WHY IS RETRIEVAL PRACTICE BENEFICIAL FOR MEMORY?
AN EVALUATION OF THE MEDIATOR SHIFT HYPOTHESIS

A dissertation submitted to
Kent State University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

by
Mary A. Pyc

December 2010
# TABLE OF CONTENTS

LIST OF FIGURES...........................................................................................................v

ACKNOWLEDGMENTS......................................................................................................vii

INTRODUCTION.............................................................................................................1

Theoretical Accounts of Testing Effects.................................................................4
Current Experiments.................................................................................................11

METHOD EXPERIMENTS 1A AND 1B.................................................................16

Participants and Design.................................................................16
Materials........................................................................................................16
Procedure....................................................................................................16
Results and Discussion........................................................................19

INTRODUCTION EXPERIMENT 2.................................................................25

METHOD EXPERIMENT 2.......................................................................................27

Participants and Design........................................................................27
Materials........................................................................................................27
Procedure....................................................................................................27
Results........................................................................................................29

INTRODUCTION EXPERIMENT 3.................................................................37

METHOD EXPERIMENT 3.......................................................................................38

Participants and Design........................................................................38
Materials........................................................................................................38
Procedure....................................................................................................38
Results and Discussion........................................................................39

GENERAL DISCUSSION..................................................................................48

REFERENCES..........................................................................................................56
APPENDIX…………………………………………………………………………………..61
LIST OF FIGURES

1  Mean proportion of items correctly recalled on the final test as a function of practice group in Experiment 1a and Experiment 1b.
2  Mean proportion of keyword shifts during practice as a function of practice group in Experiment 1a and Experiment 1b.
3  Mean proportion of keyword shifts during practice as a function of retrieval status during practice for the test-restudy group in Experiment 1a and Experiment 1b.
4  Mean proportion of items correctly recalled on the final test as a function of practice group and lag condition in Experiment 2.
5  Mean proportion of keyword shifts during practice as a function of practice group and lag condition in Experiment 2.
6  Mean proportion of keyword shifts during practice as a function of retrieval status during practice for conditions of the test-restudy group in Experiment 2.
7  Mean proportion of keywords recalled on the final keyword recall test as a function of practice group and lag in Experiment 2.
8  Mean proportion of items correctly recalled on the final test as a function of group and lag condition in Experiment 3.
9  Mean proportion of keyword shifts during practice as a function of group and lag condition in Experiment 3.
Mean proportion of keyword shifts during practice as a function of retrieval status during practice for conditions of the test-restudy group in Experiment 3.

Mean proportion of keywords recalled on the final keyword recall test as a function of group and lag condition in Experiment 3.
ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Katherine A. Rawson for her continuous support and guidance. I would also like to thank Dr. John Dunlosky, as well as members of RADlab and all of my committee members – Dr. Maria S. Zaragoza, Dr. Jeffrey A. Ciesla, Dr. Christopher Was, and Dr. Mary Ann Heiss – for their support during this process. Finally, I would like to thank my family for their constant love and support.
INTRODUCTION

From the earliest days of research on memory, cognitive scientists have been interested in factors that promote memory (e.g., Ebbinghaus, 1913/1885). Current research on memory is no exception, as an increasing number of researchers are exploring ways to promote memory. This research is especially important when considering the implications of improving memory for student learning and scholarship. Students are expected to remember a great deal of information across a broad range of topics. Therefore, it is important for researchers to identify strategies for learning that are optimal to increase the likelihood that students will form durable memory for the large amount of information they are expected to learn.

In the memory literature, tests have traditionally been used as a means of assessment. That is, researchers typically test individuals to evaluate the efficacy of various manipulations for improving memory. However, researchers have discovered that tests can be used not only as a means of assessment, but as an effective means to promote memory (i.e., the testing effect; for a review see Roediger & Karpicke, 2006). The testing effect is the robust finding that performance on a retention test is enhanced when individuals engage in test practice compared to only restudy practice (e.g., Carrier & Pashler, 1992; Chan, McDermott, & Roediger, 2006; Cull, 2000; Morris, Fritz, Jackson, Nichol, & Roberts, 2005; Roediger & Karpicke, 2006; Wheeler, Ewers, & Buonanno, 2003). Further, research has established that testing followed by restudy
during practice is a particularly potent combination for memory. Performance on retention tests is enhanced when individuals engage in test-restudy practice compared to restudy-only practice or test-only practice (e.g., Carpenter & DeLosh, 2005; Cull, 2000; Karpicke & Roediger, 2007a). Finally, previous research has shown additional memorial benefits when test practice is distributed across time or other materials compared to when test practice is massed (e.g., Carpenter & DeLosh, 2005; Cull, 2000; Cull, Shaughnessy, & Zechmeister, 1996).

For example, Cull (2000) presented word pairs for an initial study trial, followed by three practice trials that were either massed (all practice trials for a given item occurred consecutively) or distributed (other items intervened between each next practice trial for a given item). Within each schedule, practice trials involved either restudying the word pairs (restudy-only), presentation of the cue word as a prompt to retrieve the target word (test-only), or test followed by restudy (test-restudy). A short filler task was followed by a final cued recall test. Results showed higher levels of final test performance for items learned with distributed practice compared to massed practice. Additionally, final test performance was higher for items learned with test-restudy practice versus test-only practice or restudy-only practice. Most important, performance on the final retention test was highest for items learned with distributed test-restudy practice (49%, compared to an average of 26% across all other groups). Accordingly, this particular testing schedule (i.e., distributed test-restudy) will be the focus of the current research.

The testing effect is robust, and has been demonstrated across a wide range of materials, test formats, and learners. For example, the testing effect has been found for
foreign language word pairs (e.g., Karpicke & Roediger, 2008; Pyc & Rawson, 2007), essay texts (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008), word lists (e.g., Carpenter & DeLosh, 2006), general knowledge facts (e.g., Carpenter, Pashler, Wixted, & Vul, 2008), face-name pairs (e.g., Landauer & Bjork, 1978), and has even been found for non-verbal visuospatial materials such as map learning (e.g., Carpenter & Pashler, 2007). The testing effect has been demonstrated across a variety of test formats including cued recall tests (e.g., Pyc & Rawson, 2009), free recall tests (e.g., Karpicke & Roediger, 2007a), and multiple choice tests (e.g., Butler, Marsh, Goode, & Roediger, 2006). Finally, the testing effect has been demonstrated across the lifespan including preschool-aged children (e.g., Fritz, Morris, Nolan, & Singleton, 2007), young adults (e.g., Roediger & Karpicke, 2006), and older adults (e.g., Logan & Balota, 2008).

Although the testing effect has been found across a wide variety of materials, test formats, and learners, theoretical explanations for why testing is beneficial for memory have received much less attention. Very few studies have systematically evaluated theoretical explanations for the memorial benefits of testing, and those that have evaluated theory have not always compared test-restudy schedules to restudy schedules of learning (the testing schedule of interest here). Thus, the mechanisms underlying the memorial benefits of testing are not well understood. Here I discuss theories that could be relevant to explaining why performance is better on retention tests after test-restudy practice compared to restudy practice. Note, these theories are not mutually exclusive and the goal of the current research is not to competitively evaluate these theories.
Rather, I discuss them here to illustrate previous studies that have evaluated theory, and also to contrast those theories with the hypothesis I evaluated in the current experiments.

