given that flashcards exclusively revolve around repeated output and feedback, it would be reasonable to cite these findings as further evidence in support of the noticing function of the Output Hypothesis. The output stage prompted learners to notice gaps in their knowledge, while the feedback stage allowed them to compare their output with the target response. This "forced noticing" led to successful acquisition of grammatical forms. 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 forcing noticing for each item. Of course, 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 different proficiency levels, languages, and socioeconomic backgrounds. It would also be interesting to investigate which types of grammatical structures benefit from this training, and if factors such as intensity, repetition, or type of output could be manipulated to optimise the process. Furthermore, the only tests in this study were written translations from the L1. It is unknown how the treatment affected the students' other facets of language, such as spontaneous speech and open-ended writing.

6.2. Implications for Practice

In general, teachers in all contexts should be encouraged to use full-sentence flashcards with their students. In doing so, they allow every student to notice and practice forms independently, allowing more time for meaning-focused activities in class. More specifically, these 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 well-trained teacher and genuine interaction cannot be replaced, but such teachers are not as ubiquitous as one would hope. With many NGOs focusing on training local teachers, who are themselves undereducated, it may be wise to first invest in devices and internet connections so that students can access free learning apps in the short term. By doing this, learners will have access to reliable, consistent input, formulated by experts, allowing them to study at their own pace, while being guaranteed quality feedback on their work.

7. Limitations

There a number of limitations that could be improved upon in future research. First of all, although the subjects in this study represented all members of an available population, the sample was small and data were not normally distributed. Secondly, the tool used does not create filler items or activities, meaning that when one item remains in the cycle, the user reproduces it immediately after seeing feedback. This reduces the effort for the most difficult items. Lastly**,** the tests, which were designed by the researcher for the purposes of this study, could be improved. In hindsight, more attention should have been given to the complexity of item subjects, to ensure equal difficulty between paired trained and untrained items. Additionally, the pre-test was carried out on Google forms, as opposed to the pen and paper method for all post-tests. This was a logistical issue and although it does not seem to have affected results, future studies should bear this in mind.

8. Conclusion

This pilot study investigated the use of flashcards for grammar learning. Flashcards have previously been tested for their effectiveness in learning vocabulary, but the high gain scores of this study provide evidence that flashcards should also be investigated for grammar learning. Participants 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 traditional learning opportunities. It is hoped that more SLA research 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.

Word Count: 8749

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 (IJEDICT)*, *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

Bikowski, D. (2018). Technology for Teaching Grammar. In *The TESOL Encyclopedia of English Language Teaching* (pp. 1–7). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784

Bigelow, M., Delmas, R., Hansen, K., & Tarone, E. (2006). Literacy and the processing of oral recasts in SLA. TESOL Quarterly, 40, 665–689.

Bjork, E. L., Bjork, R., Roediger, H., McDermott, K. B., & McDaniel, M. A. (2010). Making Things Hard on Yoursel , But in a Good Way: Creating Desirable Difficulties to Enhance Learning.

Bjork, R. A. (1999). Assessing Our Own Competence: Heuristics And Illusions. In *Metacognition: Knowing about knowing* (pp. 185 –205). Cambridge, MA.

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., & Bjork, E. L. (1992). A New Theory Of Disuse And An Old Theory Of Stimulus Fluctuation. In S. M. Kosslyn & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (pp. 35–67). Hillsdale, NJ: Routledge.

Carrier, M., & Pashler, H. (1992). The Influence Of Retrieval On Retention. *Memory & Cognition*, *20*(6), 633–642.

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*(02), 265–291. https://doi.org/10.1017/S0272263116000139

Dizon, G. (2016). Quizlet In The EFL Classroom: Enhancing Academic Vocabulary Acquisition Of Japanese University Students: Discovery Service For Universitat De Barcelona. *Teaching English with Technology*, *16*(2), 40–56.

Cram.com, LLC. (2016). Cram.com Flashcards (1.6.1) [Mobile application]. Retrieved from https://www.apk4now.com/apk/52746/cram-com-flashcards/download

Donesch-Jezo, E. (2011). The Role Of Output And Feedback In Second Language Acquisition: A Classroom-Based Study Of Grammar Acquisition By Adult English Language Learners. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri*, *2*(2), 9.

Franciosi, S. J., Yagi, J., Tomoshige, Y., & Ye, S. (2016). The Effect of a Simple Simulation Game on Long-Term Vocabulary Retention. *CALICO Journal*. https://doi.org/10.1558/cj.v33i2.26063

Godfroid, A., & Uggen, M. S. (2013). Attention To Irregular Verbs By Beginning Learners Of German. *Studies in Second Language Acquisition*, *35*(02), 291–322. https://doi.org/10.1017/S0272263112000897

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining The Critical Period Hypothesis. *Studies in Second Language Acquisition*, *16*(01), 73. https://doi.org/10.1017/S0272263100012596

Izawa, C. (1970). Optimal Potentiating Effects And Forgetting-Prevention Effects Of Tests In Paired-Associate Learning. *Journal of Experimental Psychology*, *83*(2, Pt.1), 340–344. https://doi.org/10.1037/h0028541

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 (Vol. 21).

