UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

IMPROVING STUDY STRATEGIES THROUGH TESTING EXPERIENCE

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

MARISA CRISOSTOMO
Norman, Oklahoma
2017

IMPROVING STUDY STRATEGIES THROUGH TESTING EXPERIENCE

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

_____
Dr. Daniel Kimball, Chair

_____
Dr. Scott Gronlund

_____
Dr. Jorge Mendoza

_____
Dr. Robert Terry

_____
Dr. Sepideh Stewart

Table of Contents

List of Tables

List of Figures

Abstract

Prior research on the Testing Effect has shown that testing previously learned material can result in greater long term memory performance than if the material was restudied (Roediger & Karpicke, 2007). Although the testing can be used as a potent learning event, students seldom use testing as a main study strategy. The following studies attempted to determine if students' study decisions can be influenced by experience of two different study strategies, one more optimal than the other, such that students choose to retrieve more often. The present experiments had participants either retrieve or restudy cue-target word pairs then make a future study decision about whether to retrieve or restudy items. The results suggest that differential experience of study strategies is not necessary to improve study decisions and that study decisions may be resistant to change.

Introduction

Because a majority of students' study time is spent outside of the classroom without instructor direction, it is increasingly important that students make study decisions on their own that will benefit their performance the most. After students learn in the classroom and are tasked with preparing for an upcoming exam, they must make several study decisions such as how long to study, what to study, when to stop studying, and how to study (Kornell & Bjork, 2007). The present research will primarily focus on the last question of 'how to study, or more specifically, whether a student should test oneself while studying.

Prior research has strongly demonstrated that testing over previously studied material can be used as a means of assessment as well as a means of learning (Roediger & Karpicke, 2006). Instructors and students are likely familiar with the idea that testing can be used as a way to assess how well one knows the material after learning is complete, but *the testing effect* refers to the phenomenon that taking a test following learning the material results in better performance on a final test than when the same information is just presented again (Roediger, Putnam, & Smith, 2011). Current theories of the testing effect suggest that an initial test following study allows the learner to engage in the same effortful retrieval processes that will be used at final testing. The testing advantage has been demonstrated under a number of conditions, including with and without corrective feedback (Cull, 2000; Carpenter & Delosh, 2005). Thus, testing, which is also referred to as retrieval practice, can be used by both instructors in the classroom and students during self-study as a way to enhance long term memory. Although testing oneself has been shown to be a superior study technique, previous

research has also shown that students often choose not to test themselves. This dissertation will review the research that suggests that students are unable to make use of the benefits of testing. In addition, we will discuss the theoretical explanations as to how students change their study decisions. Furthermore, we will discuss two experiments to evaluate whether the experience of the benefits of the testing effect can influence learners to choose the more effective study strategy: testing over restudy.

Awareness of the Testing Effect

Students must rely on their own metacognition to make decisions on what is the best method to use while studying. When monitoring judgments are inaccurate, this can lead to poorer study decisions and later performance (Nelson and Dunlosky, 1994).

Through surveys of undergraduate students, researchers have found that the majority of students prefer rereading and do not use testing as a method studying. When students list their study strategies, Karpicke, Butler and Roediger (2009) found that 84% of students used rereading notes or the textbook, and 55% of students surveyed listed it as their #1 strategy of choice. They also found that when given the option to either restudy or test themselves, only 18% of students indicated that they would rather test themselves, compared to 57% of students who indicated that they would rather restudy.

Although some students do report using testing while studying, it seems that the majority of students use testing as a form of assessment, and are not aware of its benefits as a study method. When questioned why students test themselves, Kornell and Bjork (2007) found that 68% of students do so to assess what content they do and do

not know, and only 18% viewed testing as a method of learning. Kornell and Son (2009) found a similar pattern, with 66% of students reporting that that they use testing as a way to assess what information has been learned, and only 20% reporting that they test themselves because they learn more with testing than with re-presentation. Increased fluency has been used to explain this restudy preference. When one rereads the text, the fluency or ease of processing of the text is higher than if one were to be tested over the material. Students tend to base their assessments of learning as well as study decisions based upon this fluency. The ease of processing, however, has been shown to be a poor diagnostic tool of future performance (Nelson & Dunlosky, 1992). Koriat and Bjork (2005) would call this the *curse of knowledge* and would also agree that the presence of all information that learners are required to recall later makes it harder to accurately predict future performance.

Several studies have also used judgments of learning (JOLs) to assess the extent to which students are aware of testing benefits. Judgments of learning typically ask learners to predict how well they will remember something in the future or for a test. Kornell and Son (2009) had participants either restudy word pairs or be tested over them and then asked participants to predict how they would do on a later exam. Although there was a testing advantage, participants consistently predicted that they would perform better on words that had been shown again compared to word pairs that had been tested twice. Similarly, Roediger and Karpicke (2006) found that participants consistently predicted that they would do better on a later exam if they re-read a passage several times versus reading the passage once and being tested over it many times.

Agarwal, Karpicke, Kang, Roediger, and McDermott (2008) found the same pattern of learners' predictions with a cued recall test.

Choosing Study Strategies

The lack of awareness of more effective study strategies does not seem to be exclusive to testing. A large body of research has found that most learners time and time again fail to choose more effective study strategies. Desirable difficulties, such as testing and spacing, are tasks that are introduced during the learning process that create a more difficult study episode, but results in greater long-term retention (Bjork, 1994). The spacing effect is the finding that information is better recalled at a later time if its presentations are spaced out versus in succession (blocked). Kornell and Bjork (2008) had participants study category exemplars. Some category exemplars were presented three times in a row (massed/blocked practice), while other category exemplars were interspersed throughout the presentation of pairs (spaced practice). Though participants' ability to identify the category to which an exemplar belonged was better when category exemplars were spaced versus blocked, participants still believed that a blocked presentation would result in better performance on the final test. Additionally, in a study done by Kornell, Castel, Eich, and Bjork (2010), participants were presented with paintings for different artists and were instructed to learn each artist's style. For some artists, their paintings were presented as a massed presentation, and for other artists, their paintings were dispersed throughout the presentation of paintings. At final test, participants were asked to identify the artist for paintings that had not been previously presented. Although a strong spacing advantage was found, with participants correctly

identifying the paintings for which the artist's paintings were dispersed, participants still chose massed presentation when given the option.

Although the majority of studies do seem to show that students are largely unaware of testing benefits and other more effective strategies, there is some body of research to show that students can learn to use better study methods. More recent studies have examined participants' abilities to adopt superior encoding strategies with the generation effect, another desirable difficulty. The generation effect, which is the finding that producing an answer, such as unscrambling an anagram or actively guessing a target word, results in better performance on a later test compared to merely the presentation of the information. DeWinstanley and Bjork (2004) employed a fill-in-the-blank procedure in which participants read a passage. Some of the key terms were merely underlined and participants simply read them, or they were required to generate other key word using context clues. They found that when participants who experienced the generation advantage by taking a test, they were able to improve encoding strategies on a second passage. Similarly, Storm, Hickman, and Bjork (2016) also found that participants were able to improve their encoding strategies but only after a test. Simply experiencing generation without a test, and just reflecting on the differences between generated and read items, or reading about the generation advantage was not enough to significantly improve encoding strategies. These studies support Finley and Benjamin's (2012) the test-expectancy paradigm, which proposes that in order for participants to be able to employ a more effective encoding method, learners must have knowledge of the type of test and what type of information is needed for the test.

