SPACED RETRIEVAL IMPROVES RETENTION AND TRANSFER OF FOREIGN VOCABULARY

by

DURNA ALAKBAROVA

THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Experimental Psychology at the University of Texas at Arlington

May, 2022

Arlington, Texas

Supervising Committee:

B. Hunter Ball, Supervising Professor

Daniel Levine

Matthew K. Robison

ABSTRACT

Spaced Retrieval Improves Retention and Transfer of Foreign Vocabulary

Durna Alakbarova, M.S.

The University of Texas at Arlington, 2022

Supervising Professor: B. Hunter Ball

One strategy that has been shown to promote long-term retention is spaced retrieval practice. In this study, we used standard and dropout procedures of spaced retrieval training to examine whether they produce superior retention and transfer relative to repeated restudy in novel vocabulary learning. Across three sessions spaced two days apart, participants learned 30 Swahili words with their English definitions via spaced testing or by restudying. Definitions were the same (fixed) across days in Experiment 1 and different each day (variable) in Experiment 2. A week later, all participants were tested using either the same (repetition) or new (transfer) definitions from study. It was found that standard spaced retrieval practice improved recall of repetition and transfer items compared to repeated restudy with both fixed and variable learning. We also found that although the dropout procedure takes less time, standard spaced retrieval results in more transfer. These findings have important implications for choosing learning techniques to achieve the best outcomes.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Hunter Ball for his invaluable advice, continuous guidance, and patience with this thesis. I would also like to thank Dr. Daniel Levine and Dr. Matthew Robison for insightful feedback and suggestions that helped with the execution of this project.

TABLE OF CONTENTS

Spaced Retrieval Improves Retention and Transfer of Foreign Vocabulary

One of the most common study techniques employed by students is restudying material (Karpicke et al., 2009; Hartwig & Dunlosky, 2012). However, studies have shown that this technique is not the most effective at promoting long-term retention of learned material (Rivers, 2021). In contrast, retrieval practice has been shown to be a robust learning procedure (e.g., Butler et al., 2017; Roediger & Karpicke, 2006a; Agarwal et al., 2012). In a typical experiment on retrieval practice, participants first engage in a *learning phase*, wherein they study the material (e.g., the English-Swahili word pair baby – mtoto). Then, participants complete a *relearning phase*. In this phase, half of participants attempt to retrieve as much of the material as they can (e.g., baby – ???), while the other half repeatedly restudy the information (e.g., baby-mtoto). Either immediately or after a delay, all participants take a final test, the performance on which is used as an indication of learning. Studies using variations of this procedure show that engaging in retrieval by testing oneself during learning produces better recall than simply rereading the material. The memorial benefit of retrieval over restudying has been shown in a variety of materials and with different test types, and to be even more robust when used with corrective feedback (for a review see Rowland, 2014). Additionally, studies have found that this *testing effect* persists after a week in laboratory setting (Roediger & Karpicke, 2006b) and on semester and year-end exams in classrooms (Roediger et al., 2011; McDaniel et al., 2011). Moreover, retrieval has also been shown to promote the ability to apply knowledge to new tasks (i.e., *transfer*) better than restudy does (e.g., Butler, 2010; Carpenter, 2012; Butler et al., 2017; Pan & Rickard, 2018a). The present study is designed to evaluate the benefit of two spaced retrieval procedures over restudying on retention and transfer of knowledge in novel vocabulary learning.

Several theories have been developed to explain the underlying mechanisms of the testing effect. One such explanation is the *elaborative retrieval hypothesis*, which states that searching the semantic network when attempting to retrieve a correct answers on a test activates information relevant to the target, thereby creating a more elaborate memory trace for it (Carpenter, 2009; Carpenter & DeLosh, 2006). Some of this relevant information could then be reactivated by the cue during the final test and aid

in recalling the target. Restudying the material does not prompt a memory search to the same degree, and therefore does not lead to a similarly elaborate memory for the target. The elaborative retrieval hypothesis has been supported by the findings that longer lag between retrieval attempts during study results in higher recall than shorter lag (Rawson et al., 2015). Longer lag results in more forgetting than shorter lag and requires a more extensive memory search for the target. Higher recall rates with longer lag therefore show that a more extensive memory search leads to a more elaborated memory trace.

Another explanation of the testing effect is the *retrieval effort hypothesis* proposed by Pyc and Rawson (2009). This theory states that conditions that make retrieval more difficult during the relearning phase result in a better memory on the final test. Presumably, more difficult retrieval slows forgetting of information and creates a more elaborate memory for the target via a more extensive memory search. Pyc and Rawson (2009) provided support for this hypothesis with findings that increasing the number of stimuli between retrieval of each target – which allowed for more forgetting and resulted in more difficult retrieval – increased final recall. Additionally, they found that increasing the criterion level (the number of times the target must be correctly recalled) on the initial test led to diminishing returns on the final test. Higher criterion levels lead to better memory – and therefore easier retrieval – with each recall attempt. The finding that higher criterion (and by proxy, easier retrieval) lead to diminishing returns on the final test therefore supports the retrieval effort hypothesis.

To explain the testing effect in cued-recall tests, where participants study paired associates and then have to retrieve the target when presented with the cue at test, Pyc and Rawson (2010) developed the *mediator effectiveness hypothesis*. According to his hypothesis, when learning cue-target pairs, participants generate a keyword to tie the cue to the target in their memory (Carpenter, 2011). When the cue is then presented during the test, it presumably activates the mediator, which can aid in retrieval of the target information. Pyc and Rawson (2010) found that compared to restudying, retrieval practice improved memory for the mediators and the strength of the association between the mediator and the target. However, follow-up studies failed to find the relationship between mediator and target retrieval, or

the increased probability of mediator generation with testing (Cho et al., 2017; Lehman & Karpicke, 2016).

Spaced Retrieval and Long-Term Retention

While a single retrieval has been shown to benefit future recall over restudy, *spaced retrieval* improves memory even further (Roediger & Karpicke, 2006b). Spaced retrieval refers to a learning technique wherein participants retrieve information several times before the final criterion test. Three mechanisms have been used to explain the benefit of spaced retrieval for long-term memory. First, spaced retrieval includes *spaced practice*. Spacing learning over several sessions has been shown to produce better memory than completing all practice in one sitting (Cull, 2000; Pavlik & Anderson, 2005). When learning is done in different temporal contexts, each of those contexts gets associated with the target and creates a new retrieval route to that information. An alternative explanation of this mechanism is that unlike during massed retrieval practice, attempting to retrieve information during spaced practice requires reconstruction of that information and pathways leading to it from the previous study session. This results in a more meaningful organization of information, which facilitates its accessibility (Furst, 2020). Additionally, spaced retrieval introduces variability into the relearning sessions. This *variable encoding* links the target to more contextual information (not necessarily temporal), creating more retrieval routes and thereby increasing the probability of recall. Encoding variability can be introduced in many ways, such as practicing in different places (e.g., bedroom vs classroom) or with different materials (e.g., learning a concept by reading about it or by looking at a diagram). The last mechanism underlying the benefit of spaced retrieval is explained by the *mediator shift hypothesis* (Pyc & Rawson, 2012). According to this hypothesis, retrieving information repeatedly leads participants to realize that some of their mediators are not effective for recall, which causes them to shift to better mediators. Pyc and Rawson (2012) found that participants who engaged in spaced retrieval at learning shifted mediators more frequently and had higher recall on the final test than those who restudied material repeatedly. Moreover, Dikmans et al. (2020) found that spaced retrieval led participants to shift from mediated to direct retrieval, and this shift predicted better recall on the final test.

One specific form of spaced retrieval that has shown to produce strong memorial benefit is a dropout procedure (Bahrick, 1979). In a typical experiment using this paradigm, after initially studying the material, participants engage in retrieval to criterion with dropout. During this retrieval procedure, targets that are recalled to criterion (e.g., once) are dropped from the list of test items, while unrecalled targets (or those that have not yet reached criterion) are moved to the back of the testing list for another retrieval attempt. Participants then return after several days for another retrieval session with the same procedure. After several of such relearning sessions, participants take the final test. Successive relearning has been shown to be more beneficial than restudy and to produce high levels of retention both in laboratory and in naturalistic settings (Rawson et al., 2013; Rawson et al., 2018). Successive relearning benefits memory by engaging learners in elaborative retrieval, and through multiple relearning episodes, produces variability in encoding that results in multiple retrieval routes to the learned information (Rawson et al., 2013).

The main difference between the standard spaced retrieval procedure and the dropout procedure just described is that the former does not require retrieval for all items to be successful. In contrast, the dropout procedure requires successful retrieval of each item once. Thus, while both learning techniques reflect spaced retrieval, learning to a criterion of one may be more time efficient in terms of time. Although we compare the two conditions in this study, our primary interest lies in the difference in transfer between either of these spaced retrieval conditions and restudying.

Retrieval Practice and Transfer

Perhaps one of the most interesting findings from the retrieval practice literature is that testing not only benefits memory for previously studied information, but also learning of new information (Pan & Rickard, 2018b). Transfer refers to successful application of acquired skills or knowledge to new tasks that require similar skills or knowledge. For example, if participants learn that the Swahili word “mtoto” refers to “a very young child”, they could use transfer to deduce that the Swahili word for “a recently born human” would be “mtoto”. Facilitation of transfer is of great interest to researchers in the field of learning as it is a skill needed in many areas of everyday life. For example, schoolchildren must rely on concepts

learned in previous years (e.g., multiplication and division) in order to master new material (e.g., triangle area calculations) and adults must apply their skills from a previous position at work to a new position they are promoted to.