Jo, G. (2018). English Vocabulary Learning With Wordlists vs. Flashcards; L1 Definitions vs. L2 Definitions; Abstract Words vs. Concrete Words. *Culminating Projects in English*.

Juffs, A. (2006). Working Memory, Second Language Acquisition and Low-Educated Second Language and Literacy Learners. In *LOT Occasional Series* (Vol. 6, pp. 89–104). LOT, 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. 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

Krashen, S. D. (1985). *The Input Hypothesis: Issues And Implications*. Longman.

Larsen-Freeman, D., & Long, M. H. (1991). *An Introduction To Second Language Acquisition Research*.

Leow, R. P. (2000). A Study Of The Role Of Awareness In Foreign Language Behavior: Aware Versus Unaware Learners. *Studies in Second Language Acquisition*, *22*(4), S0272263100004046. https://doi.org/10.1017/S0272263100004046

Leow, R. P. (2018). ISLA: How implicit or how explicit should it be? Theoretical, empirical, and pedagogical/curricular issues. *Language Teaching Research*, 136216881877667. https://doi.org/10.1177/1362168818776674

Mackey, A. (2006). Feedback, Noticing And Instructed Second Language Learning. *Applied Linguistics*, *27*(3), 405–430. https://doi.org/10.1093/applin/ami051

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

Mennim, P. (2007). Long-Term Effects Of Noticing On Oral Output. *Language Teaching Research*, *11*(3), 265–280. https://doi.org/10.1177/1362168807077551

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. (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

O. Nelson, T., Leonesio, R., P. Shimamura, A., F. Landwehr, R., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Vol. 8) 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

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A researchsynthesis and quantitative meta-analysis.Language Learning,50, 417–528.

Nutta, Joyce. (1998). Is Computer-Based Grammar Instruction as Effective as Teacher-Directed Grammar Instruction for Teaching L2 Structures?. *CALICO Journal*. 16.

Obodoeze, F. C., & Obodoeze, N. (2011). *Computer Based Technology In Foreign Language Education In Nigeria For Sustainable Development*.

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*. https://doi.org/10.1177/1362168818819027

Plonsky, L. (2014). Sampling, power, and generalizability in L2 research (Or, why we might as well be flipping coins). Keynote presentation at the Second Language Studies Symposium, East Lansing, MI.

Pokrivcakova, S. (2014). 2.1 CALL And Teaching Vocabulary. In *CALL and Foreign Language Education* (1st ed., pp. 38–41). Constantine the Philosopher University. https://doi.org/10.17846/CALL.2014.24-28

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. (2007). Examining The Efficiency Of Schedules Of Distributed Retrieval Practice. *Memory & Cognition*, *35*(8), 1917–1927.

Rabab'ah, G. A., & AbuSeileek, A. F. (2009). The Effect Of Computer-Based Grammar Instruction On The Acquisition Of Verb Tenses In An EFL Context. *The International Arab Journal of Information Technology*, *6*(4).

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

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

Sanosi, A. B. (2018). The Effect Of Quizlet On Vocabulary Acquisition. *Asian Journal of Education and e -Learning*, *6*(4), 71–77. Retrieved from https://www.researchgate.net/publication/327108959_The_Effect_of_Quizlet_on_Vocabulary_Acquisition

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, … I. Walker (Eds.), *CLaSIC* (pp. 721–737). Singapore: National University of Singapore, Centre for Language Studies.

Soleimani, H., & Najafi, L. (2012). The Noticing Function Of Classroom Pop Quizzes And Formative Tests In The Uptake Of Lexical Items Of EFL Intermediate Learners. *International Journal of English Linguistics*, *2*(4), 73–84. https://doi.org/0.5539/ ijel.v2n4p73

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. (n.d.). The Output Hypothesis: Just Speaking And Writing Aren't Enough. *The Canadian Modern Language Review*, *50*, 158–164.

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.

APPENDIX 1

	Test A		Test B	
Item	**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	ខ្ញុំកំពុងញុំ	I am eating	ខ្ញុំកំពុងលោត	I am jumping
6	ក្មេងៗប្រុសកំពុងតែញុំ	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	តើក្មេងស្រីៗកំពុងញុំ?	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?

APPENDIX 2

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
T1A	.155	31	.057	.921	31	.025
T2A	.183	31	.010	.855	31	.001
T2B	.246	31	.000	.792	31	.000
T3A	.214	31	.001	.840	31	.000
T3B	.222	31	.000	.846	31	.000
T4A	.164	31	.034	.930	31	.045
T4B	.154	31	.059	.894	31	.005

a. Lilliefors Significance Correction