There is also some evidence within the testing effect paradigm to show that participants may choose testing over restudy under certain conditions. Kornell and Son (2009) presented participants with cue-target word pairs and allowed participants to choose a study mode between pair mode, in which pairs were presented again, or test mode, in which participants were presented with just the cue and tested themselves. Kornell and Son (2009) found that participants developed a preference for the test mode, choosing the test mode more frequently than the pair (restudy) mode as exposure to the cue-target word pairs increased. One might point to this preference for test mode as evidence that learners are aware of the memorial effects of testing, however, when asked for JOLs, participants still believed that pair mode would result in a higher score on the final test compared to the Test Mode. Although, participant study choices seemed to show that they were aware of the testing effect, their metacognitive beliefs still favored re-presentation. Students' decisions to increase test trials was likely not due to a change in metacognitive beliefs that testing is a superior study method, but rather that students were using testing as a form of assessment, A study conducted by Einstein, Mullet, and Harrison (2012) concluded that they were able to change student study habits in a real classroom setting. Students enrolled in a cognitive psychology course participated in a procedure similar to Roediger and Karpicke's (2006) lab experiment in which they were asked to re-read or be tested over passages. Students were then given a test and presented with the rationale of the testing effect and findings for the experiment. At the end of the semester, 62% of students reported that they now incorporate testing much more into their study habits compared to the beginning of the semester. 91% of students correctly answered a multiple choice question about the

benefits of self-testing, compared to only 36% at the beginning of the semester. Einstein et al. (2009) concluded that students showed an understanding of the testing effect and a change in study habits, but we would argue that these findings could be mainly the result of demand characteristics considering students were in a cognitive psychology course.

Other research regarding the effects of the testing effect on subsequent learning of repeated material has also been explored, and it has been found that testing can promote future learning of related or similar material either through the reduction of proactive interference or increased attention (Chan, McDermott & Roediger, 2006; Pastotter, Schicker, Niedernhuber, & Bauml, 2011). Essentially, after being tested, participants likely increase attention towards items that were previously tested when shown again. Although this type of testing experience has been shown to have positive effects, it still does not address future study decisions.

Improving Metacognition

An important component in the discussion of study decisions is metacognition. If a learner cannot monitor one's metacognition accurately, then we should not be surprised that the majority of students do not make the most optimal study decisions because they cannot identify when strategies are effective. Tullis, Finley, and Benjamin (2013) conducted a study exploring the testing effect and JOLs. Participants learned 64 word pairs, practiced retrieving half of the targets and restudied the intact word pair for the other half of targets. After a delay (or no delay), participants were asked to predict

how likely they would be able to recall each item on a later test, and then participants were tested over all word pairs. Tullis et al. (2013) found that participants predicted that words that had previously been retrieved would be more likely to be remembered than restudied items (demonstrating participants' knowledge of the testing advantage), but only when JOLs were made after a delay and with the cue only. However, when presented with the cue and the target, JOLs did not accurately reflect the testing effect. Rather than this being evidence that learners are sensitive to the benefits of testing, it is likely that words that were tested were better remembered (as a result of the testing effect), and when asked to make a JOL after a delay, participants rated these better remembered words with a higher JOL (Dunlosky & Nelson, 1992). Furthermore, Tullis, Finley, and Benjamin's(2013) second experiment asked participants , "From the word pairs that you restudied/ were tested on, how many words will you remember tomorrow?" Even with the indication that restudied and tested words should be recalled at different rates, participants' JOLs still did not accurately reflect the testing effect, and participants predicted that they would remember more restudied items than tested items. Restudied items were also rated as more likely to be remembered even when participants were given an indication that their answer was correct and if it had been previously restudied or retrieved. It was not until participants had received full feedback, which included an indication as to whether a single item had been retrieved or restudied as well as a display of their global score with the number of correct items answered at the end of the final test that participants were able to accurately predict that tested items would be better remembered.

Relevant Theories Regarding Study Decisions

Study decisions can be influenced by a multitude of factors other than metacognitive accuracy. For example, intrinsic motivation (Pintrich, 2000), interest in the content (Son & Metcalfe, 2000), an item's value or reward (Ariel, 2013), or even the probability that an item will appear on an exam (Ariel, Dunlosky & Bailey, 2009) all can affect study decisions. Dunlosky and Hertzog (1998) proposed the *discrepancy reduction theory* to explain how students make study decisions. It states that study decisions depend on the perceived state of learning (current state) and how it compares to their desired state (goal state), which can be the ability to remember that information at a later time or immediately. Study decisions are then made to reduce the difference between the current state and the desired state. In the context of the testing effect, if learners' perceived state is very close to their goal state due to fluency and ease of processing during restudy trials, it is not surprising that learners consistently choose to restudy over retrieve because restudy trials quickly demonstrates that their perceived state to match their goal state. The *region of proximal learning* (Metcalfe, 2011) is also often discussed in the context of study decisions. This model proposes that students will make study decisions based on one's individual assessment of whether the item can or cannot be learned given the current situation. In other words, learners do not choose, or devote less time, to study items that feel they already know or items that they feel are too difficult or "out of reach" (Kornell & Metcalfe, 2006). The *agenda-based regulated framework* (Ariel et. al, 2009) is quite similar to the *discrepancy reduction theory* in that it involves the identification of a goal, but it also proposes that learners construct

agendas in response to environmental conditions to achieve those goals. The environmental conditions can include factors such as deadlines, interest in the material, motivation, and reward, and thus the Agenda-Based Framework can explain how these factors can impact study decisions because they are taken into account in one's agenda. Although these are models that can be useful in explaining why students make certain study decisions, especially why they tend to select those that are suboptimal, the present study is more interested in how study decisions can improve through experience.

Another framework that can be useful in understanding the adoption more effective study strategies is Dunlosky and Hertzog's (2000) *updating knowledge framework*. They outline four components necessary for updating study strategies: effectiveness, monitoring, updating, and utilization. Learners must be able to experience differential effectiveness between different study strategies, and learners must be able to accurately monitor these differences in effectiveness. Furthermore, learners must then connect the differences to the study strategies and update their knowledge of a given study strategy. Lastly, learners must utilize the more effective study strategy when it is appropriate. The present experiment makes use of this framework to improve study strategies. We reasoned that if participants are able to experience the testing effect within a single study episode and are given indications as to the result of different study strategies as an aid for recognizing the differential benefits in retrieval practice compared to restudying, participants would be able to attribute improved performance to the testing strategy, and therefore choose to test oneself more often.

Present Studies

Although memory researchers boast about the robust testing effect as a highly effective study strategy, and yet that students do not typically prefer to test themselves, prior research has not adequately explored whether students can choose testing over restudying if they are shown its benefits. The present experiments explored whether mere experience of the testing effect and its explicit benefit are enough to influence study strategies. Although Tullis , Finley, and Benjamin (2013) found that learners' metacognitive beliefs can accurately reflect the benefits of testing when given a clear indication of their performance, it is unclear whether these can affect future study decisions. Furthermore, while deWinstanley and Bjork (2004) and Storm, Hickman, and Bjork (2016) have demonstrated that learners can improve encoding strategies through the experience of the generation effect, it is also unclear whether experience of the testing effect can influence strategy decisions. Experiment 1 employed a one list procedure and Experiment 2 used a two list procedure to determine if experience of the testing effect can influence future study decisions. We predicted that learners who have the opportunity to experience the benefits of testing will choose retrieval practice over restudy.