Theoretical Accounts of Testing Effects

The elaborative retrieval hypothesis is one theoretical explanation for why performance is better on retention tests after test practice compared to restudy practice (e.g., Carpenter & DeLosh, 2006; Carpenter, 2009). The elaborative retrieval hypothesis is based on the assumption that performance will be better on a final retention test when more elaborative versus less elaborative memory traces are formed during practice (cf. encoding variability hypothesis, McDaniel & Masson, 1985). The elaborative retrieval hypothesis states that more elaborative memory traces are formed during encoding for items that are learned with test practice versus restudy practice. For test trials during practice, individuals must attempt to retrieve a target answer when they are provided with a minimal amount of information to do so (e.g., for a cued recall test, only the cue is provided and the individual must generate the target answer). Thus, a search through memory is necessary to retrieve an answer. Searching through memory for the correct answer activates other related information, and connections are formed between the cue (whatever information was provided to the individual to begin the retrieval process), the target answer, and other related information in memory. For restudy trials, however, this same search through memory is not necessary because all information is provided for the individual (e.g., for word pair learning, both the cue and target are provided). Therefore, more elaborative connections are formed for items that are tested during practice compared to items that are only restudied during practice. On a final retention test,
performance will be greater for items that were tested versus restudied because more elaborative connections have been formed for tested items, which aids retrieval of the target answer (because multiple routes may be used to arrive at the correct target answer).

To test this hypothesis, Carpenter (2009) had participants learn concrete noun word pairs that were either strongly associated (e.g., toast-bread) or weakly associated (e.g., basket-bread) using either test practice or restudy practice. According to the elaborative retrieval hypothesis, for test practice trials, more elaborate traces will be formed for weakly associated versus strongly associated word pairs because a more extensive search through memory is necessary to semantically link weakly versus strongly associated items. Therefore, on a final retention test, performance will be greater for weakly associated versus strongly associated word pairs because more elaborative connections have been formed during encoding for weakly versus strongly associated word pairs. The same claim is not made for restudy practice trials because no retrieval is required, which is presumably when elaborative traces are formed. Supporting the elaborative retrieval hypothesis, final test performance was greater for weakly associated versus strongly associated word pairs that were tested during practice. However, performance did not differ for weakly associated versus strongly associated word pairs that were restudied during practice. These results suggest that one important component of test practice may be that more elaborate traces are formed during encoding, whereas the same elaborate traces are not formed during restudy practice.

Another relevant theoretical framework is the desirable difficulty framework (Bjork, 1994). One basic claim of the desirable difficulty framework is that performance
on a retention test will be greater when processing during encoding is successful but
difficult compared to when processing during encoding is successful but less difficult.

An instantiation of the desirable difficulty framework, the *retrieval effort hypothesis*, was
recently evaluated as an explanation for the memorial benefits of testing (Pyc & Rawson,
2009). The retrieval effort hypothesis states that performance on a retention test will be
greater when correct retrievals during practice are more difficult compared to when
correct retrievals during practice are less difficult.

To produce conditions involving more difficult versus less difficult retrieval
during practice, Pyc and Rawson (2009) manipulated lag (the number of items between
each next practice trial with a given item) and criterion level (the number of times an item
is correctly recalled during practice). Based on previous research, correct retrievals were
predicted to be more difficult for items learned with longer versus shorter lags during
practice (e.g., Karpicke & Roediger, 2007b). Therefore, final test performance was
predicted to be greater after practice involving longer versus shorter lags. For the
criterion level manipulation, correct retrievals were predicted to be more difficult for
those that occurred earlier during learning for a given item compared to correct retrievals
that occurred later during learning (i.e., the first correct retrieval for a given item will be
more difficult compared to the 10th correct retrieval for that same item, e.g., Karpicke &
Roediger, 2007b). Therefore, the incremental benefit to final test performance was
predicted to decrease as the number of correct retrievals during practice increased.

Difficulty of retrieval was measured by recording first keypress latency (the amount of
time from cue onset until participants began typing target answers) on test trials during
practice. Across a series of two experiments, Pyc and Rawson (2009) confirmed
predictions of the retrieval effort hypothesis, showing that the benefit to final test
performance was greater when correct retrievals during practice were more difficult
compared to when they were less difficult. Although they did not compare test-restudy
and restudy schedules of practice, the retrieval effort hypothesis would predict that
performance on retention tests is greater for test-restudy versus restudy schedules of
practice because processing is more difficult for test-restudy items compared to restudy
items. However, to date, this prediction has not been directly evaluated.

Finally, the strategy shift hypothesis (Bahrick & Hall, 2005) may be able to
explain why performance is greater on retention tests after test-restudy practice compared
to restudy practice. The basic claim of the strategy shift hypothesis is that retrieving
information from memory during practice enables individuals to evaluate the
effectiveness of strategies used to encode information. More specifically, on practice
tests, individuals may experience retrieval failure. When individuals experience retrieval
failure they evaluate the effectiveness of the strategy they used to encode an item, and
shift from a less effective to a more effective strategy during subsequent restudy, which is
beneficial for later recall. For example, for the word pair “dog-spoon”, an individual may
originally use rote repetition to encode the word pair. However, because this is not an
effective encoding strategy, a learner is likely to experience retrieval failure on a
subsequent practice trial. As a result, the individual may shift to using a more effective
strategy, such as interactive imagery (e.g., imagining a dog eating with a spoon). In
contrast, for restudy practice schedules, individuals are more likely to continue using rote
repetition because they are less likely to evaluate the effectiveness of the strategy they are using during encoding because they do not experience retrieval failure. Note that in contrast to the previously discussed theoretical accounts, this hypothesis includes a metacognitive component. That is, the strategy shift hypothesis is based on the premise that individuals evaluate the efficacy of encoding strategies, and shift from less effective to more effective encoding strategies.

The primary goal of Bahrick and Hall (2005) was to evaluate the strategy shift hypothesis for the spacing effect (i.e., the memorial benefits of distributed practice compared to massed practice), rather than the memorial benefits of test-restudy practice versus restudy practice. Nonetheless, I describe their studies in detail here because the hypothesis I evaluated in the current experiments was based on the strategy shift hypothesis. Bahrick and Hall (2005) tested the strategy shift hypothesis using Swahili-English word pairs. Participants received instructions explaining various encoding strategies they could use to learn the Swahili-English word pairs, including rote repetition, verbal elaboration (i.e., generating a word or sentence connecting the cue and target), and visual elaboration (i.e., forming an image connecting the cue and target). Participants learned word pairs across four practice sessions that were either massed (each session began when the previous session ended) or distributed (practice sessions spaced across days). After an initial study trial with each word pair, participants had a strategy trial in which they reported which strategy they used to encode the word pair. After the strategy report, participants had a self-paced test trial in which they were given the cue and had to retrieve the target. Items that were correctly recalled were dropped
from practice, whereas items that were incorrectly recalled received another restudy trial. After the restudy trial, participants again reported the strategy they used to encode the word pair. This process continued until all items were correctly recalled once. The next practice session occurred immediately for participants assigned to massed practice, whereas participants assigned to distributed practice came back for their next session either one day later or 14 days later. Each participant had four relearning sessions. Participants came back for a final cued-recall retention test 14 days after their last relearning session.

On the final retention test, performance was greater after distributed practice compared to massed practice. Importantly, strategy reports showed that participants more often used visual or verbal elaboration during distributed practice compared to massed practice. The researchers attributed this result to retrieval failures during practice. That is, participants engaged in massed practice did not experience as many retrieval failures during practice as did participants engaged in distributed practice and thus did not have the opportunity to evaluate the effectiveness of their encoding strategies. A second experiment did not collect strategy reports, so it is difficult to determine which strategies participants used to encode items (if they used any). However, participants took longer restudying items after retrieval failure trials compared to retrieval success trials, which the authors took as evidence that participants were shifting encoding strategies after retrieval failure trials during practice.