Butler (2010) examined whether the memorial benefit of retrieval can promote transfer. After studying a series of passages, participants engaged in four relearning blocks for the passages. Some passages were presented for restudy, while others were tested using questions for the information from the passages. For example, the former group reread a passage containing the concept, "A bird's wing has fairly rigid bone structure that is efficient at providing lift, whereas a bat has a much more flexible wing structure that allows for greater maneuverability", and the latter group had to answer a question, "A bat has a very different wing structure from a bird. What is the wing structure of a bat like relative to that of a bird?". A week later, all participants took a final test which consisted of questions on information from the passages (referred to as *repetition items*) and inferential questions for which the answers must be deduced (referred to as *transfer items*). For repetition items, participants saw the same questions as those asked during relearning. Transfer items involved applying information learned in the passages to a new domain. An example of a transfer counterpart to the above example was, "The U.S. Military is looking at bat wings for inspiration in developing a new type of aircraft. How would this new type of aircraft differ from traditional aircrafts like fighter jets?". The answer to this question would be that aircrafts modeled after bird wings would provide better lift, whereas those modeled after bat wings would provide better maneuverability. The results showed that retrieval practice not only enhanced memory for repetition items relative to restudy, but led to superior memory for transfer items as well. Butler posited that testing promotes transfer because retrieving information not only improves its storage strength, but also results in encoding of additional related information, thereby elaborating the memory trace and potentially creating more retrieval routes. He argued that these mechanisms allow for better understanding and ability to apply information to new contexts, thereby increasing the probability that transfer will occur.

Studies have also found that testing facilitates several *types* of transfer, such as transfer across temporal contexts (e.g., final test after a week; Coppens et al., 2011; Carpenter et al., 2008; Carpenter et

al., 2009), test formats (e.g., initial free-recall tests and final multiple-choice tests; Carpenter & DeLosh, 2006; Kang et al., 2007; McDaniel et al., 2007), and knowledge domains (e.g., previously studied but untested items; Jacoby et al., 2010; Kang et al., 2011; Chan et al., 2006; for a review see Carpenter, 2012). Other studies have also showed transfer to new related cues. For example, Carpenter (2011) showed that retrieval practice for the pair "mother-child" resulted in higher correct recall of the target "child" when cued with the new related cue "birth" than restudying the pair. In the current study, we examine transfer in a similar way: after studying and relearning Swahili words (e.g., mtoto) with one English definition (e.g., a very young child), participants are asked to retrieve the Swahili word (mtoto) in response to a new related cue (e.g., a recently born human) on the final test. Previous research has found that transfer is less likely to occur with term-definition pairs than with term-term pairs, presumably because definitions contain more words than single terms, which do not have prior strong associations to the target or a chunked memory representation (Pan & Rickard, 2017). This might not be the case in the current study, as the English definitions used are not new concepts to participants, and have representations of commonplace English words (i.e., are already chinked into a memory representation, such as "baby").

Novel related cues have not only been investigated as targets of transfer, but also as potential mechanisms of boosting transfer. For example, Butler (2010) examined whether relearning with different related cues can boost transfer to new inferential questions. During each relearning phase, participants in the testing condition were asked rephrased versions of each question. Surprisingly, Butler found that encoding variability did not boost transfer. However, that might have been due to rephrased questions not offering enough variability during learning. In fact, in a later study, Butler found that introducing variability using application questions that require different answers based on the same concept helped participants answer new application questions on the final test (Butler et al., 2017). Similarly, Goode et al. (2008) found that practice solving variations of an anagram resulted in better transfer than solving the same anagram several times. The current study similarly addresses the role of encoding variability on transfer of new knowledge, but instead using novel term-definition pairs.

A recent meta-analysis provides a mechanistic account for how retrieval practice might benefit transfer, referred to as the *three-factor framework for transfer of test-enhanced learning* (Pan & Ricard, 2018). According to the framework, for retrieval to facilitate transfer, three factors must be present: *response congruency*, *elaborated retrieval practice*, and *initial test performance*. *Response congruency* refers to the match between responses required on the initial and final tests. For example, if the cue of "a very young child" on the initial test requires a response "mtoto" and the cue of "a human offspring" on the final test requires the response "mtoto", response congruency is present, and transfer is more likely to occur. When response congruency is present, the response required on the final test is made more accessible by the initial test (via successful retrieval or feedback), which increases the probability of successful transfer.

The second factor is referred to as *elaborated retrieval practice* and can take two distinct forms: broad encoding methods and elaborative feedback. Broad encoding methods are present in initial tests that are formatted in such a way as to elicit processing of additional information related to the target. For example, an initial test may ask participants to report a detailed reason for their response or to attempt to recall everything they can that is related to the target. This type of retrieval results in greater cognitive processing, which can be driven by reactivation of memories formed during initial study, improved discrimination between answer choices (e.g., in a multiple-choice test), and/or better organization of the information in memory. *Elaborative feedback* refers to providing participants with detailed feedback on the initial tests – rather than simple corrective feedback. This can include providing participants with the reason a particular answer is correct, explaining the underlying concept of the correct answer, or allowing an opportunity to restudy all studied material. Providing elaborative feedback produces improved postretrieval restudy of the material, which can result in a more elaborated memory for the target.

The third factor – *initial test performance* – refers to the findings that higher performance on initial test during relearning predicts the magnitude of transfer. Successful retrieval during relearning is more likely to result in retrieval of aspects of the original study session (e.g., thoughts, inferences) other than the target itself than unsuccessful retrieval. These aspects, in turn, create a more elaborated memory

for the successfully retrieved targets and can serve as additional cues for successful retrieval on the transfer test. Participants with higher initial recall then are more likely to experience transfer for more items during the final test than participants with low initial recall.

This three-factor framework accounted for most of the findings in the literature on retrieval-facilitated transfer. Furthermore, while each factor alone increased probability of transfer, including all three factors had an additive effect on transfer. Thus, the framework can serve as a basis for evaluating transfer in different learning techniques. For example, although there is growing literature on the effects of spaced retrieval on transfer, no study to date has explored whether the dropout procedure promotes transfer. Additionally, as Pan and Rickard (2018b) pointed out, there are still types of materials and contexts that have not been investigated enough. One such context is novel vocabulary learning. Therefore, we aimed at filling that gap by exploring the effects of spaced retrieval on transfer in learning Swahili.

Current Study

The current study examined the benefit of spaced retrieval on memory for repetition and transfer items in novel vocabulary learning. We utilized an adapted version of the procedure used by Rawson et al. (2018). In the first Session, participants first engaged in *initial study*, during which they learned Swahili words with their English definitions. After studying all pairs, participants began the *relearning* block. In this block, a third of the participants were shown the definitions and asked to retrieve each target 5 times (standard condition), a third had to successfully recall each target once (dropout condition), and a third had to restudy the entire list 5 times (restudy condition). This number of list cycles was chosen based on a pilot study showing that it took participants an average of 35 minutes, which translates to 5 list repetitions, to successfully retrieve all words once. Participants then repeated their respective relearning tasks in Sessions 2 and 3, each spaced 48 hours after the completion of the previous session. In Session 4, which took place a week after Session 3, all participants took the same *final* test, where they had to recall the appropriate Swahili words when presented with either the same (e.g., “a very young child”; repetition items) or new (e.g., “a human offspring”; transfer items) English definitions. We chose Swahili – English-

definition paired associates as our materials because they mimic some instances of informal language acquisition in everyday life. Informal language acquisition refers to contexts outside of formal instruction in which people improve their knowledge of a language, such as reading or spending time with native speakers. This mode of learning has been shown to provide benefits to language acquisition (Gass, 1999).

Butler (2010) found that participants who were repeatedly tested on information from a learned passage experienced better recall for repetition and transfer items than participants who repeatedly restudied the passages. The first goal of the current study was to investigate whether the same effect would be found following testing of Swahili words using their English definitions relative to simply restudying these definitions. The second aim of this study was to explore whether encoding variability increased the effect of spaced retrieval on transfer. In Experiment 1 (fixed encoding) participants studied the same definition (e.g., "a very young child") associated with a single target (i.e., mtoto) across each successive session, whereas in Experiment 2 (variable encoding) participants learned a different definition (e.g., "a very young child", "a human offspring", "the product of a human reproductive process") each session. Prior research failing to show a benefit from variable encoding on transfer could reflect the lack of variability introduced by materials (i.e., slightly rephrased questions; Butler, 2010). In the current study, however, variable encoding (i.e., new definitions) orients attention to features that might be diagnostic for final transfer (i.e., new definitions). Thus, match between encoding and retrieval for transfer items might be stronger following variable compared to fixed encoding, resulting in greater transfer with variable encoding.

The third goal of the study was to compare whether standard and dropout procedures elicit the same amount of retention and transfer. As mentioned previously, the benefit of the dropout procedure on transfer has not yet been evaluated in the literature, and the lack of massed retrieval at each session makes it a candidate for a more beneficial technique in terms of time-efficiency.

Experiment 1

Experiment 1 tested whether spaced retrieval produced superior transfer to spaced restudying in novel vocabulary learning. According to the *three-factor framework of transfer of test-enhanced learning*,

the probability of transfer increases with the number of factors included during relearning. In the current study, all three of the conditions experience response congruency, as the answers required on the transfer test are the same words that served as targets at relearning. All conditions also involve the same degree of elaborative feedback, as participants in the retrieval conditions receive corrective feedback with the correct response and the restudy condition rereads the pair, all without any additional details or elaboration. However, spaced retrieval might involve a higher degree of broad encoding methods than restudy does, as retrieval might result in greater cognitive processing than simply reading the material. Additionally, participants in the spaced retrieval conditions are expected to have a higher degree of learning after the relearning sessions than restudy condition, due to the classic testing effect. Thus, a finding of greater transfer in spaced retrieval conditions relative to the restudy condition would indicate that transfer is driven by the difference in the cognitive processing that occurs at relearning.