Experiment 1

In Experiment 1, participants were presented with a list of 30 to-be-learned cue-target word pairs. Participants experienced a study phase in which they were shown the

words again or practiced retrieving the target. For one set of participants, all word pairs were shown a second time (a restudy trial). Another set of participants practiced retrieving the target word for all word pairs (retrieval practice), and a third set of participants received retrieval practice trials for half of the items and restudy trials for the other half. All participants then received a final test in which they were either shown their total score or the score was displayed as the number of correct items that had been previously retrieved versus the number of correct items that had been previously restudied. Participants were then shown a second list of words and were asked about their method of preferred study if they were to be tested. We predicted a standard testing effect, such that participants who practiced retrieving all word pairs would score the highest on the final test, followed by participants who retrieved half the items, and participants who only experienced restudy trials and did not practice retrieving any items would score the lowest. We also predicted that more effective study decisions would be exemplified in the condition in which the testing effect was most apparent. We expected that participants who experienced both restudy and retrieval practice items and were shown their total score as the number of items that they had previously retrieved or restudied would choose to retrieve a greater proportion of word pairs. This was predicted to be followed by participants who experienced both restudy and retrieval practice items but only their total score, and then participants who only experienced retrieval practice items. We hypothesized that participants who only restudied word pairs would choose to restudy items rather than to retrieve them.

Method

Participants

We recruited 138 participants (105 women and 33 men) ranging in age of 18 to 34 years (M = 19.39 years, SD = 2.1 years) from the University of Oklahoma's undergraduate subject pool. Participants received class credit for their participation. Originally, 165 subjects participated in the study, but 20 participants were excluded because they did not complete the experiment, and 7 subjects were excluded for failing to follow directions or complete a portion of the experiment (e.g. they did not participate in the study phase, did not participate in the distractor phase, did not answer any test items, etc.). An attention check question was also employed (e.g. "true or false: I do not understand a word of English) to be used as a basis for eliminating inattentive responders, but all participants who failed this question had previously been eliminated based on the previous criteria.

N = 101 participants, Age M = 36.42, SD = 10.67, 53% female, 47% male were recruited from Amazon Mechanical Turk (mTurk). 103 participants originally participated in the study, but one subject was excluded for failing to follow instructions, and one subject did not complete the experiment. The lower amount of participants eliminated due to inattention or failure to follow instructions compared to the undergraduate sample is expected, because mTurk allows experimenters to withhold payment if the quality of work is not satisfactory. However, subjects recruited from mTurk for this experiment were always compensated the full amount, regardless if they had failed an attention check. Workers were allowed to participate in the study if they

were located in the United States, had completed greater than 1000 surveys on mTurk, and had an approval rating of greater than 95%. MTurk subjects were compensated $2.00 for their time.

The study was administered online for both undergraduate and mTurk participants. Participants were required to complete the experiment in one sitting, and they were only allowed to participate once.

Materials

Sixty word pairs were derived from the University of South Florida Free Association Norms. The average forward (M =0.23, SD = 0.19) and backward (M = 0.00017, SD = 0.01) strengths between cues and targets were small. Targets were also not related to any of the other cues in the list. The 60 word pairs were randomly distributed to create 2 different word lists.

The experiment consisted of 5 phases: list 1 presentation, study phase, distractor phase, list 2 presentation, and study decisions. Participants were randomly assigned to view either List 1 or List 2 first, and were shown the other list second.

Following the study decision phase, a study habit questionnaire from Kornell and Bjork's (2007) study and a question from Einstein, Mullet, and Harrison(2009) was used to assess participants' typical study habits (See Table 1).

Procedure

Participants were instructed that they would be shown a list of cue target word pairs. They were asked to memorize each cuetarget pair and were told that they would

be tested on them later. Participants were also instructed that they might experience restudy and/or retrieval practice trials. Participants were given instructions and an example of each type of trial. In List 1, cue target pairs appeared individually on the screen in a random order for four seconds each. After being presented with the cue-target word pairs, participants were randomly assigned to 1 of 3 study phase conditions. In the retrieve only study phase, participants were shown the cue word followed by the first two letters of the target word (e.g. DARE – BR_____). Participants were given 10 seconds to type the entire target in the space provided. All cuetarget word pairs were shown in this manner for the Retrieve Only group and were presented in a random order. For the restudy only condition, participants were shown all cuetarget pairs intact (E.g. DARE – BRAVE) for the second time. These pairs appeared in a random order, and participants were instructed to type in the whole target word in the space provided with 10 seconds to do so. Because restudy trials asked participants to type in the target word, the only difference between the restudy and retrieval practice trials was that participants had to retrieve the target word from memory in retrieval practice trials. Participants who were assigned to the both condition restudied half of the cue-target word pairs and retrieved the other half. Pairs were randomly assigned to either be presented as a restudy or retrieval practice trial, resulting in 15 restudy trials and 15 retrieval trials. Presentation of study trials for the Both condition were presented in a random order. Following the study phase, participants completed simple arithmetic problems and read an article for a total of 15 minutes. Participants were randomly assigned to read one of the following articles: “Physical Attractiveness Bias in Hiring: What is Beautiful is Good”, “Flying Dinos and Baby Birds Offer New Clues About

How Avians Took Wing", "Liar: It Takes One to Know One", "How Sleep Deprivation Could Add Extra Pounds". Following the distraction phase, participants were given the final test. For the final test for all conditions, only the cue was displayed (E.g. DARE - _____). Participants were given 10 seconds to type in the target in the space provided. While participants were taking their test, their total score was displayed. For participants who were in the Retrieve Only and the Restudy Only conditions, the total score of correct items out of 30 was displayed above each test item. Their total score out of 30 appeared one final time by itself after all pairs had been tested. Participants in the Both condition were randomly assigned to either view their aggregated total score or disaggregated total score. Participants who viewed their Aggregated Score viewed their score in the same manner of the retrieve only and restudy only groups. Participants who were displayed their disaggregated score were shown their correct score as thetotal number of correct items that had been previously retrieved and the total number of correct items that had been previously restudied. These scores were displayed above every test item, and once more after all word pairs had been tested. Following the Final Test, participants were shown a second list of words. Following the presentation of the second list, participants were instructed to make a hypothetical study decision, "If you were to be tested on this second list of words, what percentage of word pairs would you like retrieve or restudy?". Participants were given a description, including an example, of a restudy trial and a retrieval practice trial. They were presented with a slider for Restudy Items and another slider for Retrieve items. Participants could use the sliders to indicate 0% - 100% for each type of item. The total for both Restudy and Retrieve Items was required to equal 100%. Following the study decisions, participants were given a

questionnaire, derived from the Kornell and Bjork's (2007) study, about typical study habits (See Table 1).

Results

Because the average age (and presumably occupation) is different between the mTurk and undergraduate sample, results were analyzed with sample population as a condition. However, Bayes Factor was calculated for the percentage on the final test (t=.243, n1= 138, n2 = 101, Scaled JZS Bayes Factor = 6.79) as well as the percentage of words wished to be retrieved in the study decision phase (t=.416, Scaled JZS Bayes Factor = 6.44) between mTurk and undergraduate participations, suggesting evidence for the null (Rouder, Speckman, Sun, & Morey (2009). In other words, the two samples did not score significantly different on percent correct on the final test and study decisions.

The Study Phase

MTurk participants in the Both Condition scored 75.29% (SD = .23),and SONA participants scored 77.36% (SD = .16), on the 15 retrieval practice trials. mTurk participants in the Retrieval Only condition scored 78.08% (SD = .14), and undergraduate participants scored 74.03% (SD = .16) on 30 retrieval practice trials. Combining samples, the Retrieve only condition correctly answered 75.87% (SD = 26.22) of the 30 retrieval practice trials, and the Both condition correctly answered M = 76.50% (SD = 19.41). We expected performance on these retrieval practice trials to be high because of the two letter stem. We employed this procedure based upon Rowland

and Delosh's (2014) study, which found that the testing effect can be observed even at short intervals if the retrieval practice performance is highly accurate.