Results from Bahrick and Hall (2005) provide some support for the strategy shift hypothesis for the spacing effect. That is, participants engaged in distributed practice
more often used effective strategies, and latency results showed that participants took longer encoding items after retrieval failure trials versus retrieval success trials. However, their experiments had some limitations that I discuss here to highlight important changes for the current experiments. First, Bahrick and Hall (2005) evaluated shifting of global strategies; they measured the extent to which participants shifted from one strategy to another. It is possible, however, that participants shifted specific mediators within a given strategy. For example, participants may have shifted from a less effective visual elaboration to a more effective visual elaboration after retrieval failure trials during practice (e.g., shifting from a non-interactive image of a dog sitting next to a spoon to an interactive image of a dog eating with a spoon). A second limitation of Bahrick and Hall (2005) is that Experiment 2 used latencies after retrieval failure trials versus retrieval success trials to infer that participants were shifting strategies after retrieval failure trials during practice, but they did not collect strategy reports to confirm this assumption. Therefore, it is difficult to determine the extent to which participants were using a longer amount of time restudying items after retrieval failure trials because they were generating new strategies or if they were engaged in some other activity. Finally, because Bahrick and Hall’s (2005) goal was to evaluate the strategy shift hypothesis for the spacing effect, their experiments were not designed to compare test-restudy and restudy schedules of practice. Therefore, the current experiments compared shifting of mediators during practice for test-restudy and restudy schedules of practice.
Current Experiments

The current experiments were designed to evaluate a variant of the strategy shift hypothesis, the *mediator shift hypothesis*. The mediator shift hypothesis states that individuals will shift to more effective mediators within a given strategy as a function of the status of retrieval (i.e., retrieval failure or retrieval success) on test trials during practice. According to the mediator shift hypothesis, the memorial benefits of test-restudy practice are greater than the memorial benefits of restudy practice because individuals monitor and modify mediators when they cannot correctly recall an item during practice, whereas individuals engaged in restudy practice cannot do so because they do not experience retrieval failure.

The goal of the current experiments was to evaluate the mediator shift hypothesis for learning foreign language word pairs. Therefore, it was important to instruct individuals to use a global strategy that is amenable to learning foreign language word pairs. One effective strategy for learning these word pairs is a keyword encoding strategy (e.g., McDaniel, Pressley, & Dunay, 1987; Raugh & Atkinson, 1975; Fritz, Morris, Acton, Voelkel, & Etkind, 2007), which involves two basic steps. First, learners are instructed to generate an English word that is orthographically or phonologically similar to the foreign language cue. Second, learners are instructed to relate the generated English keyword to the English target so that the two words are semantically connected. For example, consider the Swahili-English word pair “wingu – cloud”. An English word that is similar to “wingu” is “wing”, and the connection between “wingu” and “cloud” could be “birds have wings and fly in clouds”. Therefore, when prompted with “wingu”,
a learner would first retrieve the word “wing” and then relate it to birds, which fly in clouds, to retrieve the target “cloud”.

The keyword encoding strategy may be especially useful for foreign language learning compared to other effective encoding strategies such as sentence generation. More specifically, it may be difficult to generate a sentence for a foreign language word pair that an individual has never been exposed to before. However, the keyword encoding strategy can be applied to almost any language because the task involves generating an English word that is similar to the foreign language word. Therefore, the current experiments will exclusively use this strategy to evaluate the efficacy of the mediator shift hypothesis for explaining the memorial benefits of testing.

To facilitate understanding of the predictions of the mediator shift hypothesis, I now describe the basic method for the current experiments. First, participants received an initial study trial and initial keyword trial for each item. For the initial study trials both the cue and target were presented and participants were told to use the keyword encoding strategy to help them remember the word pairs. For initial keyword trials, participants were asked to type in the keyword they used to associate the cue and target. Second, items were presented for either test-restudy practice or restudy practice. For test trials, the cue (e.g., wingu) appeared on the computer screen and participants typed in the target answer (e.g., cloud). For restudy trials, both the cue and target words appeared on the computer screen. After each restudy trial in both groups, participants completed a keyword report trial in which they reported the specific keyword used to encode a given item. Third, participants returned two days later for a final cued-recall retention test,
followed by a test of their memory for the keyword mediators they generated during Session 1.

For all experiments, final cued-recall performance was predicted to be greater for items learned with test-restudy practice versus restudy practice. The mediator shift hypothesis makes three basic predictions regarding why final test performance will be greater after test-restudy practice versus restudy practice. Prediction 1 is related to keywords generated during practice, and states that a greater proportion of keyword shifts (i.e., changing from a keyword that was previously generated for a given word pair to a new keyword for that word pair) will occur during practice for items learned with test-restudy practice versus restudy practice. A greater proportion of keyword shifts are predicted for items learned with test-restudy practice versus restudy practice because individuals engaged in test-restudy practice will experience retrieval failure during practice. These individuals will be better able to evaluate the effectiveness of keywords (compared to individuals engaged in restudy practice), and will shift to effective keywords more often than will individuals engaged in only restudy practice.

Prediction 2 specifically relates to keyword use for items learned with test-restudy practice. Prediction 2 states that a greater proportion of keyword shifts will occur after retrieval failure trials versus retrieval success trials during practice. On retrieval failure trials during practice, individuals will evaluate the effectiveness of the keyword they generated for a given item and will shift from a less effective keyword to a more effective keyword.
Prediction 3 is related to keyword recall after the final test. Prediction 3 states that a greater proportion of keywords generated during practice will be recalled on the final keyword recall test for items learned with test-restudy practice compared to items learned with restudy practice. The premise of the mediator shift hypothesis is that testing during practice leads to the formulation of more effective keywords, which facilitates retrieval of targets on retention tests. Two factors have been shown to influence the effectiveness of mediators, mediator retrieval and mediator decoding (e.g., Dunlosky, Hertzog, & Powell-Moman, 2005). For a mediator to be effective it must be recallable when prompted with a cue and it must be decoded to elicit the target from memory. The final keyword recall test in the current research provides a measure of mediator retrieval, one of the key components of mediator effectiveness.

I now provide a brief roadmap for the remainder of the paper. Experiments 1a and 1b evaluated these three predictions of the mediator shift hypothesis. Experiments 1a and 1b differed only in the number of items learned and the number of practice trials each item received. These experiments are reported together below because they produced similar results. To foreshadow, results from these first two experiments were in line with Predictions 1 and 2 of the mediator shift hypothesis, but failed to support Prediction 3. To provide a stronger test of the mediator shift hypothesis, Experiment 2 was designed to examine more difficult learning conditions which were predicted to increase the proportion of retrieval failures during practice. Importantly, Experiment 2 was also designed to provide a better test of Prediction 3 by revising the format of the final keyword recall test. Results from Experiment 2 supported all predictions from the
mediator shift hypothesis. Experiment 3 was designed to rule out a potential alternative explanation for the testing effect observed in Experiment 2.
METHOD EXPERIMENTS 1A AND 1B

Participants and design

Participants included 53 Kent State University undergraduates in Experiment 1a and 54 in Experiment 1b who participated in return for course credit. Participants were randomly assigned to test-restudy or restudy practice groups, with 26 to 28 participants in each group for Experiments 1a and 1b, respectively.

Materials

Items included 30 (Experiment 1a) and 45 (Experiment 1b) Swahili-English word pairs previously normed for item difficulty (Nelson & Dunlosky, 1994). Target items for the current experiments were selected from a larger list based on results from a pilot study. In the pilot study, participants were asked to generate keywords for the full set of normed Swahili-English word pairs. The word pairs selected for the current experiments were those for which participants generated the most keywords in the pilot study. These word pairs were selected to increase the likelihood that participants would be able to come up with a keyword for each word pair during practice, and also to ensure that it was possible to come up with more than one keyword for each word pair.

Procedure

Task instructions and item presentation were administered via computer. All trials were self-paced. On the instruction screen, participants received detailed
instructions regarding the keyword encoding strategy, which they were told they should use during the task. The instructions were as follows:

Your primary task today is to learn the English translation for Swahili words (e.g., *Mshoni-Tailor*). Previous research has indicated that the keyword method is an effective way to promote memory for word pairs such as the Swahili-English pairs you will learn today (e.g., *Mshoni-Tailor*).

The keyword method involves generating a word to associate a foreign language word (e.g., *Mshoni*) with an English word that you already know that will help you remember the correct English translation for the Swahili word (e.g., *SHOE* for *MSHONI* because they sound similar. The word *SHOE* is then related to *TAILOR* because at one time TAILORS made SHOES).