We are also interested in comparing the benefit of the two spaced retrieval procedures on transfer. There were two reasons for including both retrieval conditions. The first was for practical reasons, as the standard condition is more directly comparable to the restudy condition because the relearning phase includes five list-learning cycles in each. When considering the dropout method, the number of exposures varies for each item, and so choosing an appropriate restudy comparison group is difficult. The second reason was for application to real-world study contexts. Despite differences in exposure between the two testing conditions, presumably the dropout method may be more efficient in terms of time. Thus, if recall is identical between the two conditions, yet the dropout method is faster, this would suggest that perhaps a dropout method is the ideal method for learning. No prior research has examined whether the dropout procedure promotes transfer, and no studies directly compared the two procedures, so we did not form predictions regarding the benefit of one over the other.

Method

Participants

An a priori power analysis using GPower 3.1 ($\alpha = .05$, Power = .80) showed that a total of 150 participants (50 per condition) would be required to detect a medium effect size for the Condition x Item

Type interaction. However, due to the pandemic forcing the study online, we experienced low rates of signups for participation and high rates of attrition. As this study was a part of a Master's Thesis, we decided to collect only 30 participants per condition (resulting in total of 90 participants in the study). Participants (aged 17 – 47; M = 20.30, SE = 5.17) were undergraduate students from the University of Texas at Arlington recruited through the SONA systems for class credit. After accounting for exclusionary criteria (detailed below), the final sample consisted of 95 participants (restudy: 33; standard: 29; dropout: 33).

Design and Materials

A 3 (*condition: restudy, standard, dropout; between-subjects*) x 2 (*item type: repetition, transfer*; *within-subjects*) mixed design ANOVA was used to assess memory. Materials included 30 Swahili words (e.g., mtoto) with their English definitions (e.g., a very young child). The English-Swahili word pairs for the stimuli were obtained from Nelson and Dunlosky (1994) and Carpenter et al. (2008). For each word, we created four converging definitions using online dictionaries. Four definitions were used in anticipation of using variable encoding during Experiment 2, which requires a total of four definitions for each Swahili word.

A pilot study (N = 197) was conducted with 60 words to test how well the created definitions evoked the correct word. We excluded 12 Swahili words that did not have at least four definitions averaging at least 75% accuracy. Then, we selected 30 words for which the definitions, on average, elicited the correct response most frequently. For those words, the four of five definitions with the highest average recall were selected for the final stimulus materials. Thus, for the final 30 selected Swahili words, each had four converging definitions. For example, for the Swahili word "mtoto", the four definitions would be (1) a very young child, (2) the product of a human reproductive process, (3) a recently born human, and (4) human offspring.

To counterbalance the procedure, we created two versions of the final test. In one version, half of the definitions were previously studied by participants (repetition items) and half were new definitions

(transfer items). In the second version, repetition items from the first version appeared as transfer items and transfer items from the first version appeared as definition items.

All data collection was completed online using PsychoPy3 software. QuestionPro survey software obtained participant consent and randomly assigned participants to each condition. The PsychoPy3 experiment was hosted on Pavlovia.com.

Procedure

Initial Learning (Session 1). During Session 1, all participants engaged in two phases. Phase 1 consisted of learning and Phase 2 consisted of relearning. Before beginning Session 1, participants filled out a demographic questionnaire. After the questionnaire, participants were shown task instructions for both phases. Participants were told that they would be studying Swahili words with their English definitions.

Participants in the restudy condition were told that after the initial learning, they would be shown each Swahili word with its English definition for 10 seconds for restudy. They were instructed to press the spacebar once they feel they have sufficiently studied the pair, after which they were be presented with the pair for 4 more seconds (to equate presentation time between conditions). Instructions included examples of what the participants would see and what they were to do in in this session.

Participants in the retrieval conditions were told that after initial study, they would be shown the definitions and given up to 10 seconds to retrieve the corresponding Swahili word. If they retrieve the word before that time is up, they were instructed to press the ENTER key. Once that key was pressed, or the time has run out, participants were presented with the definition, the correct answer, and colored text to indicate whether their answer was correct (green) or incorrect (red). Before moving on to the task itself, participants engaged in a practice trial corresponding to their condition.

Phase 1 (Learning)**.** During Phase 1, all participants studied the 30 words and their definitions. Each word-definition pair was presented for 10 seconds for study (DeWinstanley & Bjork, 2004). Then participants engaged in Phase 2.

Phase 2 (Relearning). During Phase 2, participants in the *restudy* condition were shown the same word-definition pairs from Phase 1, in random order. They were told to study each pair for up to 10 seconds and press the space bar once they felt that they had learned the word. After 10 seconds had passed or participants had pressed the space bar, they were presented with each word again for 4 seconds for further study. This was done to equate the length of presentation of each stimulus between the conditions (Pan & Rickard, 2017). Participants restudied the entire list for 5 cycles or 35 minutes, whichever came first, so as to equate the length of relearning to the other conditions.

Participants in the *standard* condition engaged in repeated testing for Phase 2. During relearning, participants were presented with the definitions, one at a time and in a random order (Rawson et al., 2018), and given up to 10 seconds to type out the corresponding Swahili words. After each retrieval attempt, regardless of success, participants were presented with the correct word and its definition for 4 seconds for restudy. They were tested on entire list 5 times, so as to equate the length of relearning to the other conditions.

Participants in the *dropout* condition engaged in retrieval. During retrieval, they were presented with the definitions, one at a time and in a random order (Rawson et al., 2018), and given up to 10 seconds to type out the corresponding Swahili words (Rawson et al., 2018). On both correct and incorrect trials, participants were presented with the correct answer for 4 seconds for restudy (Pan & Rickard, 2017). If participants responded incorrectly or omitted a response, the items in those trials were then placed at the end of the stack for another recall trial. Participants had up to 35 minutes to recall all of the words. This was done to equate the total amount of exposure time between the conditions. Pilot studies showed that 35 minutes, which roughly translates to 5 cycles, was on average enough time for participants to recall all words correctly.

Relearning (Sessions 2 and 3). Sessions 2 and 3 were relearning sessions. Session 2 was administered 2 days after initial learning and Session 3 was administered 2 days after Session 2. In each of these sessions, participants in the retrieval conditions were told that they would be tested on their memory for the Swahili words from the previous session, and the participants in the restudy condition

were told that they would be restudying the pairs from the previous session. The rest of the instructions were the same as the ones in Phase 2 of the Initial Learning session. Just as in Session 1, participants engaged in a practice trial corresponding to their condition before completing the task itself. During these sessions, participants engaged in the same relearning procedure as the one they completed in Phase 2 of the initial study session using the same definitions they studied. That is, participants in the dropout condition engaged in the same cued-recall test with feedback for up to 35 minutes as they did at initial study, participants in the standard condition were tested on the entire list 5 times, and participants in the restudy condition restudied all materials for 35 minutes/5 cycles.

Final Test (Session 4). A week after Session 3 (Rawson et al., 2018), all participants took the same final test, in which half of the items were tested using previously seen definition cues (*repetition*; e.g., a very young child) and half were tested using new definition cues (*transfer*; e.g., the product of a reproductive process). Participants were told that they would be shown English definitions and that they had unlimited time to retrieve the corresponding Swahili words. The instructions stated that some of the definitions would be from previous sessions, and some would be new English definitions corresponding to previously studied Swahili words. Participants in the retrieval conditions were told that the test would be in a similar format to tests from relearning, but that each word would only be tested once, and no feedback would be provided.

Post-Session Questionnaire. At the end of each session, participants filled out a questionnaire asking them if they a) had used any outside aid (such as an online dictionary), b) written any words or definitions down, and c) to provide any other information that might have affected their performance.

Attention Checks. For all experiments, each session included an attention check interspersed with other questions that asked for demographic information. The demographic questions were added so that participants did not catch on to attention checks and only paid attention at the end of the experiment. The answers to these questions were obvious, and the attention checks were created in order to exclude data of participants who were not paying attention during the tasks.

Data Scoring

For all experiments, recall was hand-scored by the experimenter. As the target words that participants had to type in were in Swahili and might have been challenging to spell, a separate strict and lenient scoring systems were utilized. For the strict scoring system, a word was only counted as correct if it exactly matched the correct Swahili word. For the lenient scoring system, a word was counted as correct if the word a) was only misspelled by one letter (e.g., mtot*a*); b) was missing one letter (e.g., mtot); c) had two consecutive letters that were switched (e.g., m*ot*to); d) had one extra letter (e.g., m*a*toto); or e) was inputted correctly, but because of a technical issue was scored as incorrect. All analyses were conducted using each scoring system separately. However, as they produced a very similar pattern of results, only the lenient scoring results are presented below.

Exclusionary Criteria

Exclusionary criteria were preregistered. The exclusionary criteria were as follows, with the number of subjects fitting each criteria noted in parentheses: a) native language was not English or had previous knowledge of Swahili, as determined by responses in the Demographics questionnaire at the beginning of the study (n = 10); b) failed 75% (3 out of 4) attention checks (n = 0); and c) reported writing words or definitions down, using outside aid, or otherwise cheating (n = 7). Additionally, three participants were excluded due to system error with the experiment. This criterion was not preregistered or foreseen, but the experimental error compromised the data obtained from these participants.

We originally preregistered that we would exclude participants in the restudy condition that did not press space at least once during a single relearning session (as an index of "accuracy") or for participants whose average study time was less than 1 second. However, many participants never pressed the spacebar, while others had brief durations. In hindsight, we realized that is possible that 10 seconds was not sufficient time for some participants to feel that they have learned the words, while for others 1 second may have been enough (since after terminating study the information was presented again for 4 seconds) these. We therefore analyzed the data with and without these participants. As there was no difference in the results of the two analyses, we report analyses including all of these participants.

Attrition

Attrition rates were high for this experiment (52% in total). In the *restudy* condition, 75 participants started the study, and 39 participants completed all sessions. In the *dropout* condition, 76 participants started the study, and 41 participants completed all sessions. In the *standard* condition, 67 participants started the study, and 34 participants completed all sessions. Importantly, a chi-square test revealed that there were no differences in attrition rates between the 3 conditions, [$X^2(4, N = 221) = 2.26$, $p = .69$]. We are not sure of the reason for such high attrition rates, but it is possible that students struggled to participate in four online experimental sessions that were spread out over two weeks.