Final Test Performance and the Testing Effect Between Participants

Final test performances for each condition are displayed in Figure 1. We predicted that participants who practiced retrieval would score better on the final test. More specifically, we expected the Retrieve Only group to score better than the Both group, as only half on the items were tested, followed by the lowest performance in the Restudy only group. We found a significant effect of condition ($F(5) = 6.185$, $p < .002$) but there was neither a significant effect of sample ($F(5) = .22$, $p = .65$) nor significant interaction between sample and condition ($F(3) = 1.037$, $p = .356$). Post-hoc comparisons indicated that participants in the Retrieve Only group scored significantly higher ($M = 64.47\%$, $SD = 18.22$) than the Restudy only condition ($M = 52.02\%$, $SD = 22.34$, $p = .001$), and the Both Condition scored higher than the restudy condition ($M = 60.49\%$, $SD = 21.65$), $p = 0.034$). Though the Retrieve Only condition scored numerically higher than that of the Both condition, the difference was not significant ($p = 0.422$). The same pattern emerged when collapsed across all participants, such that the Retrieve Only group ($M = 64.47\%$, $SD = 18.22$) scored significantly higher than that of the Both ($M = 60.49\%$, $SD = 21.56$) and the Restudy only group ($M = 52.03\%$, $SD = 22.34$, $F(2) = 7.618$, $p = 0.001$)

Study Decisions

Study decision results can be seen in Figure 2. We hypothesized that participants who were given the opportunity to experience the difference in study strategies would report wanting to retrieve more word pairs on an upcoming exam. Results are reported as the percentage participants would want to retrieve. The percentage that participants wished to restudy can be found by subtracting the listed percentage from 100. We found no significant effect of population ($F(5) = .0242$, $p = .623$) or a significant interaction between condition and population ($F(5) = 1.76$, $p = .174$). A significant main effect for condition was found ($F(5) = 5.23$, $p = .006$), such that the retrieve only group indicated wanting to practice retrieving more word pairs ($M = 71.26\%$, $SD = 28.57$) than the both Condition ($M = 60.40\%$, $SD = 27.03$, $p = .035$) and the restudy condition ($M = 60.57\%$, $SD = 26.23$, $p = .023$). However, there was no statistical difference, and practically no numerical difference between the restudy and Both Condition ($p = 0.99$). These findings are counter to what we had predicted, as participants in the both condition should have experienced the differential benefits and detriments of each study strategy, which should have had a greater impact on study decisions (Tullis et al., 2013; Dunlosky & Hertzog (2000). It is worth noting that study decisions for all groups did not fall below 50%, suggesting that study decisions did not show a strong preference for restudy items in any group.

Experiencing Both Types of Study Items

We also predicted that a testing effect would arise within a list for the Both group. Words that were previously retrieved were recalled at significantly higher rates

(M = 64.23%, SD =23.08) than words that were previously restudied (M = 56.75%, SD = 23.50, t(81) = 3.85, p = 0.001).

In the analysis to determine if the aggregated versus disaggregated presentation of results affected percentage correct on the final test, we did not find a significant main effect of sample type (F(3)=1.15, p = .285), nor a main effect of results presentation (F(3)=.008, p=.927), or a significant interaction (F(3)=1.43, p =.067). These null results, however, were expected because the results are displayed after a final test trial, and thus results should not have had an impact on initial learning and consequently should not have had an impact on final performance.

We also predicted that participants would increase retrieval practice decisions if they were given external support, or the clear identification whether correctly answered items had been previously recalled or retrieved. As seen in Figure 3, we found no significant effect of population (F(3)=2.89, p = .093) or significant effect of results presentation (F(3) = .1.92, p =.169); however, there was a significant interaction between population and results presentation (F(3) = 4.23, p = .043). For mTurk participants, percentage of words requested to be retrieved later on was marginally lower when shown disaggregated scores (M=44.20%, SD = 20.86) than when shown just the total score (M=64.00%, SD = 33.21, t(32) = 2.01, p = .053). However, the undergraduate sample showed the opposite pattern, such that the disaggregated results produced a numerically, but not significantly, higher percentage of words to be desired to be retrieved (M = 65.81%, SD = 24.38) than when results were aggregated into a single total score (M = 61.95%, SD = 21.77, t(46) = -.573, p = .570). While prior

research suggests that increasing external support should promote more effective study decisions, this experiment did not show such support.

Study Habit Survey

Questions for the survey and participant response percentages can be found in Table 1. The first survey question demonstrates, when given the option, most students will test oneself over the word pairs rather than restudy, and most students did not choose to restudy during the study decision phase. Our surveys also seems to support the notion that the majority of students use testing as a means of assessing one's knowledge rather than as a learning event. Furthermore, the last question, which was taken from Einstein et al's (2012) experiment, seems to indicate that most learners are willing to employ more testing during study, with few learners choosing to restudy more. One other result that should be noted, however, is the higher percentage (50 – 74%) of learners who indicate that they re-read sections of the textbook or material. This may indicate the mismatch between laboratory and real world study decisions, in that students may choose to test over simple word pairs, but would rather restudy more complex material.

Experiment 1 Discussion

As predicted, a testing effect emerged such that participants who retrieved more word pairs had better performance. Counter to what was predicted, external support in

the form of results presentation did not affect study decisions. Tullis et al's (2012) study found that external support was able to shift JOLs to be more accurate, however, Experiment 1 failed to show that results presentation have any effect on study decisions. Our findings seem to offer support for Storm, Hickman, and Bjork's (2016) study that showed that experience of the testing effect can influence study strategies. However, Experiment 1 demonstrated that experience of both strategies (testing as a superior strategy, and restudying as an inferior strategy), was not necessary to change study habits. In fact, our findings seem to better support Benjamin and Finley's (2015) test expectancy paradigm, in that participants are able to shift encoding strategies due to experience of the test, and thus study decisions are based on knowledge of the demands of the final test. Participants who were in the Retrieval Only group seemed to be more aware of the memorial effects of retrieval, resulting in significantly higher retrieval percentages. It is possible that instead of the restudy trials acting as a way to compare effectiveness of study strategies, the restudy trials gave subjects the opportunity to assess the strategy as they have in the past, thus basing its perceived effectiveness on its current ease of processing (Dunlosky & Nelson, 1992). This may resulted in the desire to restudy some word pairs. If participants are not given the opportunity to view restudy trials, then they may be less likely to choose to the restudy method.

Experiment 2

To further evaluate whether testing experience can influence study decisions and if the differential effectiveness of study strategies is necessary, we employed a 2 list

procedure in Experiment 2. We reasoned that if participants are able to experience a stronger benefit of testing, this may lead participants to choose a greater proportion of tested items; one that is greater than that of the Retrieval Only group. In Experiment 1, participant scores were out of 30, and thus subjects could correctly recall 15 items that had been previously restudied and 15 items that were previously retrieved. With a two list paradigm, the difference between the two study items on the final performance could be more apparent.

As with Experiment 1, we predicted a testing effect, such that participants who practiced retrieving items would score higher on the final test than when words were just restudied. We predicted that participants who experience both the detriment of restudying and the clear advantage of the testing effect would demonstrate superior future study decisions. More specifically we predicted that those participants who experience both a study episode with all retrieval practice trials and another study episode with all restudy trials would become more sensitive the benefits of testing, thus influencing them to choose to test oneself more than those participants who experienced only retrieval practice trials, and especially more than those who only experienced restudy trials.

Method

Participants

N = 67 participants (Age M = 19.32, SD = 1.45, 53% female, 47% male) were recruited for the University of Oklahoma's undergraduate subject pool. Subjects

received class credit for their participation. 91 subjects originally participated in the experiment, but 22 subjects were removed from data analysis because they not finish the experiment, and 2 participants were removed for not following instructions. N = 114 participants, Age M =39.07, SD = 12.10, 64% female, 36% male, were recruited mTurk. 117 participants had started the experiment, but 1 was removed for data analysis because she did not finish the experiment, and 2 were removed because they failed to complete a portion of the experiment. Qualifications for mTurk workers to participate as well as compensation for participating were the same as in Experiment 1. Subjects who had participated in Experiment 1 were not allowed to participate in Experiment 2.