Participants were provided two additional examples of the keyword encoding strategy, which was followed by instructions regarding the remainder of the task (i.e., that they would have multiple restudy or test-restudy trials with each item, depending on the group to which they were assigned). After reading the instructions, participants were prompted to see the experimenter. At this time, the experimenter had participants describe what they would be doing in the task to ensure that all participants fully understood the keyword encoding strategy and the task that they would be completing.

Each word pair was first presented for an initial study trial and an initial keyword report trial. For initial study, the Swahili-English word pair was presented on the screen until the participant pressed a button indicating that they were done studying the item. After the initial study trial for a given item, participants immediately completed a keyword report trial, in which they were asked to report the specific keyword mediator they used to associate the Swahili-English word pair. Upon completing initial study and initial keyword report trials for all items, items received five (Experiment 1a) or three (Experiment 1b) blocks of distributed restudy or distributed test-restudy practice trials.
This distributed practice involved each next practice trial with a given item being separated by a practice trial with all other to-be-learned items. For the restudy groups, the Swahili-English word pair was presented on the screen until the participant pressed a button indicating that they were done studying. After each restudy trial with a given word pair, participants had a keyword report trial. For the test-restudy groups, participants had a self-paced test trial before each restudy trial. For test trials, only the Swahili word was presented and participants typed the English translation in a field provided on the screen.

Upon completion of the experimental task, participants were dismissed and reminded to return two days later. The second session began with a participant-paced cued-recall final test. On each final test trial, participants were prompted with a Swahili cue and asked to report the English target. Upon completion of the cued-recall test, participants had a final keyword recall test for all items. For Experiment 1a, participants were prompted with each Swahili-English word pair and asked to recall all of the keywords from Session 1 that were used for each word pair. For Experiment 1b, participants were prompted with the Swahili-English word pair but were given separate response boxes for each trial they had during practice. For both experiments, if participants could not remember the keyword they used for a given word pair, they were instructed to type “I cannot remember the keyword” in the keyword response box. When participants completed the final keyword recall test, they were debriefed, thanked, and dismissed.
Results and Discussion

The mean proportion of items correctly recalled on the final test as a function of practice group is reported in Figure 1 for Experiments 1a and 1b. As expected, performance was greater for items learned with test-restudy practice versus restudy practice. Independent samples t-tests revealed significant effects of practice group, \( t(51) = 3.48, p < .001 \) (Experiment 1a) and \( t(52) = 3.00, p = .002 \) (Experiment 1b). Thus, both Experiments 1a and 1b replicated the well-established testing effect, which permits evaluation of predictions of the mediator shift hypothesis.

![Figure 1](image.png)

*Figure 1.* Mean proportion of items correctly recalled on the final test as a function of practice group in Experiment 1a and Experiment 1b.
Keywords during practice. To revisit, Prediction 1 states that a greater proportion of keyword shifts will occur during practice for items learned with test-restudy practice versus restudy practice. To test this prediction, analyses were conducted on practice trials that met two criteria reflecting the logically necessary conditions for examining keyword shifting: (a) a keyword had been generated for that item on an earlier practice trial, and (b) a keyword was generated on the current practice trial. Keyword shifts were operationalized as the mean proportion of these trials in which the keyword generated on the current practice trial was different from the keyword generated on the previous trial during practice.

The mean proportion of keyword shifts during practice is reported in Figure 2 for Experiments 1a and 1b. Supporting Prediction 1, a greater proportion of keyword shifts occurred for items learned with test-restudy practice versus restudy practice, \( t(51) = 2.99, p = .002 \) (Experiment 1a) and \( t(52) = 1.75, p = .043 \) (Experiment 1b).
Figure 2. Mean proportion of keyword shifts during practice as a function of practice group in Experiment 1a and Experiment 1b.

Prediction 2 states that a greater proportion of keyword shifts will occur after retrieval failure trials compared to retrieval success trials during practice for items learned with test-restudy practice. The proportions of keyword shifts for retrieval failure trials and retrieval success trials in Experiments 1a and 1b is reported in Figure 3. Consistent with Prediction 2, keywords were more likely to shift after retrieval failure trials versus retrieval success trials, $t(26) = 5.16, p < .001$ (Experiment 1a) and $t(27) = 4.51, p < .001$ (Experiment 1b).
Figure 3. Mean proportion of keyword shifts during practice as a function of retrieval status during practice for the test-restudy group in Experiment 1a and Experiment 1b.

*Keyword recall test.* Prediction 3 states that a greater proportion of keywords from Session 1 will be recalled on the final keyword recall test for items learned with test-restudy practice versus restudy practice. Contrary to this prediction, the mean proportion of keywords recalled was similar for test-restudy and restudy groups in Experiment 1a (for restudy group = .89, SE = .03; for test-restudy group = .88, SE = .02) and Experiment 1b (for restudy group = .80, SE = .04; for test-restudy group = .71, SE = .03). An independent samples t-test showed no significant difference between restudy and test-restudy groups in Experiment 1a \(t(51) = .298, p = .384\). An independent
samples t-test showed significant differences between restudy and test-restudy groups in Experiment 1b \[t(52) = 1.91, p = .062\] but surprisingly in the opposite direction, with a greater proportion of keywords reported on the final keyword recall test in the restudy group compared to the test-restudy group.

One potential explanation for why keyword recall on the final keyword recall test was not in line with Prediction 3 concerns the format of the final keyword recall test. Specifically, participants received both the Swahili cue and English target and were asked to recall the keyword they used for the word pair during Session 1. Therefore, participants may have been able to reconstruct the keyword from Session 1 instead of retrieving it from memory. Thus, one important goal of Experiment 2 was to better evaluate Prediction 3 by using cue-only prompts for the final keyword recall test.

*Encoding time.* Mean encoding time for all trials during practice (i.e., initial study trials, initial keyword trials and all practice trials) in Experiments 1a and 1b was greater for items learned with test-restudy practice versus restudy practice (Experiment 1a: 93.73 minutes, SE = 4.43 versus 62.99 minutes, SE = 3.86, respectively; Experiment 1b: 69.49 minutes, SE = 3.96 versus 52.79 minutes, SE = 3.66, respectively), \(t(51) = 5.23, p < .001\) (Experiment 1a) and \(t(52) = 3.09, p = .003\) (Experiment 1b). This difference in encoding time was due to the additional task of a test trial for items learned with test-restudy practice versus restudy practice. The mean time for restudy trials during practice for test-restudy and restudy groups was 23.03 minutes (SE = 2.18) and 27.43 minutes (SE = 3.18) for Experiment 1a and 19.65 minutes (SE = 2.17) and 23.27 minutes (SE = 2.23) for Experiment 1b, respectively. Additionally, the mean time for test trials during practice
for the test-restudy group was 33.46 minutes (SE = 1.81) and 21.49 minutes (SE = 1.92) in Experiments 1a and 1b.

Although differences for encoding time obtained, it is unlikely that this difference was responsible for the testing effect observed here. Previous research has demonstrated that the testing effect obtains even when time on task is equated for test-restudy and restudy groups (e.g., Carrier & Pashler, 1992); in fact, one study showed that one test trial during practice is as effective as five restudy trials (e.g., Allen, Mahler, & Estes, 1969; see also Roediger & Karpicke, 2006 for a review). Nonetheless, to foreshadow, Experiment 3 will evaluate this alternative interpretation.