Results

Data from 95 participants are reported below. We analyzed final recall using strict and lenient scoring schemes. All results from both analyses were significant at an alpha level of .05. Therefore, analyses shown below have been conducted using the lenient scoring scheme.

Final Test Performance

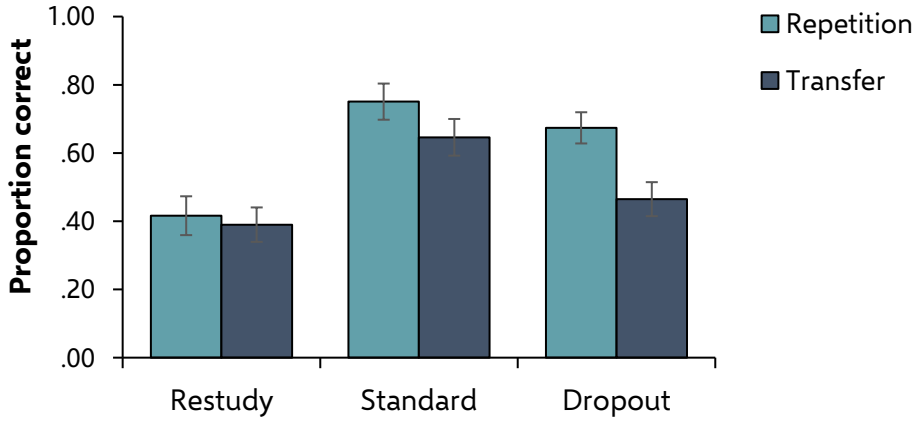
Figure 1 displays means and standard errors for retention and transfer in the dropout, standard, and restudy conditions. A 3 (condition: restudy, standard, dropout; *between-subjects*) X 2 (item type: repetition, transfer; *within-subjects*) ANOVA revealed that participants recalled more repetition than transfer words on the final test [*Item Type:* $F(1, 92) = 30.74$, $p < .001$, $\eta_p^2 = .25$]. There was also a significant main effect of *Condition* [$F(2, 92) = 9.11$, $p < .001$, $\eta_p^2 = .17$]. These main effects were qualified by a significant interaction [*Condition x Item Type:* $F(2, 92) = 7.02$ $p = .001$, $\eta_p^2 = .13$].

To probe the interaction, we conducted separate one-way ANOVAs for repetition and transfer items. For repetition items, there was a significant main effect of condition [*Condition:* $F(2, 92) = 11.32$, $p < .001$, $\eta_p^2 = .20$]. Participants in the restudy condition (M = .42, SE = .05) recalled significantly fewer words than participants in the dropout condition (M = .67, SE = .05) and those in the standard condition (M = .75, SE = .05). There was no difference in recall between the dropout and standard conditions. For transfer items, there was a main effect of condition [*Condition:* $F(2, 92) = 6.33$, $p = .003$, $\eta_p^2 = .12$]. Participants in the restudy condition (M = .39, SE = .05) recalled significantly fewer words than

participants in the standard condition (M = .65, SE = .05), but not significantly fewer than those in the dropout condition (M = .47, SE = .05). Additionally, participants in the dropout condition recalled significantly fewer transfer items than participants in the standard condition.

Figure 1

Retention and Transfer in Experiment 1



Note. Spaced retrieval produced better recall for repetition items than restudy, but only the standard condition produced better transfer than the restudy condition. Bars represent standard error of the mean.

Rate of Learning

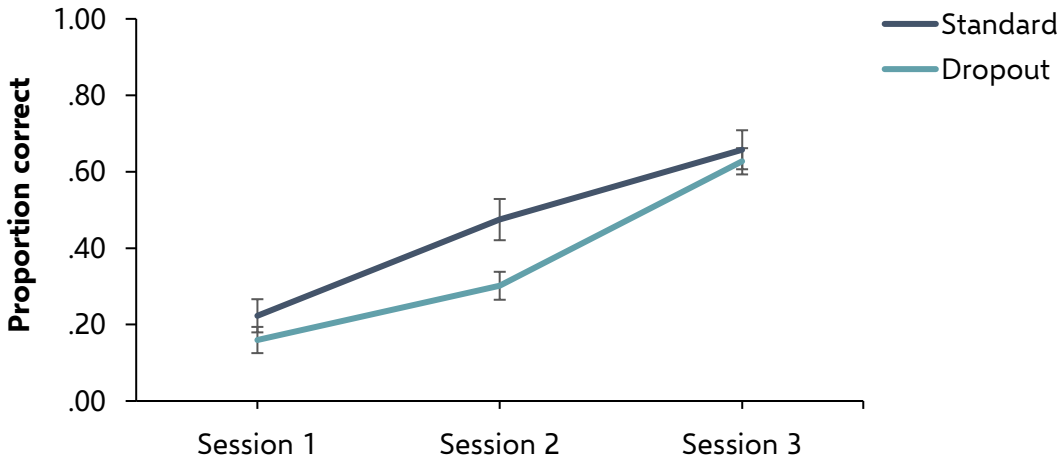
Figure 2 displays means and standard errors for recall accuracy during each study session in the dropout and standard conditions. We conducted a preregistered analysis of rate of learning for the two standard conditions to examine whether different relearning techniques yielded differences in how quickly participants acquired material during learning. To do that, we calculated the proportion of words (out of 30) participants in each condition recalled the first list cycle they were tested on in each session[1]. A 2 (condition: standard, dropout; between-subjects) x 3 (session: 1, 2, 3; within-subjects) mixed-effects ANOVA revealed that number of items recalled on the first test each session increased across sessions [*Session: F*(2, 120) = 108.57, $p < .001$, $\eta_p^2 = .64$]. Recall did not differ between the two conditions

[1] Note that this comparison cannot be done for subsequent cycles (e.g., cycle 5) because recalled items are not included in later cycles in the dropout condition.

[*Condition:* $F(1, 60) = 3.42$, $p = .07$, $\eta_p^2 = .05$] and there was no interaction between the two, [*Session x Condition:* $F(2, 120) = 2.97$, $p = .055$, $\eta_p^2 = .05$].

Figure 2

Relearning Accuracy Across Days in Experiment 1



Note. Overall, there was no difference in learning rates between the dropout and standard conditions. Bars represent standard error of the mean.

We also conducted exploratory (but pre-registered) correlational analysis between accuracy in the first recall cycle of each relearning session and final recall of repetition and transfer items, collapsed across the two retrieval conditions. For final recall of repetition items, there was no correlation with Session 1 performance, but there were sizable positive correlations with Session 2 and 3 performance, [*Session 1:* $r(60) = .12$, $p = .36$; *Session 2:* $r(60) = .50$, $p < .001$; *Session 3:* $r(60) = .75$, $p < .001$. The same pattern appeared for correlations between final recall for transfer items and the three sessions, [*Session 1:* $r(60) = .10$, $p = .44$; *Session 2:* $r(60) = .48$, $p < .001$; *Session 3:* $r(60) = .75$, $p < .001$]. Additionally, there was a strong positive correlation between final recall for repetition and transfer items [$r(60) = .75$, $p < .001$].

Total Relearning Time

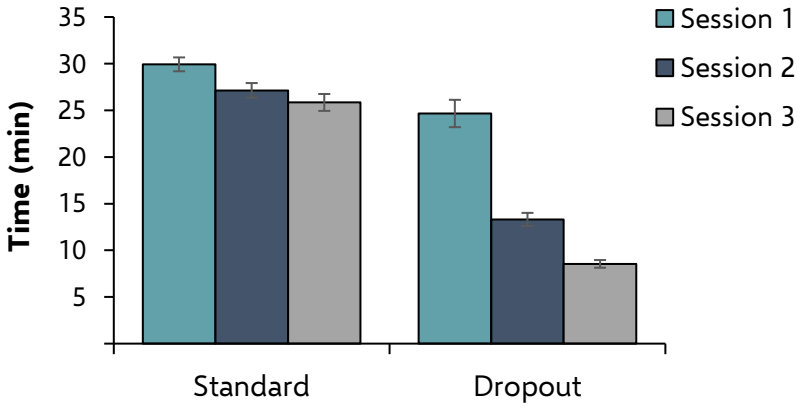
Differences in the duration of relearning can be an important consideration in research and learning environments, where time is a limited resource. We therefore conducted exploratory analyses to

examine the time spent in relearning each session in the standard and dropout retrieval conditions. Relearning time reflects the relearning duration in each session and does not include time spent on instructions, practice, or post-session questionnaires.

Figure 3 displays means and standard errors for time spent on relearning in each session in the standard and dropout conditions. A 2 (condition: standard, dropout; between-subjects) x 3 (session: 1, 2, 3; within-subjects) mixed-effects ANOVA on relearning time revealed that participants in the standard condition spent more time relearning overall [*Condition:* $F(1, 60) = 139.74$, $p < .001$, $\eta_p^2 = .70$]. There was also a significant main effect of session, [*Session:* $F(2, 120) = 120.29$, $p < .001$, $\eta_p^2 = .67$], with relearning duration decreasing each session. These main effects were qualified by a significant interaction, [*Condition x Session:* $F(2, 120) = 42.95$, $p < .001$, $\eta_p^2 = .42$]. This interaction reflects that the dropout procedure was faster in each session, but this difference increased across sessions.

Figure 3

Relearning time Experiment 1



Note. Participants in the dropout condition spent significantly less time training than participants in the standard condition. Bars represent Standard error of the mean.