Materials

The materials for Experiment 2 were exactly the same as materials used for Experiment 1.

Procedure

The procedure for Experiment 2 was very similar for Experiment 1, with the exception of the elimination of the Both Retrieve and Restudy condition for the first list and the addition of the second test. Participants were randomly assigned to see List 1 first and List 2 second, or vice versa. After the presentation of the first list, participants were randomly assigned to either the Restudy or Retrieve group. These conditions were the same as the Restudy Only and Retrieve Only conditions, respectively, in Experiment 1. Participants were then shown the same distractor phase, followed by the Final Test. Subjects' total score for the first list was displayed as it was in Experiment 1.

Participants were then presented with the second list, then randomly assigned to either the Restudy group or Retrieve group. This resulted in 4 conditions: Restudy-Restudy, Retrieve-Retrieve, Retrieve-Restudy, and Restudy-Retrieve. Participants then had a distractor phase in which they completed different, but similar simple arithmetic problems and an article that they had not previously read. The Final Test for List 2 was then presented. Following the final test, total scores for both lists were displayed. For participants in the Restudy-Restudy and the Retrieve-Retrieve group, they were displayed their correct scores from List 1 and List 2. Participants in the Retrieve-Restudy or Restudy-Retrieve group either viewed their final scores as the 2 other conditions viewed it, or they viewed their scores as labeled as Total Correct Items that had been previously retrieved ___, Total Items Correct for items that had been previously restudied. Participants were then displayed the study decision sliders from Experiment 1, except it asked them to make a judgment if they were to be shown and tested over a third list of words.

Results

Participants were collapsed across samples because sample did not result in significant differences in Experiment 1. In addition, a Bayes Factor was calculated for the total score for the first list ($t=1.005$, $n1= 67$, $n2 = 114$, Scaled JZS Bayes Factor = 3.76) as well as the total score for the second list ($t=.613$, Scaled JZS Bayes Factor = 5.93) between mTurk and undergraduate participations, in which both suggested that there is no difference between the samples.

Study Phase

Performance on retrieval practice trials was similar to performance in Experiment 1 for both the first list (M = 73.33%, SD = 18.65) and the second list (M = 73.89%, SD = 20.42), which were not significantly different from one another (t(59) = -.234, p = .816).

Testing Effect Between Participants

These results can be viewed in Figure 4.We hypothesized that we would see a testing effect between participants, such that participants who experienced testing effect for both lists would outperform participants in the Both condition, who we expected to outperform those in the Restudy condition. We calculated a Grand Total for all participants, which was comprised of the total score from both lists out of 60. There were no significant differences between the three conditions (Retrieve Only M = 56.56%, SD = 20.12; Restudy Only M = 52.63%, SD = 22.70; Both M = 58.85%, SD = 23.33) for Grand Total (F(2) = 1.159, p = .316). A non-significant difference between the Both condition and the other two conditions is not surprising, given that the Both condition employs both strategies, and therefore is inherently similar to the other conditions. However, it is surprising that a testing advantage was not found.

We also averaged performance across both tests for the Retrieve Only and Restudy Only conditions, and compared them to both of the Both condition's Retrieve and Restudy lists. The only significant difference was that between the Both's restudy (M = 53.00%, SD = 26.36) and retrieve conditions (M = 64.69%, SD = 24.01, t(70 = -

2.659, p = 0.010). Though the Retrieve Only condition was numerically higher than the Restudy Only condition, all other comparisons were non-significant (Retrieve Only M= 58.96%, SD = 19.96; Restudy Only M = 55.70%, SD = 24.50).

Because these analyses required the combining of lists, we hypothesized that perhaps the decrease in the testing effect was due to fatigue, as participants were required to complete 2 of everything, including the distractor. If this were the case, we should expect to see a significant difference between participants' performance on the first list and the second list; however, this was not the case. Results can be found on Figure 5. Scores on the first test that participants completed were not significantly different than scores on the second list within any condition (Retrieve Only List 1 M = 56.28%, SD = 22.85, List 2 M = 56.83%, SD – 21.61, t(59) = -.227, p = .821; Restudy Only List 1 M = 51.60%, SD = 25.62, List 2 M = 53.67%, SD = 24.61, t(49) = -.714, p = .478; Both List 1 M = 58.5%, SD = 28.33, List 2 M = 59.20%, SD = 23.18, t(70) = -.265, p = .792)

It is interesting to note that performance on these tests was lower than the average performance in Experiment 1. Experiment 2 was conducted at the end of the semester, and the beginning of the summer. It is possible that the lower performance rates could be due to differences in motivation as summer and the end of the school year draws near.

Study Decisions

Study decision results are shown in Figure 6. We predicted that if participants were able to experience a greater benefit of testing, they would choose to retrieve more

word pairs. Results again are reported as the percentage of word pairs a participant indicated he or she would like to retrieve if there was an upcoming test. However, we did not find any effect of condition; retrieval study decisions were similar across all conditions (Retrieve Only M = 61.95, SD = 24.25, Restudy Only M = 61.64, SD = 27.36, Both M = 62.47%, SD = 28.97, F(2) = 0.015, p = 0.985). Results are in Figure 7. Experience alone of the test's demands was not enough the change encoding strategies. Numerical support, however, would show support for Dunlowsky and Herzog's (2000) Effectiveness component, such that differences between study strategies are necessary to impact study decisions. As in Experiment 1, none of the conditions showed a strong restudy preference.

We also did not find an effect for results presentation on study decisions for those participants in the Both condition. Study decisions were not significantly different when results were presented as a Total Score (M = 63.66, SD =29.7) compared to when they were presented with the indication which list had been previously retrieved or restudied (M = 61.12, SD = 28.51, t(69) = .366, p = .766). As in Experiment 1, the external support provided did not seem to guide learners to make more effective study decisions.

Comparing Experiment 1 and 2

Because of the lack of significant results in Experiment 2, we decided to compare results across experiments. Caution should be heeded, however, because conditions are not truly equated across experiments. Furthermore, the average performance between two lists was used for Experiment 2 final test performance. For

performance on the final test, we found a significant main effect of condition, such that participants in the Retrieve Only group (M = 61.26%, SD = .20) scored significantly higher than both the Restudy (M = 52.28%, SD = 22.39) and Both group (M = 59.73%, SD = 22.34, F= 5.56, p =.004); however, the Restudy condition was not significantly different than the Both group (p=.807). There was no significant main effect of experiment (F=1.98, p = 0.159) nor interaction between condition and experiment (F=1.439, p =.238). Results are shown in Figure 8. For study decisions, we did not find a significant main effect of condition (F=1.83, p = .161, main effect of experiment (F=.594, p =.411), or interaction (F=1.95, p = .143). Results are displayed in Figure 9. After comparing Experiment 1 and 2, we found similar patterns as Experiment 1 for final test performance, and although Experiment 2 performance seemed to be lower than that of Experiment 1, Experiment did not have a significant effect on either performance or study decisions. Study decisions, however, did not significantly differ between conditions, unlike Experiment 1 which showed that the Retrieve group having a higher proportion of retrieval practice trials in their study decision.

Experiment 2 Discussion

Experiment 2 was designed to increase the Testing Effect, such that its benefits would be more apparent; however, the two list design seemed to eliminate the testing advantage between the restudy and retrieval only groups; however, the testing effect was still observed within the Both condition. Although there is a research to show that prior testing does have a positive effect for future tests, this explanation seems unlikely

given that performance on both tests was not significantly different. Participants who were able to experience the testing effect between the two lists did indicate a numerically higher, albeit small, preference for testing compared to the other conditions.