Taken together, results from Experiments 1a and 1b confirmed Prediction 1 and Prediction 2 of the mediator shift hypothesis. However, one potential limitation of the current experiments is related to the mean proportion of retrieval failures for each word pair during practice. The mean proportion of retrieval failures was relatively low for Experiment 1a (mean proportion = .21, SE = .04, or about 1 out of 5 trials) and Experiment 1b (mean = .33, SE = .03, or about 1 out of 3 trials), which limits the extent to which keywords can shift during practice (and the extent to which I can evaluate the efficacy of the mediator shift hypothesis for explaining testing effects). Therefore, one goal of Experiment 2 was to examine more difficult learning conditions which were predicted to increase the proportion of retrieval failures for each item during practice for the test-restudy group.
INTRODUCTION EXPERIMENT 2

Results from Experiments 1a and 1b provide some support for the mediator shift hypothesis. Confirming Prediction 1, during practice a greater proportion of keyword shifts occurred for items learned with test-restudy practice versus restudy practice. Confirming Prediction 2, a greater proportion of keyword shifts occurred after retrieval failure trials versus retrieval success trials for items learned with test-restudy practice. However, results failed to support Prediction 3. Additionally, items in Experiments 1a and 1b had relatively few retrieval failure trials during practice, which limits the extent to which keyword shifting can occur.

The goals of Experiment 2 were three fold: Importantly, the first goal was to better evaluate Prediction 3 of the mediator shift hypothesis. The second goal was to increase the proportion of retrieval failures per item during practice by examining more difficult learning conditions. The third goal was to provide a stronger test of the mediator shift hypothesis by comparing two different test-restudy schedules of practice. I briefly describe the method of Experiment 2 here to highlight how it addresses each goal.

As in Experiments 1a and 1b, participants learned Swahili-English word pairs with restudy or test-restudy schedules of practice, and returned two days later for a final cued-recall test and a final keyword recall test for all items. Importantly, the format of the final keyword recall test in Experiment 2 was cue-only (as opposed to cue-target), which provides a better test of Prediction 3 because participants are required to retrieve
keywords from memory. Additionally, the practice phase of Experiment 2 involved either a short lag (9 intervening items) or a long lag (69 intervening items) between each next practice trial with a given item. Based on previous research showing slower learning with longer lags (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006 for a review on the spacing effect literature), I expected a greater proportion of retrieval failures during practice for the long lag condition compared to the short lag condition for the test-restudy group. If so, the mediator shift hypothesis predicts a greater proportion of keyword shifts during practice for the long lag condition compared to the short lag condition in the test-restudy group.

Concerning effects of lag on keyword use in the restudy group, the most conservative prediction from the mediator shift hypothesis is that no differences will exist in the proportion of keyword shifts during practice. Because items are not being retrieved from memory there would not be differences in retrieval failure during practice, which is presumably when keyword shifting occurs. However, a more liberal prediction (one that allows other mechanisms to influence metacognitive decisions during encoding) would allow for some effect of lag on keyword shifting in the restudy group.
METHOD EXPERIMENT 2

Participants and Design

Participants included 59 Kent State University undergraduates who did not participate in Experiments 1a or 1b. Lag (short versus long) was manipulated within participants. Practice group (test-restudy versus restudy) was manipulated between participants, with 29-30 participants randomly assigned to each group.

Materials

Items included 70 Swahili-English word pairs, with 60 target items and 10 filler items. The 60 target items included the 45 Swahili-English word pairs from Experiment 1b, with an additional 15 target Swahili-English word pairs, which were selected based on the pilot study described in Experiment 1. Items were randomly assigned to one of six practice lists. Three of these lists were assigned to the short lag condition, and the other three lists were assigned to the long lag condition, with list assignment to lag condition counterbalanced across participants. Ten fillers were used as primacy and recency buffer items.

Procedure

The general procedure was the same as Experiments 1a and 1b with the following exceptions. First, items received three restudy or test-restudy practice trials after their initial study and keyword report trials. Second, lag was a within participant
manipulation. Third, on the final keyword recall test, only the cue was provided to prompt recall of the keyword(s) for each word pair.

Given that one goal of the current experiment was to increase the proportion of retrieval failures during practice, Experiment 2 was designed to optimize lag differences. Previous research has shown that longer versus shorter lags lead to more difficult learning, and therefore will produce more retrieval failures during practice. Thus, with the goals of maximizing lag differences between conditions and keeping the length of the experiment manageable, the order of item presentation for the current experiment was blocked by list (see Appendix for detailed description of schedule of item presentation). Items in each of the three short lag lists were randomly sorted. The three lists assigned to the long lag condition were combined, and randomly sorted. The five items assigned to be primacy items had an initial study trial, immediately followed by an initial keyword report trial. Primacy buffer items then had one practice trial (either restudy practice or test-restudy practice depending on the group to which the participant was assigned). After practice trials for the primacy buffer items, target items received an initial study trial, immediately followed by an initial keyword report trial. Initial study and keyword report trials were followed by three practice trials (either restudy or test-restudy practice), in which each restudy trial was immediately followed by a keyword report trial, as in Experiment 1. As illustrated in Appendix, the only difference between short lag and long lag items was that items assigned to the short lag lists completed the initial study trial and all practice trials (i.e., practice trials 1-3) for items in a block before moving on to the next list of to-be-learned items. For items in the long lag lists, each practice trial was
followed by a block of trials for one set of short lag items. Upon completion of the last practice trial for long lag items, participants completed an initial study trial, initial keyword report trial, and one practice trial for the remaining five filler items to provide a recency buffer.

Results

*Final test performance.* Mean final test performance as a function of lag condition and practice group is reported in Figure 4. Results of a 2 (practice) x 2 (lag) mixed factor ANOVA showed a significant main effect of practice group, with higher levels of final test performance for items learned with test-restudy practice versus restudy practice, F(1,57) = 10.80, p = .002. Results also showed a significant main effect of lag, with higher levels of final test performance for items learned with a long lag versus a short lag between practice trials, F(1,57) = 108.69, p < .001. The interaction was not significant, F < 1.
Manipulation check. Before evaluating the extent to which the testing effect may be explained by keyword shifting during practice, I conducted a paired samples t-test to evaluate the extent to which the lag manipulation successfully increased the proportion of retrieval failures during practice for the test-restudy group. Results showed a significant difference between short lag and long lag conditions, with a greater proportion of retrieval failures during practice for items learned with a long lag versus a short lag, $t(28) = 9.61, p < .001$. Because the lag manipulation successfully led to differences in the
proportion of retrieval failures during practice, I now evaluate predictions of the mediator shift hypothesis.

*Keywords during practice.* To revisit, Prediction 1 states that a greater proportion of keyword shifts will occur for items learned with test-restudy practice versus restudy practice. The mean proportion of keyword shifts during practice was computed as in Experiments 1a and 1b and is reported in Figure 5. Results of a 2 (practice) x 2 (lag) mixed factor ANOVA showed a significant main effect of practice group, with a greater proportion of keyword shifts during practice for items learned with test-restudy practice versus restudy practice, F(1,57) = 8.91, \( p = .004 \). Results also showed a main effect of lag, with a greater proportion of keyword shifts for items learned with a long lag versus a short lag, F(1,57) = 24.51, \( p < .001 \). The interaction was not significant, F(1,57) = 2.14, \( p = .15 \).
To revisit, Prediction 2 states that a greater proportion of keyword shifts will occur after retrieval failure trials compared to retrieval success trials during practice for items learned with test-restudy practice. The proportion of keyword shifts as a function of retrieval status during practice is reported in Figure 6. Results of a 2 (lag) x 2 (retrieval status) repeated measures ANOVA showed a significant main effect of retrieval status, $F(1,28) = 66.48, p < .001$. Keywords were more likely to shift after retrieval failure trials versus retrieval success trials. The main effect of lag was not significant ($F$
< 2). The interaction was significant, $F(28) = 38.71, p < .001$, and will be discussed further in the General Discussion.

*Figure 6.* Mean proportion of keyword shifts during practice as a function of retrieval status during practice for conditions of the test-restudy group in Experiment 2.