Discussion

The purpose of Experiment 1 was to examine whether spaced retrieval produced superior retention and transfer relative to restudying in novel vocabulary learning. A second aim was to compare the two spaced retrieval conditions in terms of retention and relearning time, as this comparison has not been conducted previously (but see Karpicke & Roediger, 2007 for comparison of different single-session criterion levels with dropout). We found that both retrieval procedures improved recall of repetition items over restudy. In contrast, only the standard procedure improved memory for transfer items over restudy. To examine why standard, but not dropout procedure of relearning promotes transfer, we refer to the three-factor framework for transfer of test-enhanced learning. Any differences in the levels of factors between the two testing procedures could point to the reason for these results.

The first factor, response congruency, was equated between the two retrieval conditions, as both conditions were required to produce the same responses on the transfer test as they did during testing. The factor of elaborative feedback was also equated between the two conditions: both received non-elaborative feedback at relearning. The results of our analysis of learning rates showed that there were no differences in overall recall between the two conditions.

This leaves broad encoding methods as a possible imbalanced factor driving the difference in transfer between the two conditions. Broad encoding methods are present in relearning procedures that orient participants to engage in deeper cognitive processes of the material relatively to simply restudying. One possibility is that the standard, but not the dropout, procedure induced deeper cognitive processing in our experiment by repeatedly retrieving information in each session. Our dropout condition only had to retrieve each word correctly once, limiting the number of retrieval attempts for some words. Experiment 2 was designed to examine whether increased levels of broad encoding through encoding variability might reduce transfer differences between the two retrieval conditions and improve transfer for the dropout condition relative to the restudy condition.

Finally, our analysis of the relearning time in the two retrieval conditions in each session revealed that participants in the standard condition spent more time on learning overall. Thus, the improved

memory for repetition and transfer items in the standard conditions comes at a cost in terms of time taken to complete the relearning.

Experiment 2

The purpose of Experiment 2 was to investigate whether encoding variability would promote transfer in novel vocabulary learning. To do that, we used the same procedure as in Experiment 1, with the exception that participants were presented with different English definitions of the same Swahili words in each relearning session. On the final test, half of the words were tested with definitions from the last relearning session (repetition items) and half were tested with new, never-before-seen definitions (transfer items).

One possible way to promote cognitive processing is encoding variability. Introducing variability at encoding has been shown to improve learning by creating more retrieval routes (Estes, 1955; Martin, 1968; Kukull et al., 2002). It has also been shown to promote transfer to novel tasks (e.g., Wahlheim et al., 2012; Barcroft & Sommers, 2005). Some studies have found the benefit of variable encoding on transfer following retrieval practice (Goode et al., 2008; Smith & Handy, 2014, 2016; Butler et al., 2017). However, this effect is only found when learning stimuli introduce a sufficient degree of variability (Butler, 2010). In the current experiment, we explored whether practicing with different definitions of the same Swahili word would introduce enough variability to promote transfer and whether encoding variability would induce greater cognitive processing in the dropout condition.

We expected that varied definitions will benefit recall of transfer items more than repetition items. Transfer is dependent on elaborative encoding of learned material (i.e., broad encoding methods), which encoding variability supports. Retrieval of learned information without application to a new task, while benefitting from elaborative encoding, does not necessarily depend on it to the same degree. If so, we should find that variable learning benefits transfer items more than repetition items.

We also hypothesized that varied definitions would benefit transfer more in the spaced retrieval conditions relative to the restudy condition. The reason for this prediction is that we expected the greater cognitive processing introduced by varied definitions to have an additive effect with the already present

greater cognitive processing in the spaced retrieval (versus restudy) conditions. According to the elaborative retrieval hypothesis, testing results in a more extensive memory search than restudying, which creates a more elaborative memory trace (Carpenter & DeLosh, 2006). Encoding variability (i.e., learning with different definitions) improves memory by creating more retrieval routes. Therefore, testing and variability improve learning by two different processes which might have an independent boosting effect on items. Alternatively, it might be that testing and elaborative encoding both enhance cognitive processing through the same mechanisms of greater elaboration and strengthening of memory traces. If that is the case, the restudy condition, which engages in less cognitive processing than the testing conditions do, might see more benefit from encoding variability.

The second aim of this experiment was to examine whether encoding variability would have a different effect on the two spaced retrieval conditions. Results of Experiment 1 showed that the dropout method did not show improved memory for transfer memory. Assuming that the testing condition is already engaging in broad encoding following multiple retrievals, we would expect that encoding variability would boost transfer more in the dropout than the restudy condition. We note, however, that this is a post-hoc hypothesis.

Method

Participants

An a priori power analysis using GPower 3.1 ($\alpha = .05$, Power = .80) showed that a total of 150 participants (50 per condition) would be required to detect a medium effect size. However, due to the pandemic forcing the study online, we have experienced low rates of signups for participation and high rates of attrition. We decided to collect only 30 participants per condition (resulting in total of 90 participants in the study). Participants (aged 17 – 45; $M = 19.48$, $SE = 3.93$) were undergraduate students from the University of Texas at Arlington recruited through the SONA systems for class credit. After accounting for exclusionary criteria (detailed below), the final sample consisted of 95 participants.

Design

A 3 (condition: restudy, standard, dropout; between-subjects) X 2 (item type: repetition, transfer; within-subjects) mixed design ANOVA was used to analyze the proportion of words correctly recalled on the criterion test.

Materials and Procedure

The materials and procedure were the same as in Experiment 1, with one exception. During each session, participants were presented with a different English definition for the same Swahili word. The instructions for each session were similar to those in Experiment1, but in Sessions 2 and 3 participants were told that all of the definitions would be new English definitions corresponding to the Swahili words from the previous session.

To counterbalance the procedure, we created two versions of the materials. In one version, half of the words on the final test were tested with the definitions from Session 3, and half were tested with new, never before seen definitions. In the second version, the former words were tested with new definitions and the latter words were tested with definitions from Session 3.

Exclusionary Criteria.

Exclusionary criteria were preregistered. The exclusionary criteria were as follows, with the number of subjects fitting each criterion noted in parentheses: a) native language was not English or had previous knowledge of Swahili, as determined by responses in the Demographics questionnaire at the beginning of the study (n = 5); and b) reported writing words or definitions down, using outside aid, or otherwise cheating (n = 2).

Like in Experiment 1, we did not exclude participants that scored 0% accuracy on any single relearning session and participants whose study duration was less than 1 second in the restudy condition. Analyses with and without those participants showed no difference in the results.

Attrition

Attrition rates were high for this experiment (50% in total). In the *restudy* condition, 75 participants started the study, and 37 participants completed all sessions. In the *standard* condition, 60

participants started the study, and 33 participants completed all sessions. In the *dropout* condition, 71 participants started the study, and 33 participants completed all sessions. A chi-squared test of independence showed there were no differences in the attrition rates between conditions [$X^2(4, N = 207) = 3.17, p = .53$].

Results

Data from 95 participants are reported below. We analyzed final recall using strict and lenient scoring schemes, which produced the same results. We therefore report analyses using the lenient scoring scheme.

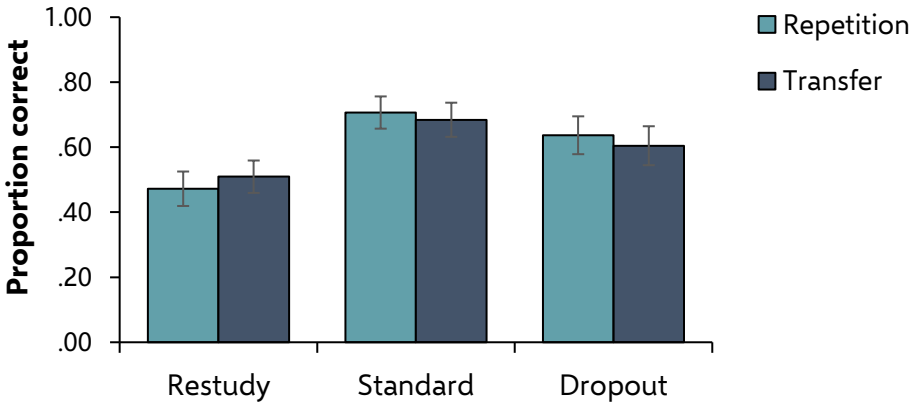
Final Test Performance

Figure 5 displays means and standard errors for retention and transfer in the dropout, standard, and restudy conditions. A 3 (condition: restudy, standard, dropout; between-subjects) X 2 (item type: repetition, transfer; within-subjects) ANOVA revealed no effect of test, [$F < 1$] and no *Item Type x Condition* interaction [$F(2, 92) = 1.94, p = .15, \eta_p^2 = .04$]. However, there was a significant effect of *Condition* [$F(2, 92) = 4.17, p = .02, \eta_p^2 = .08$]. The condition effect reflects that participants in the restudy condition (M = .49, SE = .05) recalled fewer words than participants in the standard condition (M = .70, SE = .05), but not significantly so than those in the dropout condition (M = .62, SE = .05). There was no difference in recall between the dropout and standard conditions.

Although there was no test x condition interaction, we conducted preregistered analyses for recall differences between conditions separately for repetition and transfer items. As can be seen in Figure 4, for repetition items, participants in the restudy condition recalled significantly fewer words than participants in the standard condition, [$F(1, 64) = 10.14, p = .002, \eta_p^2 = .14$], and dropout condition, [$F(1, 63) = 4.35, p = .04, \eta_p^2 = .07$]. Recall in the standard condition did not differ significantly from the dropout condition [$F < 1$]. For transfer items, participants in the restudy condition recalled significantly fewer words than participants in the standard condition, [$F(1, 64) = 5.81, p = .02, \eta_p^2 = .08$], but not the dropout condition, [$F(1, 63) = 1.52, p = .22, \eta_p^2 = .02$]. Recall in the standard condition did not differ significantly from the dropout condition, [$F(1, 57) = 1.01, p = .32, \eta_p^2 = .02$].