When we compared across Experiments, the same pattern for the testing effect emerged, such that the retrieval group scored better than the Both group and restudy group. However, unlike the pattern in Experiment 1, study decisions of the Retrieve Only group was not significantly higher. This again showed evidence against the concept that differential experience of study strategies can influence study decisions.

General Discussion

Experiment 1 demonstrated that experience of testing can guide learners to pick more effective study strategies, whereas Experiment 2 did not show any differences across groups for study decisions. Experiment 1 demonstrated that experience of retrieval practice itself can influence study decisions. Storm, Hickman, and Bjork (2016) suggested that experiencing different study strategies, one more optimal than the other, can contribute significantly to changing encoding strategies, but Experiment 1 demonstrated that experience of restudy trials is not necessary to improve study decisions. Although we did not find the expected testing effect between conditions in Experiment 2, the testing effect did occur when participants experienced both restudy and a retrieval practice phase, and this experience seemed to have a least a numerical effect on study decisions.

These experiments also explored the role of external diagnostic support. We employed a procedure in which participants were shown how many correctly recalled words had been either retrieved or restudied previously, and we predicted that an increased amount of diagnostic support would result in better study decisions. We did not, however, find such support. Both performance and study decisions did not differ on the amount of external diagnostic support. It is possible that participants failed to notice, or give weight, to the scores that were displayed. If participants did not care of the outcome of the final test or felt that their performance was irrelevant, then it is likely that participants did not process the information. Although learners were given the support to make judgements over the effectiveness of each study method, they likely did not use the information, or were unable to make a connection to future study decisions.

Although Tullis, Finley, and Benjamin (2013) found that diagnostic support during study can improve JOLS, we did not find evidence to suggest that providing diagnostic feedback can change future study decisions. Carpenter, Lund, Coffman, Armstrong, Lamm, and Reason (2016) also found similar results as Tullis and colleagues (2013) in an applied classroom study. They found that when students engaged in activities requiring retrieval, they performed better on exams and also had more accurate metacognitive judgements when predicting future performance. It is possible that JOLs are more sensitive to diagnostic features and capable of changing more so than study habits. By the time students are in college, they have likely developed a preferred study method and are therefore resistant to change. A future study should use Tullis et al.'s (2013) procedure in addition to a study decision to investigate whether JOLs and study decisions are independent.

DeWinstanley and Bjork's (2004) report on the generation effect suggests that when learners recognize the benefit that generation provides, learners engage in superior encoding strategies when presented with a similar task. If testing experience and the knowledge of its memorial advantages is enough to change encoding strategies, we should have observed an increase of performance on the second list. The present experiments did not explore this because a second list was not employed when testing condition was manipulated within a list, however, we did not find support that the differential experience between test and restudy significantly improves study decisions. Finley and Benjamin (2012) found that learners can change and improve encoding strategies if they are aware of the type of test and its test demands; however, if encoding strategies were improved because learners were aware of the type of test after the List 1 test, it was not apparent in either study decisions or test performance for Experiment 2. Experiment 1, however, may provide some evidence for the Test-Expectancy paradigm, such that learners who received retrieval practice trials essentially had extra exposure to the type of test demands. This extra exposure to the type of test may have resulted in improved encoding strategies, and could explain why participants in the retrieval only condition requested to retrieve more items than the other conditions.

The core assumptions of the Dunlowsky and Hertzog's (2000) Updating Knowledge framework included effectiveness, monitoring, updating, and utilization; however, Experiment 1 demonstrated that experience of differential benefits of the study strategies was not necessary to increase study decisions towards increased testing trials because participants in the Retrieve Only condition chose to test more so than the other groups. Further, the results presentation, which was designed to aid monitoring,

did not have an effect in either experiment. It is possible that participants who were not shown how many words were correct based on study strategy were able to accurately monitor without the external support, however, this seems unlikely given that Tullis, Finley, and Benjamin (2012) found that participants were unable to keep track of their scores on their own. In order for participants to monitor effectiveness on their own, they would be required to remember if the target had been previously retrieved or restudied as well as keep a tally as how many of each they had correctly answered. Again, this seems unlikely, especially given that participants were not even able to remember all of the targets.

While the present experiments shed light on the possibility that retrieval practice may be sufficient in improving study strategies, there are several ways in which our experiments can be improved. For one, our procedure only asked students for a hypothetical immediate study decision. We could ask participants for a prospective study decision such as, "If you were to be tested on these words in 2 weeks, how would you like to study? Although, a decision like this would have to interact with another desirable difficulty – Spaced study. Future research should use actual future study decisions at longer delays to more accurately evaluate study choices, rather than just a hypothetical decision. Our procedure also evaluated the testing effect at a short retention interval. Although we were able to observe a testing effect, it may be possible that the same procedure with a longer delay would heed different results because the testing advantage may be more apparent to learners. Furthermore, the ability to generalize our findings to a real classroom setting may be limited. Not only are word pairs much less complex than texts, but we also used a two letter word stem in the

retrieval practice trials. These two letter word stems are unlikely to occur if one were to employ self-testing; however, we will argue that the two letter word stem is akin to peeking at the first part of an answer.

There are number of reasons why students may not choose to test themselves. One possibility is that students have difficulty assessing their own responses, making it difficult to test oneself. Rawson and Dunlowsky (2007) conducted a study in which students took an exam and were tasked to grade their work and award themselves full credit, partial credit, or no credit. When presented with their own response and could directly compare it to the correct response, students' assessment of their own work was incorrect 43% of the time. If students are unable to accurately evaluate their own work, this may be a reason why students do not test more often. Rawson and Dunlowsky's (2007) study perhaps points to the importance and growing need of computer based programs that can offer students immediate, constructive, and accurate feedback.

Perhaps one of the main reasons why students choose to restudy so often is because it can be a more effective study strategy in the short term (Roediger & Karpicke, 2006). If students experience an advantage of restudying when taking an immediate test, then this likely perpetuates the notion that restudying is an effective strategy for long term retention. As demonstrated in our survey of study habits, many students (48-73%) do not return to course material after a course has ended. If students feel they do not or will not have the need for the information at a later, they never have the opportunity to see why restudying results in poor long term retention as compared to testing. Another reason why participants may choose to not test themselves is that it takes longer and is more difficult (hence the term, "desirable difficulty"). Presentation

requires only one-step to display information, whereas testing requires 2 steps: partial presentation and retrieval. An extra step is also added if feedback is given (Izawa, 1992).

As mentioned previously, students may also use testing as a means of assessments instead of as a potent learning event, and thus do not incorporate it into their study habits. One's ability to accurately monitor one's memory has been discussed in a variety of studies in the context of how it impacts study decisions (Kornell & Metcalfe, 2006, Son & Metcalfe, 2000). These studies propose that as metacognitive accuracy increases, more effective study decisions are able to be utilized. However, Kornell and Son (2009) demonstrated a metacognitive dissociation, such that participants will choose to test even though they believed that restudying was more effective than testing. Their results suggest that study decisions may be independent of metacognitive assessments. Our procedure did not require metacognitive judgements, but rather relied on the experience of testing to influence later study decisions. Perhaps it is not important as to the reason why students engage in certain study decisions, but rather greater importance should be placed on determining the factors that can guide them toward using more effective study methods.

Experiencing the testing effect, or as this research demonstrated in Experiment 1, the experience of retrieval practice trials, may be one factor that can guide students toward increased use of testing. Instructors should therefore encourage self-testing by emphasizing its long term memorial benefits. Instructors should also include testing and retrieval practice exercises within lessons not only so that students can benefit from the

testing, but also because experience of testing may enhance study decisions, as these current studies have shown.