*Keyword recall test.* Recall that one important goal of the current experiment was to provide a better test of Prediction 3 by implementing a cue-only prompt for the final keyword recall test. The mean proportion of keywords recalled on the final keyword recall test as a function of lag condition and practice group is reported in Figure 7. Consistent with predictions of the mediator shift hypothesis, results of a 2 (practice) x 2 (lag) mixed factor ANOVA showed a significant main effect of practice group, with a
greater proportion of keywords recalled on the final keyword recall test for items learned with test-restudy practice versus restudy practice, $F(1,57) = 10.44, p = .002$. Results also showed a main effect of lag, with a greater proportion of keywords recalled for items learned with a long lag versus a short lag, $F(1,57) = 78.79, p < .001$. The interaction was not significant, $F < 1$. These results are consistent with higher levels of final test performance for items learned with test-restudy practice versus restudy practice, as well as higher levels of final test performance for items learned with a long lag versus a short lag.

Figure 7. Mean proportion of keywords recalled on the final keyword recall test as a function of practice group and lag in Experiment 2.
Encoding time. As in Experiment 1, encoding time was greater for the test-restudy group than for the restudy group [mean encoding time 115.93 (SE = 6.69) versus 84.12 (SE = 5.01)], t(57) = 3.79, p < .001. Experiment 3 was designed to rule out encoding time as an explanation for the memorial benefits of testing. To do so, in addition to the groups from Experiment 2, Experiment 3 included test-restudy and restudy groups for which encoding time was equated for test and restudy trials during practice.

In sum, results replicated the pattern of results from Experiments 1a and 1b, supporting Prediction 1 and Prediction 2 of the mediator shift hypothesis. Importantly, when a more appropriate cue-only keyword recall test was utilized, results also supported Prediction 3 of the mediator shift hypothesis; a greater proportion of keywords from Session 1 were recalled on the final keyword recall test for items learned with test-restudy practice versus restudy practice.

In addition to supporting each prediction of the mediator shift hypothesis, the lag manipulation provided a stronger test of the mediator shift hypothesis by comparing two test-restudy conditions, which were predicted to differentially influence final test performance and keyword shifting during practice. Replicating previous spacing effect research, a greater proportion of items were correctly recalled on the final cued-recall test for items learned with a long lag versus a short lag. Importantly, the pattern of results for keyword shifting during practice and keyword recall on the final keyword recall test were in line with predictions from the mediator shift hypothesis. Specifically, a greater proportion of keyword shifts occurred during practice for items learned with a long lag versus a short lag. Similarly, a greater proportion of keywords from Session 1 were
recalled on the final keyword recall test for items learned with a long lag versus a short lag. These results are consistent with higher levels of final test performance for the long lag condition compared to the short lag condition in the test-restudy group.

Similar effects of lag on keyword shifting during practice and keyword recall on the final keyword recall test obtained in the restudy group, which is not predicted by the mediator shift hypothesis. Although the pattern of results was similar for conditions of the restudy and test-restudy groups, keyword shifting and keyword recall was greater for items learned with test-restudy practice versus restudy practice. These results suggest the possibility that the same mechanism may influence keyword shifting and keyword recall for test-restudy and restudy groups. However, results also demonstrate that an additional factor influences keyword shifting and keyword recall for items learned with test-restudy practice, retrieval failures during encoding. Most important, test-restudy practice (versus restudy practice) led to the formation of more effective keywords during practice, which benefitted later recall of target items.
INTRODUCTION EXPERIMENT 3

Taken together, results from Experiment 2 support predictions of the mediator shift hypothesis as an explanation for the memorial benefits of testing. Specifically, retrieval failure trials during test practice afford an evaluation of the effectiveness of keywords generated during encoding, whereas restudy practice does not. The generation of more effective keywords facilitates later retrieval of targets.

Experiment 3 was designed to rule out encoding time as an alternative explanation for the testing effect. Experiment 3 included two additional groups for which encoding time during practice trials was equated for test-restudy and restudy groups. If encoding time underlies the memorial benefits of testing observed here, final test performance will be similar for the test-restudy and restudy groups for which encoding time was equated. In contrast, the mediator shift hypothesis predicts that final test performance will be greater for the test-restudy group compared to the restudy group even when encoding time is equated.
METHOD EXPERIMENT 3

Participants and Design

Participants included 98 Kent State undergraduates who did not participate in any of the previous experiments. The design of Experiment 3 was the same as Experiment 2 with the following exception. Experiment 3 also included experimenter-paced test-restudy and restudy groups, for which encoding time was equated. Thus, Experiment 3 was a 2 (lag: short versus long) x 2 (practice: test-restudy versus restudy) x 2 (pacing: self versus experimenter) design, with 22-27 participants randomly assigned to each group.

Materials

Materials were the same as those used in Experiment 2.

Procedure

For all four groups, initial study and all keyword report trials were self-paced to ensure that participants had sufficient time to report keywords during practice. Practice trials in the self-paced test-restudy group (hereafter referred to as SP test-restudy) and the self-paced restudy group (hereafter referred to as SP restudy) were the same as for the test-restudy and restudy groups in Experiment 2.

For practice trials in the experimenter-paced groups (hereafter referred to as EP test-restudy and EP restudy), I used the encoding time results from Experiment 2 to determine the amount of time allotted for each trial during practice. For trials 1, 2, and 3
in Experiment 2, mean test-restudy time plus one standard deviation was 11.17, 7.15, and 6.59 seconds, respectively. Accordingly, participants in the EP groups in Experiment 3 were given 11.17, 7.15, and 6.59 seconds for practice trials 1, 2, and 3.

Participants in the EP restudy group were instructed that the computer would advance to the next trial when the appropriate amount of time had elapsed. Participants in the EP test-restudy group were instructed that they had a certain amount of time to complete both test and restudy trials. A button was provided on the bottom of the screen for test trials so that participants could advance to the study trial when they had finished retrieving an item during a test trial. Thus, participants in the EP test-restudy group could advance the screen to restudy trials, but once they were on the study screen they had to wait until the computer advanced them to their strategy report trial for that item.

Results and Discussion

*Final test performance.* Mean final test performance as a function of lag condition, practice group, and pacing group is reported in Figure 8. Results of a 2 (lag) x 2 (practice) x 2 (pacing) mixed factor ANOVA showed a significant main effect of practice group, with higher levels of final test performance for items learned with test-restudy practice versus restudy practice, F(1,94) = 37.47, p < .001. Results also showed a significant main effect of lag, with higher levels of final test performance for items learned with a long lag versus a short lag, F(1,94) = 165.75, p < .001. Inconsistent with an encoding time interpretation of the testing effect, neither the main effect of pacing nor any interaction was significant. Furthermore, results of an independent samples t-test showed significantly higher levels of final test performance for the EP test-restudy group
versus the EP restudy group, $t(50) = 3.56, p < .001$. Taken together, results showed that encoding time did not influence final test performance, but testing during practice did influence final test performance, with higher levels of final test performance after test-restudy practice versus restudy practice.

To evaluate the extent to which keyword shifting during practice influenced the memorial benefits of test-restudy practice, I now evaluate each prediction of the mediator shift hypothesis.

Figure 8. Mean proportion of items correctly recalled on the final test as a function of group and lag condition in Experiment 3.
Keywords during practice. The proportion of trials for which participants shifted to new keywords during practice is reported in Figure 9. Results of a 2 (lag) x 2 (practice) x 2 (pacing) mixed factor ANOVA showed a significant main effect of lag, with a greater proportion of keyword shifts for items learned with a long lag versus a short lag, $F(1,93) = 38.59, p < .001$. The main effect of practice group approached significance, with a greater proportion of keyword shifts for items learned with test-restudy practice versus restudy practice, $F(1,93) = 2.94, p = .09$. The main effect of pacing was not significant ($F < 1$), and no interaction was significant, $Fs < 1.9$.