Figure 4

Retention and Transfer in Experiment 2



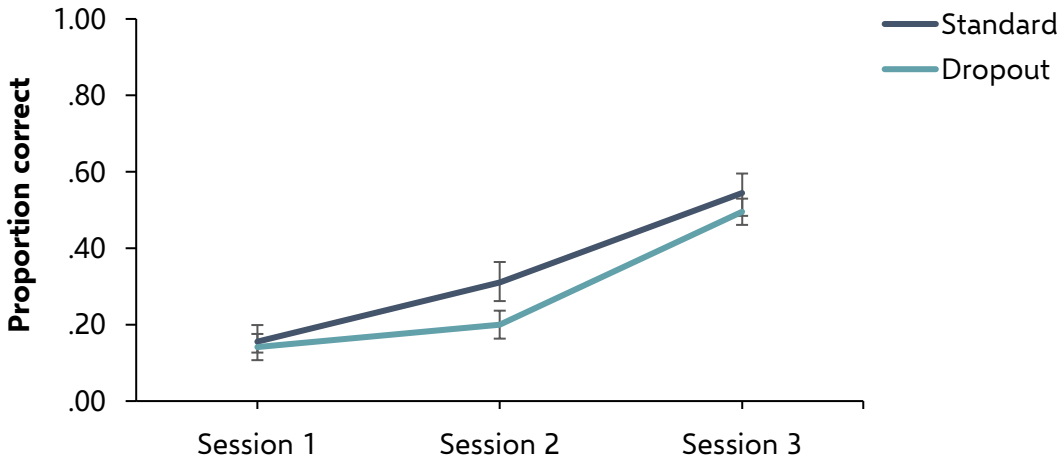
Note. Standard condition produced better recall for repetition, but not transfer items than the restudy condition. Dropout condition did not produce better recall of repetition or transfer items relative to restudy. Bars represent standard error of the mean.

Rate of Learning

Figure 5 displays means and standard errors for recall accuracy in each session in the dropout and standard conditions. A 2 (condition: standard, dropout; between-subjects) x 3 (session: 1, 2, 3; within-subjects) mixed-effects ANOVA revealed that recall increased with each new session, [*Session: F*(2, 114) = 50.20, $p < .001$, $\eta_p^2 = .47$]. There was no effect of condition, [*Condition: F*(1, 57) = 1.70, $p = .20$, $\eta_p^2 = .03$] and no significant interaction between the two, [*Session x Condition: F* < 1].

Figure 5

Learning Accuracy Across Days in Experiment 2



Note. Overall, there were no differences in the rate of learning between the dropout and standard conditions. Bars represent standard error of the mean.

We also conducted correlational analysis between accuracy in the first recall cycle of each relearning session and final recall of repetition and transfer items, collapsed across retrieval conditions. As with Experiment 1, this analysis showed no correlation of final recall for repetition items with initial performance in session 1, but there was a strong positive correlation in sessions 2 and 3, [*Session 1: r*(56) = .06, $p = .67$; *Session 2: r*(56) = .68, $p < .001$; *Session 3: r*(56) = .82, $p < .001$]. The same pattern appeared for correlations between final recall for transfer items and the three sessions, [*Session 1: r*(56) = -.07, $p = .63$; *Session 2: r*(56) = .72, $p < .001$; *Session 3: r*(56) = .89, $p < .001$]. Additionally, there was a strong positive correlation between final recall for repetition and transfer items, [$r(56) = .87$, $p < .001$].

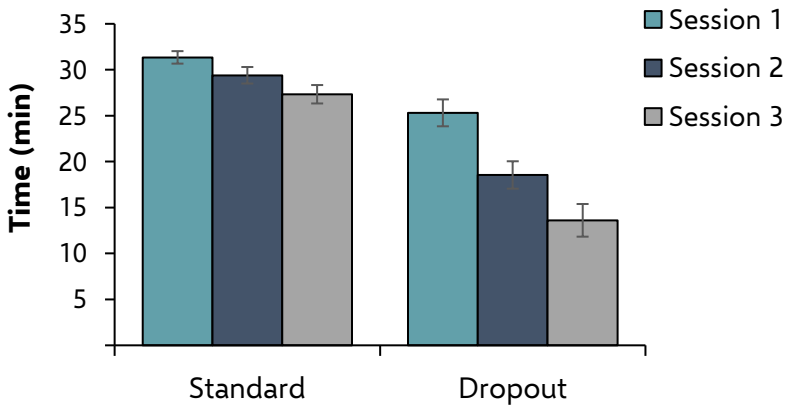
Total Relearning Time.

Figure 6 displays means and standard errors for time spent on relearning in each session in the dropout and standard conditions A 2 (condition: standard, dropout; between-subjects) x 3 (session: 1, 2, 3; within-subjects) mixed-effects ANOVA on relearning time revealed that participants in the standard condition spent more time relearning overall [*Condition: F*(1, 56) = 41.03, $p < .001$, $\eta_p^2 = .42$]. There was also a significant main effect of session, [*Session: F*(2, 112) = 50.15, $p < .001$, $\eta_p^2 = .47$]. These main

effects were qualified by a significant interaction, [*Condition x Session:* $F(2, 112) = 11.60$, $p < .001$, $\eta_p^2 = .17$]. This interaction reflects that the dropout procedure was faster in each session, but this difference increased across sessions.

Figure 6

Relearning time Experiment 2



Note. Participants in the dropout condition spent significantly less time training than participants in the standard condition. Bars represent Standard error of the mean.

Discussion

Experiment 2 was designed to examine whether learning novel vocabulary using varied definitions promotes greater transfer relative to recall of repetition items and to a greater degree in spaced retrieval relative to repeated restudy conditions. We were also interested in whether this encoding variability had a different effect on the two spaced retrieval conditions. The results are largely consistent with Experiment 1, such that overall, the standard condition produced the best recall, the restudy condition had the worst recall, and the dropout condition fell somewhere in between. When examining item types separately as a function of condition, the results showed that the standard and dropout procedures produced superior recall of repetition items compared to the restudy condition. In contrast, only the standard condition showed superior transfer over the restudy condition (although the dropout condition was numerically greater than the restudy condition and did not differ from the standard

condition). We hypothesized that studying with different definitions would promote greater cognitive processing and accentuate the differences between the retrieval and restudy conditions for both repetition and transfer items. However, it appears that introducing variability had relatively little influence on performance, other than producing a numerical increase in transfer for the dropout condition. However, as encoding variability was manipulated across different experiments, it is not possible to directly explore this idea. In the following section we therefore conducted preregistered exploratory cross-experimental analyses.

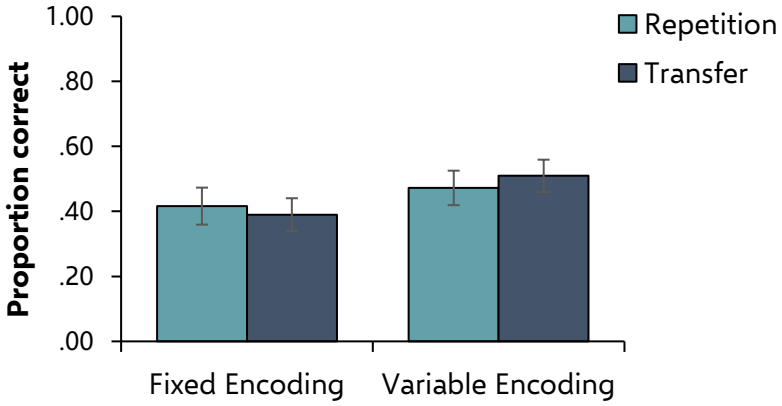
Cross-Experimental Analyses

To better understand how variable encoding affects retention and transfer, we conducted an analysis including data from both experiments. Direct comparison of these two encoding manipulations can provide a better insight into the mechanisms underlying retention and transfer in the three conditions. Therefore, we conducted separate 2 (item type: repetition, transfer; within-subjects) x 2 (encoding: fixed, variable) ANOVAs for each condition. Figure 7 displays means and standard errors for retention and transfer in the restudy (A), standard (B), and dropout conditions (C).

Analysis for the restudy condition revealed no main effect of *Item Type*, [$F < 1$], or *Encoding*, [$F(1, 67) = 1.51$, $p = .22$, $\eta_p^2 = .02$], and there was no interaction between the two, [*Item Type x Encoding:* $F(1, 67) = 2.16$, $p = .15$, $\eta_p^2 = .03$]. Thus, encoding variability did not influence memory.

Figure 7 (A)

Restudy Condition

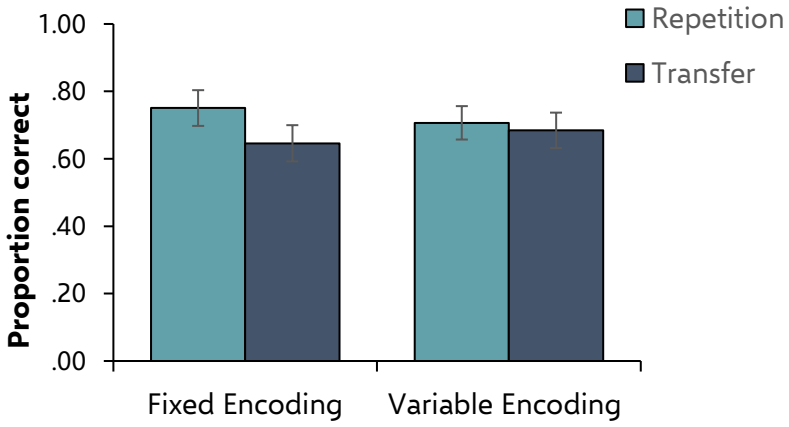


Note. Encoding variability had no effect on retention and transfer in the restudy condition. Bars represent standard error of the mean.

Analysis for the standard condition revealed that participants recalled more repetition items [*Item Type:* $F(1, 57) = 7.71$, $p = .01$, $\eta_p^2 = .12$]. Recall did not differ across encoding conditions [*Encoding:* $F < 1$] and there was no interaction between the two [*Item Type x Encoding:* $F(1, 57) = 3.25$, $p = .08$, $\eta_p^2 = .05$]. Thus, encoding variability did not influence memory.