Lastly, laboratory research of the testing effect should also be evaluated in a real classroom setting. Effectiveness of the testing effect has varied across applied settings. For example, Roediger et al. (2011) found that a 9% advantage on a final test had students received a quiz immediately following the lesson. However, Karpicke, Blunt, Smith, and Karpicke (2014) found that retrieval exercises, such as the use of flashcards, did not significantly improve later performance when compared to other activities that required in depth interaction with the information. On the other hand, Carpenter et al (2016) found that retrieval practice enhanced quiz performance, and Einstein et al. (2012) concluded that experience of the testing effect contributed to improved study strategies.

The present research has once again demonstrated the value of testing and the robust testing effect even at shorter delays, but it also touches on the difficulties in changing study decisions and strategies. Current strategies, which students often have used for years and perhaps have grown accustomed to, may be more resistant to change. Further research should explore other ways in which instructors can improve study strategies and thus performance.

Table 1: Study Habits Part 1

Questions		Retrieve Only		Restudy only		Both	
		Undergrads	mturk	undergrads	mturk	undergrads	mturk
If shown another list of 30 cue-target word pairs, how would you study them for an upcoming test?							
	Restudy the intact cue-target word pairs	27%	36%	17%	37%	21%	24%
	Test myself by covering up the target and seeing if I can come up with the answer	67%	57%	60%	56%	75%	56%
	Another study strategy	6%	7%	10%	4%	4%	3%
Would you say you study the way you do because a teacher taught you to study that way?							
	Yes	40%	31%	55%	26%	37%	47%
	No	60%	69%	45%	74%	63%	53%
How do you decide to study next?							
	Whatever is due soonest/overdue	54%	45%	54%	30%	40%	32%
	Whatever I haven't studied in a long time	0%	0%	4%	0%	10%	15%
	Whatever I find interesting	2%	17%	0%	26%	2%	29%
	Whatever I feel like I'm doing the worst in	28%	24%	15%	33%	29%	15%
	I plan my study schedule ahead of time and study whatever I've scheduled	17%	14%	15%	11%	19%	9%
Do you usually return to course material to review it after a course as ended							
	Yes	33%	43%	27%	52%	23%	47%
	No	67%	57%	73%	48%	77%	53%

Table 2. Study Habits Part 2

		Retrieve Only		Restudy Only		Both	
		Undergrads	mTurk	undergrads	mTurk	Undergrads	mTurk
Of all things being equal, what do you study more for?							
	Essay/short answer exams	23%	21%	21%	29%	13%	15%
	Multiple choice exams	27%	24%	29%	22%	35%	35%
	About the same	49%	55%	4%	48%	35%	50%
When you study, do you typically read a textbook more than once?							
	Yes I reread whole chapters	15%	19%	15%	33%	13%	29%
	Yes I reread sections that underlined/highlighted/marked	58%	74%	50%	60%	60%	47%
	Not usually	27%	5%	23%	7%	27%	24%
If you quiz yourself while you study (either using a quiz, flashcards, etc.) why do you do so?							
	I learn more than way than through rereading	23%	29%	27%	15%	19%	29%
	To figure out how well I have learned the information I'm studying	52%	40%	40%	67%	63%	47%
	I find quizzing more enjoyable than rereading	20%	26%	13%	4%	8%	24%
	I usually do not quiz myself	4%	5%	8%	15%	10%	8%
How often will you incorporate testing in reading and studying in the future?							
	Substantially more often	19%	14%	27%	19%	19%	9%
	Somewhat more often	48%	40%	40%	30%	35%	26%
	No more or no less than I currently do	33%	40%	17%	48%	46%	59%
	Somewhat less often	0%	5%	4%	0%	0%	6%
	Substantially less often	0%	0%	0%	3%	0%	0%