![Figure 9](image.png)

*Figure 9.* Mean proportion of keyword shifts during practice as a function of group and lag condition in Experiment 3.
Why did the main effect of practice group fail to reach significance? Visual inspection of Figure 9 suggests that unexpectedly low levels of keyword shifting in the EP test-restudy group may be responsible. Supporting this possibility, results of a 2 (lag) x 2 (practice) mixed factor ANOVA for just the SP groups replicated the pattern of results from Experiment 2, with a main effect of practice group and lag condition, \( F(1,44) = 156.64, p < .001 \) and \( F(1,44) = 15.28, p < .001 \). That is, a greater proportion of keyword shifts occurred for items learned with test-restudy practice versus restudy practice and for items learned with a long lag versus a short lag. However, this pattern did not obtain for the EP groups. Results of a 2 (lag) x 2 (practice) mixed factor ANOVA showed a main effect of lag \( [F(1,49) = 24.06, p < .001] \), but did not show a main effect of practice group, \( F < 1 \). One potential explanation for why the proportion of keyword shifts was lower for the EP test-restudy group compared to the SP test-restudy group is that the encoding time I imposed (i.e., 1 standard deviation above the mean) was too short for participants in the EP group to complete both test and restudy trials. This would be especially problematic for retrieval failure trials. For example, a participant might spend time trying to retrieve an item, fail to retrieve the item, and then advance to the study trial for that item. However, at this point, much of their time may have been used trying to retrieve the item, leaving minimal time to restudy the word pair (which is presumably when new keywords are formed). Thus, I may have unintentionally disrupted encoding of new keywords for this group on restudy trials during practice.

Prediction 2 states that a greater proportion of keyword shifts will occur after retrieval failure trials compared to retrieval success trials during practice for items
learned with test-restudy practice. The mean proportion of keyword shifts as a function of retrieval status during practice is reported in Figure 10. Results of a 2 (lag) x 2 (retrieval status) x 2 (pacing) mixed factor ANOVA showed a significant main effect of retrieval status, with a greater proportion of keyword shifts after retrieval failure trials versus retrieval success trials, $F(1,45) = 96.49, p < .001$. The main effect of lag was also significant, with a greater proportion of keyword shifts for items learned with a short lag versus a long lag (this outcome will be discussed further in the General Discussion), $F(1,45) = 5.48, p = .02$. The main effect of pacing was not significant, $F(1,45) = 2.43, p = .13$. The lag X retrieval status interaction was significant, $F(1,45) = 87.81, p < .001$, and the retrieval status X pacing interaction approached significance, $F(1,45) = 3.96, p = .05$, respectively. No other interaction was significant, $Fs < 2$. 
Figure 10. Mean proportion of keyword shifts during practice as a function of retrieval status during practice for conditions of the test-restudy group in Experiment 3.

Keyword recall test. Prediction 3 states that a greater proportion of keywords from Session 1 will be recalled on the final keyword recall test for items learned with test-restudy practice versus restudy practice. The mean proportion of keywords recalled on the final keyword recall test as a function of lag condition, practice group, and pacing group is reported in Figure 11. Results of a 2 (lag) x 2 (practice) x 2 (pacing) mixed factor ANOVA showed a greater proportion of keywords recalled for items learned with test-restudy practice versus restudy practice, F(1,94) = 4.10, p = .05. The main effect of
lag showed a greater proportion of keywords recalled for items learned with a long lag versus a short lag, \( F(1,94) = 120.21, p < .001 \). For presentation control, a greater proportion of keywords were recalled for the experimenter-paced group versus the self-paced group, \( F(1,94) = 3.74, p = .06 \). The lag X pacing interaction was significant, \( F(1,94) = 4.19, p = .04 \). No other interaction was significant, \( Fs < 1 \).

Figure 11. Mean proportion of keywords recalled on the final keyword recall test as a function of group and lag condition in Experiment 3.

Encoding time. Mean encoding time was 74.15 (SE = 4.03), 117.19 (SE = 4.43), 122.77 (SE = 6.17), and 136.67 (SE = 5.87) minutes for the SP restudy, EP restudy, SP test-restudy, and EP test-restudy groups, respectively. Results of a 2 (practice) x 2 (pacing) between participant ANOVA showed significant main effects of practice group and pacing group, with significantly longer encoding times for items learned with test-
restudy practice and for the experimenter-paced group, $F(1,94) = 42.03, p < .001$ and $F(1,94) = 29.39, p < .001$. Results also showed a significant interaction, $F(1,94) = 7.70, p = .01$. Even though encoding time was equated for test and restudy trials for the experimenter-paced group, participants had self-paced strategy trials during practice. Results of an independent samples t-test for the EP test-restudy group and the EP restudy group showed significantly longer encoding times for strategy report trials for items learned with test-restudy practice versus restudy practice (which explains the overall longer encoding time for the EP test-restudy group versus the EP restudy group), $t(49) = 2.67, p = .01$. These results suggest that when encoding time was equated, participants in the EP test-restudy group may not have had enough time on study trials to encode items, and therefore spent more time on keyword report trials (perhaps because they were trying to generate new keywords).

Although one goal of Experiment 3 was to equate encoding time for the EP test-restudy group and the EP restudy group, results show that this manipulation was not completely successful (functional encoding time differences were reduced but not eliminated). It is possible that if I had used a longer encoding time (e.g., the mean plus two standard deviations), differences between these groups would not have emerged. Importantly, even though the pacing manipulation did not completely equate functional encoding time for these groups, the pattern of results for final test performance suggest that encoding time cannot entirely explain the memorial benefits of testing. Specifically, the SP restudy group and the EP restudy group showed similar levels of final test performance even though encoding time was greater for the EP versus the SP restudy
group. Similarly, for the EP test-restudy group and the SP test-restudy group, final test performance was similar even though encoding time was greater for the EP versus the SP test-restudy group.
GENERAL DISCUSSION

Across a series of four experiments, I tested the mediator shift hypothesis for explaining the memorial benefits of testing. According to the mediator shift hypothesis, final test performance is greater after test-restudy practice versus restudy practice because individuals monitor and modify mediators when they cannot correctly recall an item during practice. Individuals engaged in restudy practice cannot evaluate the effectiveness of mediators as well as individuals engaged in test-restudy practice because they do not experience retrieval failure during encoding. Across experiments, three predictions of the mediator shift hypothesis were confirmed. Supporting Prediction 1, a greater proportion of keyword shifts occurred during practice for items learned with test-restudy practice versus restudy practice. Supporting Prediction 2, a greater proportion of keyword shifts occurred after retrieval failure trials versus retrieval success trials during practice for items learned with test-restudy practice. Supporting Prediction 3, a greater proportion of keywords from Session 1 were recalled on the keyword recall test in Session 2 for items learned with test-restudy practice versus restudy practice in Experiments 2 and 3.

Thus, the current experiments demonstrated that keywords were more likely to shift during practice for items learned with test-restudy practice versus restudy practice, presumably because retrieval failures during practice allowed individuals to evaluate the effectiveness of keywords and to shift to more effective ones. These more effective keywords facilitated recall of targets on the final retention test. To revisit, previous
research has demonstrated that the effectiveness of mediators depends on at least two factors, mediator retrieval and mediator decoding (e.g., Dunlosky et al., 2005). For a mediator to be effective it must be retrieved when prompted by a cue on a test trial. Additionally, the mediator must be decoded to elicit the correct target response.

Pyc and Rawson (in press) provided evidence for a contribution of both of these factors to the memorial benefits of testing. In brief, they tested the mediator effectiveness hypothesis, which states that mediators generated during test-restudy practice (compared to restudy practice) are more likely to be retrieved and decoded, which facilitates recall of target answers on a retention test for items learned with test-restudy practice versus restudy practice. Pyc and Rawson (in press) had participants learn Swahili-English word pairs using the keyword encoding strategy and a procedure very similar to the one used in the current experiments. First, participants had an initial study and initial keyword report trial for each item. Participants then engaged in either test-restudy or restudy practice. Each restudy trial was immediately followed by a keyword report trial. Participants returned one week later for a final retention test.