Figure 7 (B)

Standard Condition

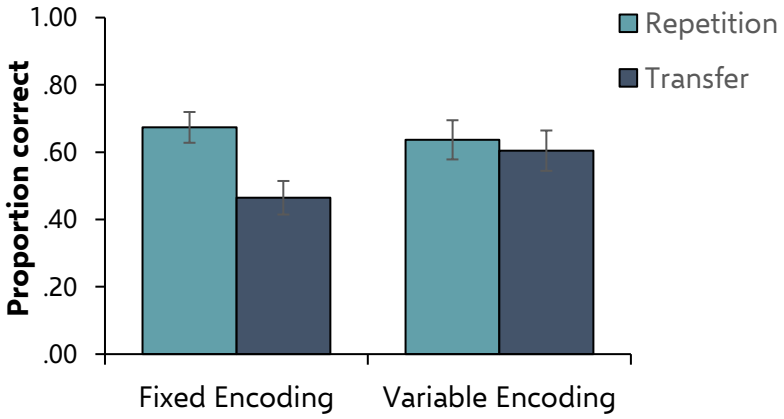


Note. Encoding variability had no effect on retention and transfer in the standard condition. Bars represent standard error of the mean.

Analysis for the dropout condition revealed that participants recalled more repetition than transfer items, [*Item Type: F*(1, 60) = 27.45, $p < .001$, $\eta_p^2 = .31$]. Although there was no effect of encoding, [*Encoding: F* < 1], there was a significant interaction between the two factors, [*Item Type x Encoding: F*(1, 60) = 14.76, $p < .001$, $\eta_p^2 = .20$]. This reflects that variable encoding did not affect recall of repetition items [*F* < 1], but numerically (but not significantly) improved recall for transfer items [*F*(1, 60) = 3.27, $p = .075$, $\eta_p^2 = .05$]. Thus, encoding variability does numerically increase transfer.

Figure 7 (C)

Dropout Condition



Note. Encoding variability improved transfer but not retention in the dropout condition. Bars represent standard error of the mean.

The results of our condition-level analyses showed that variable encoding numerically promotes transfer in dropout, but not standard or restudy conditions. This finding is in line with our prediction that variable encoding promotes greater cognitive processing (i.e., broad encoding methods) in the dropout condition. It is possible that the manipulation did not have the same effect in the standard condition because retrieving items repeatedly in a single session might have already forced participants to engage in greater cognitive processing. For the restudy condition, it seems that encoding variability failed to induce greater cognitive processing, as evident by the lack of the effect of encoding.

General Discussion

In this study we explored whether spaced retrieval promotes learning of Swahili words from their English definitions compared to repeated restudy and whether this transfers to improved memory for vocabulary using never-before seen definitions. Learning novel vocabulary is a universal human experience: from toddlers learning their mother tongue, to schoolchildren acquiring more complex vocabulary, to anyone trying to master a new language. Much of language acquisition occurs in informal settings, such as from conversing with those who already have a mastery of the language or from reading books. To simulate such learning, we used Swahili words with the English definitions as our stimuli.

Although there is previous research on the benefits of retrieval practice for retention and transfer, no studies have examined these processes with such materials. We found that the standard spaced retrieval procedure resulted in higher recall of words cued with the same (repetition) and new (transfer) definitions than studying by rereading the material regardless of whether the same (fixed) or different (variable) definitions were studied across relearning sessions. Additionally, we found that although more efficient in terms of relearning duration, the dropout procedure of spaced retrieval produces less transfer than the standard procedure. Below we discuss the theoretical and applied ramifications of these findings.

Memorial Benefit of Spaced Retrieval on Repetition Items

The first aim of this study was to investigate the effect of spaced retrieval and repeated restudy on retention of repeated definitions and terms during new language acquisition. In line with our predictions, both Experiment 1 and 2 showed that both spaced retrieval procedures promoted retention over restudying. The benefit of retrieval over restudy was not surprising. According to the elaborative retrieval hypothesis, when attempting to retrieve information, participants engage in a memory search that reactivates information that is relevant to the target information, creating a more elaborative memory trace for the target. At test, some of this relevant information can be used to cue the target information, making retrieval more likely. Restudying information, on the other hand, does not result in a memory search, and thus does not produce an equally elaborated memory trace for target information that can be helpful at test (Carpenter & DeLosh, 2006). Additionally, the desirable difficulty hypothesis poses that conditions that make initial learning more difficult result in better retention of information (Bjork, 1994, 1999). Retrieval practice is more difficult than restudying, because it forces participants to engage in memory search and recall the target, thereby improving retention. Finally, the mediator effectiveness hypothesis states that in cued-recall paradigms, such as the one used in this study, participants generate mediators to link the cue to the target at study. During retrieval, the mediator can be activated by the presented cue and aid in retrieval of the target. Retrieval leads to creation of more effective mediators than restudy, thereby promoting recall (Carpenter, 2011).

Memorial Benefit of Spaced Retrieval on Transfer Items

The second aim of the study was to investigate whether spaced retrieval would benefit recall of Swahili terms using definitions that were semantically related to previously studied definitions, but never actually learned. Given that prior research has shown that transfer can occur for different types of materials (for a meta-analysis see Adesope et al., 2017), we hypothesized that spaced retrieval would be particularly beneficial for transfer using term-definition pairs. Our results supported this prediction, with the qualification that it depends on the type of spaced retrieval used. Regardless of whether the same definitions were used during each relearning session (Experiment 1) or different definitions were used (Experiment 2), we found that the standard condition promoted transfer over restudy. The finding of superior transfer in the standard condition compared to restudy is in line with previous research on the transfer benefits of retrieval (Pan & Rickard, 2018b), as well as previous findings on improved transfer following spaced retrieval compared to repeated restudy (Butler et al., 2017; Butler, 2010). In contrast, we were surprised to find the lack of transfer benefit in the dropout condition. It is possible that learning to a criterion of one is not sufficient for transfer to occur.

According to the three-factor framework of transfer of test-enhanced learning, three factors affect transfer: response congruency, elaborated retrieval practice, and initial learning. Response congruency refers to the similarity between responses required at relearning and responses required on the final test (e.g., “mtoto” as a response to “a very young child” at relearning and to “a human offspring” at final test). Elaborated retrieval practice contains broad encoding methods, which refers to study designs that orient participants to information from study session other than the target (e.g., “mtoto” refers to “baby”), and elaborative feedback, which refers to feedback that explains the underlying concept of the correct response (e.g., both “a very young child” and “a human offspring” refer to a baby). Finally, initial learning refers to the amount of information retained at the end of relearning.

In the current study, the restudy, standard, and dropout condition all had the same degree of response congruency, as they were all required to retrieve the same targets on the final test as the ones they encountered at relearning. The three conditions were also equated on the degree of elaborative feedback. While the retrieval groups received corrective feedback (i.e., “correct!” or “incorrect!”) with

the opportunity to restudy the correct answer, and the restudy group was given the same length of time to study the correct answer, no condition was given any further information about why a particular response is correct. In contrast, retrieval conditions presumably had a higher degree of the factor of initial learning than the restudy condition. This assumption is based on the classic testing effect that testing results in better memory than restudy does. The analysis of relearning accuracy on the first cycle of each session showed that initial learning (i.e., recall accuracy in Session 3) did not differ between the standard and dropout conditions. Broad encoding methods (i.e., greater cognitive processing) were also assumed to be greater in the retrieval conditions compared to the restudy condition, as retrieving information requires more cognitive processing (e.g., memory search) than simply restudying it. The results of Experiment 1 showed that the standard, but not dropout, procedure promoted transfer over restudy led us to explore whether there was an imbalance in the degree of broad encoding methods between the two retrieval conditions, as it was the only factor that might have been different between them. Broad encoding methods are supported by study designs that orient participants to engage in more elaborative processing (Pan & Rickard, 2018b). Therefore, we argued that lack of the transfer benefit in the dropout condition was due to the procedure not inducing as much elaborative processing as the standard procedure does. Indeed, Experiment 2 showed that when greater cognitive processing is promoted via encoding variability, the differences in retention and transfer between the two conditions were attenuated. In other words, it seems that multiple retrievals (at each session) in the standard condition produce substantial "broad encoding" to promote transfer, whereas a single retrieval does not.

Although the dropout method did not result in greater transfer than the restudy condition, it is worth noting that cross-experimental comparison revealed that variable encoding increased transfer for the dropout methods (albeit only marginally), but not the standard or restudy conditions. This suggests the variable encoding alone does not produce transfer if participants are simply restudying materials. Likewise, if participants have engaged in multiple retrievals of the same item, as in the standard condition, variable encoding may have little influence on memory for transfer items. However, variable encoding combined with minimal retrieval (i.e., a criterion of one) may at least have some benefit for

performance. To address this issue, in future work we plan to compare fixed versus variable encoding within-subjects and to manipulate the criterion level (e.g., 1 vs. 5) required in the dropout condition. In any manner, the failure to find benefits of variable encoding in standard and restudy conditions, and only a marginal benefit in the dropout condition, could be because the varied definitions used in our experiment did not introduce enough encoding variability to produce an effect. Butler (2010) found that rephrasing questions participants had to answer at relearning did not boost transfer over relearning with the same questions. A later study using similar design overcame this problem by creating questions that required participants to apply a concept embedded in previously learned information to a new example (Butler et al., 2017). Perhaps the definitions we created did not differ enough from each other to simulate applying knowledge to a new example.

Comparing the Two Repeated Retrieval Techniques

We were also interested in examining differences between two spaced retrieval techniques: standard spaced retrieval and dropout spaced retrieval. The robust memorial benefit of spaced retrieval has prompted researchers to recommend its use in school settings (Roediger & Pyc, 2012; Dunlosky & Rawson, 2015; Roediger & Karpicke, 2006a). Time consumption of any learning method is an important consideration for students, and therefore finding a learning technique that is most time efficient while still producing long-term retention and transfer is important for naturalistic applications of learning research. The results from the current study demonstrated the dropout procedure was more time-efficient and resulted in comparable memory for repetition items to the standard condition. However, this was not the case for the transfer items. Thus, recommendations for which procedure is most optimal depends on the goal of the learner (i.e., retention versus transfer). In addition to application, the results from the current study have important implications for the theoretical mechanisms underlying standard versus dropout methods.

Both the standard and dropout techniques provide a robust memorial benefit on long-term learning over restudying (Rowland, 2014; Rawson et al., 2013; Rawson et al., 2018). However, these two techniques have never been compared to each other with regard to their benefits on transfer. While both

techniques involve several sessions of retrieval, *within* in each session, the standard procedure requires spaced retrieval of all items regardless of retrieval success, and the dropout procedure only requires that items be retrieved until the criterion is met. This leads to an imbalance of exposure and retrieval attempts for some targets between the conditions. For example, if the word "mtoto" is successfully retrieved on the first cycle of the spaced retrieval condition, it will have to be retrieved several more times. In contrast, "mtoto" only has to be successfully retrieved in the dropout condition until it is recalled to criterion (e.g., once), after which it is dropped from further testing and restudy. In essence, the difference between the two conditions is one of intersession repeated retrieval. As a result, the dropout method might be less time consuming, but the question remained whether it is as beneficial as the standard method.

We found that when learning to a criterion of one, the dropout procedure did take less time and resulted in the same degree of initial relearning (i.e., on cycle 1 of each relearning day). However, unlike the standard condition, performance on the final recall for transfer items in the dropout condition was not better than repeated restudy when studying with the same definitions or different definitions (although performance was at least numerically greater following variable encoding). One possibility for the differences in performance across retrieval conditions is memory strength. Every time information is successfully retrieved, the memory trace for that information is strengthened and decays at a slower rate (Pavlik & Anderson, 2005; Bjork & Bjork, 1992). Information in the dropout condition could only be successfully retrieved once at each relearning session, and therefore could only be strengthened once. In contrast, because participants in the standard condition had to attempt retrieval of each item 5 times in each relearning session, there is a possibility that some items were successfully retrieved more than once. If this is the case, more targets had stronger memory traces in the standard condition than in the dropout condition at the end of each session. Moreover, these memory traces would decay at a slower rate than the memory traces for targets in the dropout condition. A week after the last relearning session, then, the standard condition would have more memory for the studied material than the dropout condition, which would result in better retention and a higher probability of transfer occurring.

An alternative explanation of our findings is proposed by the desirable difficulties hypothesis (Bjork, 1994, 1999). The hypothesis states that learning that is slower and more effortful results in better long-term retention than learning that is fast and easy. Karpicke and Roediger (2007) had participants in one condition engage in several restudy sessions, after each of which they took a free-recall test on all of the material (STST condition). For participants in another condition, the words that were recalled correctly were dropped from future restudy and testing blocks (STS_nT_n condition). A week after the initial session, all participants returned for a final free-recall test on all of the words studied initially. Their results showed that recalling each word only once resulted in worse retention after a week than retrieving all words several times. These findings suggest that retrieving items several times successfully creates more difficult initial relearning conditions than successfully recalling each item once, as participants have to engage in memory search and produce a correct response several times. Similarly, Rawson et al. (2020) found that the dropout procedure provided only a marginal benefit on retention over single-session learning. They proposed that learning to a criterion of one stops learning before functional mastery is achieved. This explanation is supported by the literature on overlearning, which suggests that continuing learning after material has been learned enough to be recalled enhances long-term retention (Postman, 1962; for a meta-analysis, see Driskell et al., 1992).

These two explanations are not mutually exclusive. It is possible that the mechanism driving the learning benefit of overlearning is that it introduces desirable difficulties by forcing learners to engage in more elaborative processing. This would explain superior transfer in the standard condition relative to the dropout condition we found in Experiment 1. To explore whether this is the case, future research should manipulate learning criterion in the dropout condition to examine learning and transfer.

Conclusion

The results of the current study showed that spaced retrieval improve retention and transfer over restudy in novel vocabulary learning. Moreover, the comparison of the standard and dropout procedure of spaced retrieval revealed that while standard spaced retrieval takes more time, it results in more transfer. Our findings suggest that if the goal of learning is retention, the dropout method is a time efficient

procedure that yields long-term retention, even without variability in learning materials. To achieve the goal of transfer, the standard procedure is more beneficial. However, when used with variable materials, the dropout method provides a faster alternative to the standard method. These findings have important implications for future research on mechanisms underlying spaced retrieval as well as recommendations made to instructors.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659-701.

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The Value of Applied Research: Retrieval Practice Improves Classroom Learning and Recommendations from a Teacher, a Principal, and a Scientist. *Educational Psychology Review*, 12.

Bahrick. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*(3).

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in second language acquisition, 27*(3), 387-414.

Bjork, R. A. (1994). Memory and metamemory considerations in the. *Metacognition: Knowing about knowing, 185*(7.2).

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. *Attention and Performance, 17*.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes, 2*, 35-67.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118-1133. https://doi.org/10.1037/a0019902

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *J Exp Psychol Appl, 23*(4), 433-446. https://doi.org/10.1037/xap0000142

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *J Exp Psychol Learn Mem Cogn, 37*(6), 1547-1552. https://doi.org/10.1037/a0024140

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & cognition, 34*(2), 268-276.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23*(6), 760-771.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & cognition, 36*(2), 438-448.

Chan, J. C., McDermott, K. B., & Roediger III, H. L. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553.

Cho, K. W., Neely, J. H., Brennan, M. K., Vitrano, D., & Crocco, S. (2017). Does testing increase spontaneous mediation in learning semantically related paired associates? *J Exp Psychol Learn Mem Cogn, 43*(11), 1768-1778. https://doi.org/10.1037/xlm0000414

Coppens, L. C., Verkoeijen, P. P., & Rikers, R. M. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology, 23*(3), 351-357.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 14*(3), 215-235.

Dikmans, M. E., van den Broek, G. S. E., & Klatter-Folmer, J. (2020). Effects of repeated retrieval on keyword mediator use: shifting to direct retrieval predicts better learning outcomes. *Memory, 28*(7), 908-917. https://doi.org/10.1080/09658211.2020.1797094

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*(5), 615.

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology, 1*(1), 72-78. https://doi.org/10.1037/stl0000024

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological review, 62*(5), 369.

Furst, E. (2020). Successive Relearning aka Spaced Retrieval Practice. *Teaching with learning in mind*.

Gass, S. (1999). Discussion: Incidental vocabulary learning. *Studies in second language acquisition, 21*(2), 319-333.

Goode, M. K., Geraci, L., & Roediger, H. L. (2008). Superiority of variable to repeated practice in transfer on anagram solution. *Psychonomic bulletin & review, 15*(3), 662-666.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic bulletin & review, 19*(1), 126-134. https://doi.org/10.3758/s13423-011-0181-y

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1441.

Kang, S. H., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic bulletin & review, 18*(5), 998-1005.

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528-558.

Karpicke, J. D., Butler, A. C., & Iii, H. L. R. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*(4), 471-479.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*(2), 151-162.

Kukull, W. A., Higdon, R., Bowen, J. D., McCormick, W. C., Teri, L., Schellenberg, G. D., van Belle, G., Jolley, L., & Larson, E. B. (2002). Dementia and Alzheimer disease incidence: a prospective cohort study. *Arch Neurol, 59*(11), 1737-1746. https://doi.org/10.1001/archneur.59.11.1737

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *J Exp Psychol Learn Mem Cogn, 42*(10), 1573-1591. https://doi.org/10.1037/xlm0000267

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: an encoding variability hypothesis. *Psychological review, 75*(5), 421.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399-414. https://doi.org/10.1037/a0021782

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4-5), 494-513.

Pan, S. C., & Rickard, T. C. (2017). Does Retrieval Practice Enhance Learning and Transfer Relative to Restudy for Term-Definition Facts? *Journal of Experimental Psychology: Applied, 23*(3).

Pan, S. C., & Rickard, T. C. (2018a). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied, 23*.

Pan, S. C., & Rickard, T. C. (2018b). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710-756. https://doi.org/10.1037/bul0000151

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.

Postman, L. (1962). Retention as a function of degree of overlearning. *Science, 135*(3504), 666-667.

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. (2010). Why testing improves memory: mediator effectiveness hypothesis. *Science, 330*(6002), 335. https://doi.org/10.1126/science.1191465

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *J Exp Psychol Learn Mem Cogn, 38*(3), 737-746. https://doi.org/10.1037/a0026166

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., Dunlosky, J., & Sciartelli, S. M. (2013). The Power of Successive Relearning: Improving Performance on Course Exams and Long-Term Retention. *Educational Psychology Review, 25*(4), 523-548. https://doi.org/10.1007/s10648-013-9240-4

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Mem Cognit, 43*(4), 619-633. https://doi.org/10.3758/s13421-014-0477-z

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

Rivers, M. L. (2021). Metacognition about practice testing: a review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review, 33*(3), 823-862.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *J Exp Psychol Appl, 17*(4), 382-395. https://doi.org/10.1037/a0026252

Roediger, H. L., III, & Karpicke, J. D. (2006a). 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

Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248. https://doi.org/10.1016/j.jarmac.2012.09.002

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol Bull, 140*(6), 1432-1463. https://doi.org/10.1037/a0037559

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1582.

Smith, S. M., & Handy, J. D. (2016). The crutch of context-dependency: Effects of contextual support and constancy on acquisition and retention. *Memory, 24*(8), 1134-1141.

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & cognition, 40*(5), 703-716.