As in prior research, participants in one group were prompted with only the cue word for each item on the final test, and a greater proportion of targets were correctly recalled for items learned with test-restudy practice versus restudy practice. Importantly, results supported predictions from the mediator effectiveness hypothesis. Concerning mediator retrieval, participants in a second group were prompted with a cue, but were asked to retrieve their own mediator (i.e., keyword) for a given item prior to recalling the target answer. A greater proportion of mediators were recalled for items learned with
test-restudy practice versus restudy practice. Concerning mediator decoding, participants in a third group were prompted with a cue, but were also provided with the last mediator (i.e., keyword) that they had generated for a given item in Session 1. A greater proportion of targets were correctly recalled for items learned with test-restudy practice versus restudy practice. That is, when differences in mediator recall were eliminated (because participants were provided with their own mediators), mediators generated in the test-restudy group (versus the restudy group) were more likely to elicit the correct target answer. These results demonstrate that mediators generated during encoding are more likely to be retrieved and decoded after test-restudy practice compared to restudy practice.

In addition to identifying factors that influence mediator effectiveness, the mediator effectiveness hypothesis may be able to explain the unexpected pattern of results for the final keyword recall test in Experiment 3. To revisit, the EP restudy group and the two test-restudy groups recalled a similar proportion of keywords on the final keyword recall test, which is not predicted by the mediator shift hypothesis. However, also note that final test performance was greater for items learned with test-restudy practice versus restudy practice. Taken together, these results suggest that although mediator retrieval may have been similar for the EP restudy and test-restudy groups, mediator decoding may have been deficient in the EP restudy group.

The mediator shift hypothesis compliments the mediator effectiveness hypothesis by stating that more effective mediators arise at least in part from shifting mediators after retrieval failure trials during practice. Taken together, these two hypotheses suggest that
mediator use during test practice is one theoretical explanation for the memorial benefits of testing. That is, shifting mediators after retrieval failure trials during practice leads to the generation of more effective mediators (i.e., mediators that are more likely to be recalled and decoded), which subsequently facilitates retrieval of target answers.

Although the current experiments provided support for predictions of the mediator shift hypothesis, here I discuss two potential limitations of these experiments. First, the pacing manipulation in Experiment 3 was not completely successful at equating functional encoding time for test-restudy and restudy groups. Encoding time results for the EP test-restudy group suggest that participants did not always have enough time to retrieve and restudy items. Thus, future research should consider allowing more time for practice trials (e.g., using the mean plus two standard deviations) when equating test-restudy and restudy practice groups. Although differences in encoding time emerged for the EP test-restudy and EP restudy group, the pattern of results for final test performance suggest that encoding time did not influence the memorial benefits of testing. Additionally, previous research has consistently demonstrated that encoding time does not influence the memorial benefits of test practice versus restudy practice (e.g., Carrier & Pashler, 1992).

Second, the particular way that I instructed participants to use the keyword encoding strategy could have limited the extent to which keywords could shift during practice. Participants were instructed to generate an English word (i.e., keyword mediator) that sounded like the Swahili cue and then to semantically relate the keyword to the English target. Constraining participants to semantically relate keywords to
English targets may have limited the number of keyword mediators participants could generate for a given item during encoding because it may be difficult to generate keywords that sound like the Swahili word and that are semantically related to the English target. By contrast, instructing participants to generate an English keyword but then form an interactive image between the English keyword and the English target may have afforded a greater number of viable mediators (and hence more shifting) during practice. Almost any two English words can be formed into an interactive image, whereas it is much more difficult to semantically relate any two English words.

Consistent with the possibility that the participants were limited by the specific keyword strategy I instructed them to use, consider keyword shifting as a function of retrieval status for the test-restudy practice groups in Experiments 2 and 3. For illustrative purposes, consider results from Experiment 3 (Figure 10); replicating previous research, learning was more difficult for items learned with a long lag versus a short lag between practice trials (mean proportion of retrieval failures during practice was .43 compared to .24). Although results supported Prediction 2 in that a greater proportion of keyword shifts occurred after retrieval failure versus retrieval success trials, the proportion of keyword shifts after retrieval failure trials was lower for items learned with a long lag versus a short lag. If a greater proportion of retrieval failures occurred for the long lag versus the short lag condition, why was the proportion of keyword shifts lower for retrieval failure trials in the long lag condition versus the short lag condition? One plausible explanation is that a limited number of keywords could be generated for a given word pair. As the number of times an item was incorrectly recalled increased, it would
become increasingly difficult to come up with a new (more effective) keyword. This explanation is in line with the greater proportion of keyword shifts after retrieval failure trials for items learned with a short lag versus a long lag. Thus, an important direction for future research is to evaluate the efficacy of the mediator shift hypothesis using variations of the keyword encoding strategy implemented here. For example, future research could use a keyword encoding strategy in which individuals generate keywords but instead of being instructed to semantically relate the English mediator and English target, individuals are instructed to form an interactive image (a normatively effective encoding strategy; e.g., Richardson, 1998) between the English mediator and the English target.

More generally, future research needs to explore other theoretical mechanisms underlying the testing effect because very few studies have evaluated theoretical mechanisms underlying the memorial benefits of testing. Furthermore, researchers have proposed two different kinds of effect that testing can have on memory, which could have different underlying mechanisms. Specifically, Roediger and Karpicke (2006) have discussed the indirect effects of testing and direct effects of testing. The indirect effects of testing are related to the way in which tests can influence subsequent practice. Specifically, the outcome of testing influences and guides future practice and learning. In contrast to the indirect effects of testing, the direct effects of testing are related to the act of testing itself. Specifically, retrieving a piece of information from memory modifies and enhances memory.
The mediator shift hypothesis provides one theoretical explanation for the indirect effects of testing, in that the outcome of the practice tests presumably improved memory by influencing the generation of mediators during the subsequent restudy opportunity. Although results of the current experiments confirmed predictions of the mediator shift hypothesis, it is likely not the only theoretical explanation for the indirect effects of testing. Additionally, mediator use could also underlie the direct effects of testing. Specifically, test practice (compared to restudy practice) may be beneficial for memory not only because tests increase the memory strength of targets, but test practice could also increase the memory strength of mediators. If test practice increases the memory strength of mediators as well as targets, mediator strength could be one theoretical explanation for the direct effects of testing. Therefore, it is important for future research to continue evaluating the extent to which mediator use influences the direct and indirect effects of test practice.

Many researchers are interested in the memorial benefits of testing because of the important implications of testing for student learning and scholarship. In contrast to many theory-based fields of research, the testing effect literature is largely empirical, which may limit the extent to which researchers can identify the most beneficial schedules of learning. That is, instead of evaluating schedules of practice based on factors that are theoretically predicted to lead to differences in performance, many researchers simply manipulate one or more factors that they believe will influence learning (e.g., number of practice trials or spacing of practice trials). Given that very few studies have evaluated theory in the testing literature, future theoretical research will not
only improve understanding of the memorial benefits of testing, but may also enhance the educational implications of this research.

To conclude, the current experiments evaluated and confirmed predictions of the mediator shift hypothesis for testing effects. These results suggest that testing is beneficial for memory because testing during practice affords an evaluation of the effectiveness of mediators. On retrieval failure trials during practice, individuals evaluate the effectiveness of mediators and shift from less effective mediators to more effective mediators, which facilitates retrieval of information on a later retention test. Most important, the current studies support one theoretical explanation for the memorial benefits of testing in a literature that is empirically well-established but theoretically not well-understood.
REFERENCES


### Schedule of item presentation in Experiment 2 and Experiment 3.

<table>
<thead>
<tr>
<th>Items</th>
<th>Block of Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB1</td>
<td></td>
</tr>
<tr>
<td>PB2</td>
<td></td>
</tr>
<tr>
<td>RB1</td>
<td></td>
</tr>
<tr>
<td>RB2</td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td></td>
</tr>
<tr>
<td>B4</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td></td>
</tr>
</tbody>
</table>

Note: Number in subscript refers to trial number, in which 1 = initial study and initial keyword trials, and 2-4 = subsequent practice trials (either restudy or test-restudy). Letter A-D refers to set of items. PB = primacy buffer items; RB = recency buffer items.