Figure 1. Exp 1 Final Test Performance

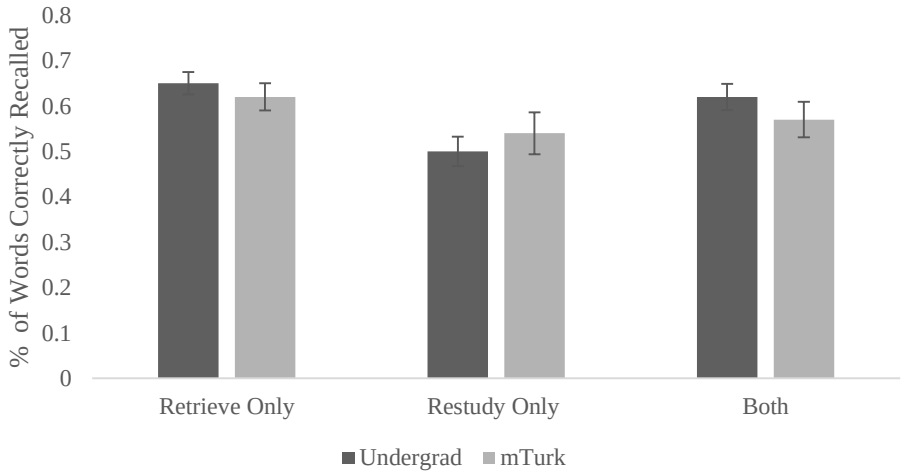


Figure 2. Exp 1 Study Decisions

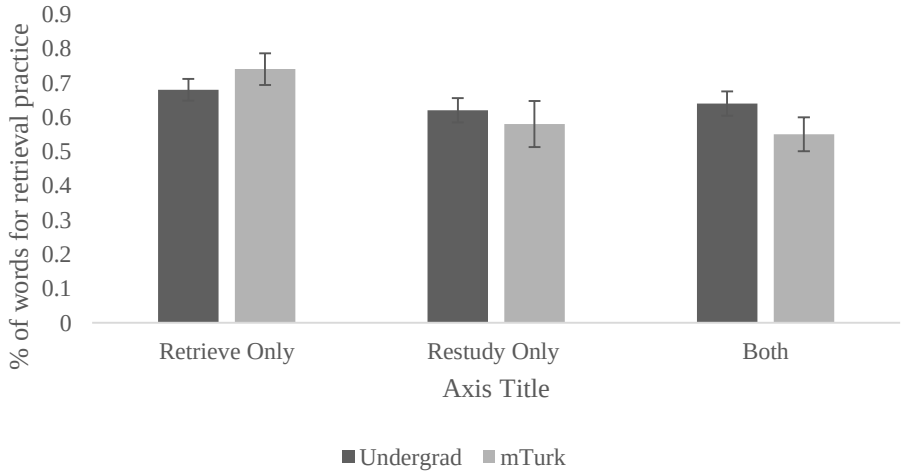


Figure 3. Exp 1 Results Presentation

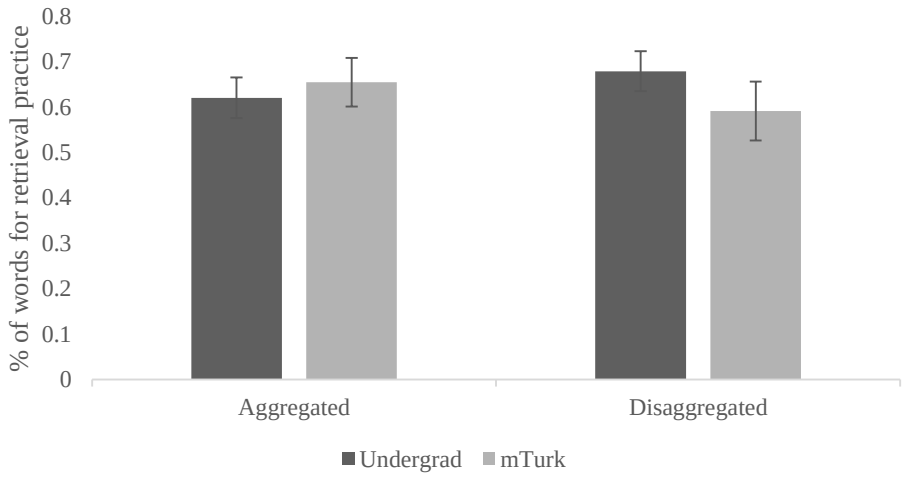


Figure 4. Exp 2 Final Test Performance

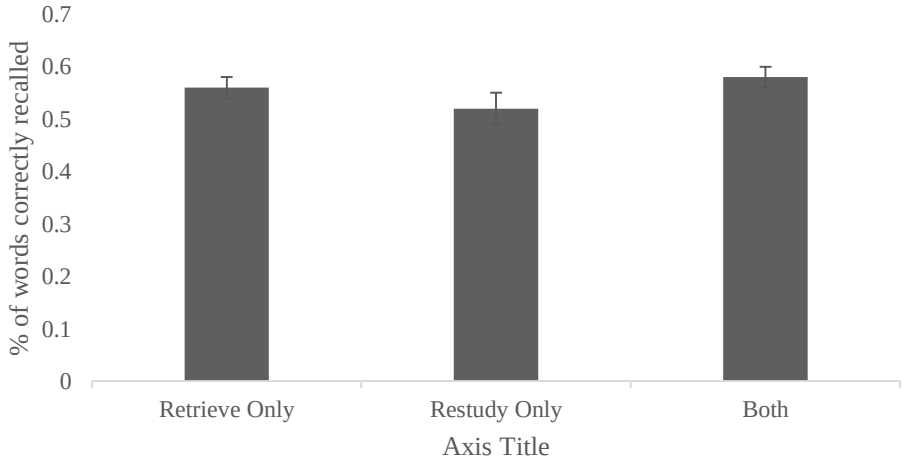


Figure 5. Exp 2 List Differences

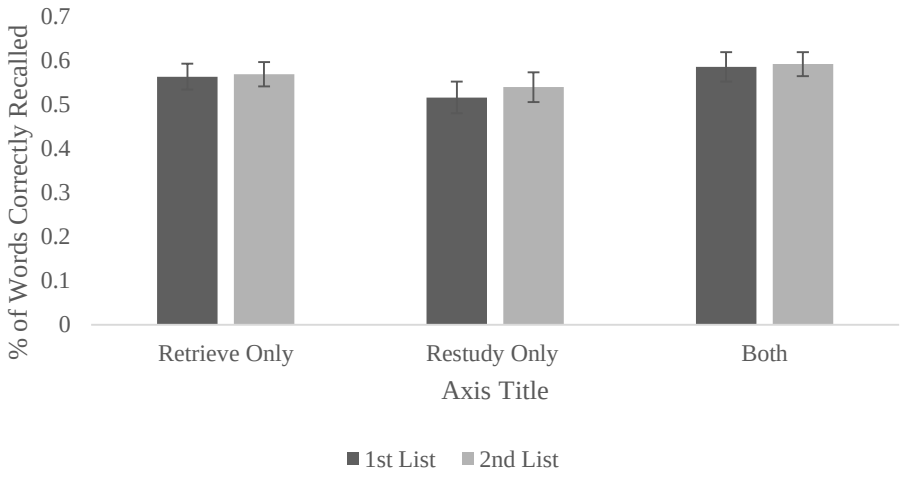


Figure 6. Exp 2 Study Decisions

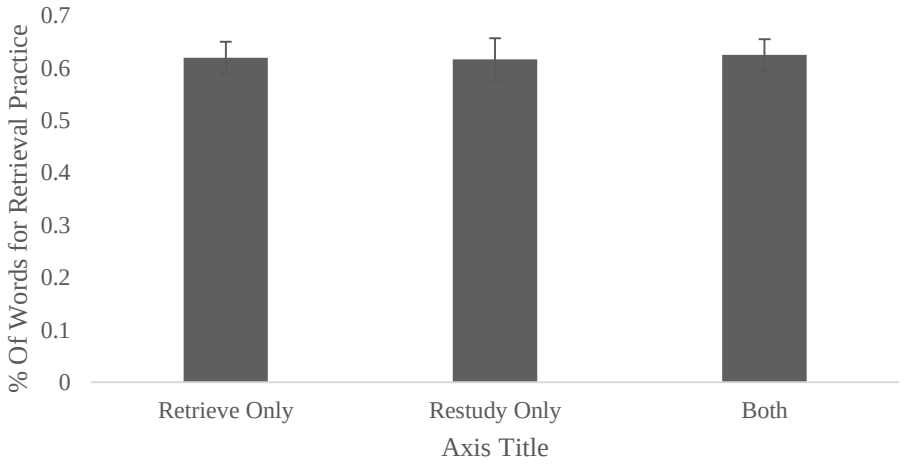


Figure 7. Experiment 2 Results Presentation

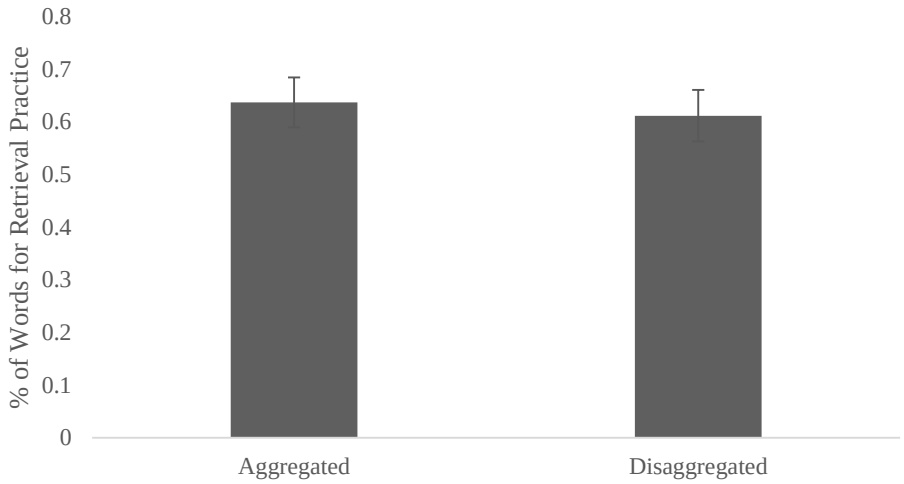


Figure 8. Performance Across Experiments

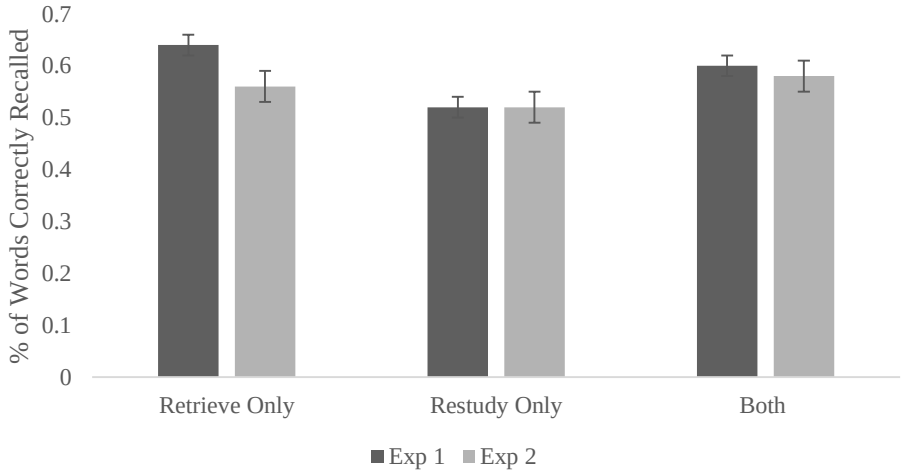
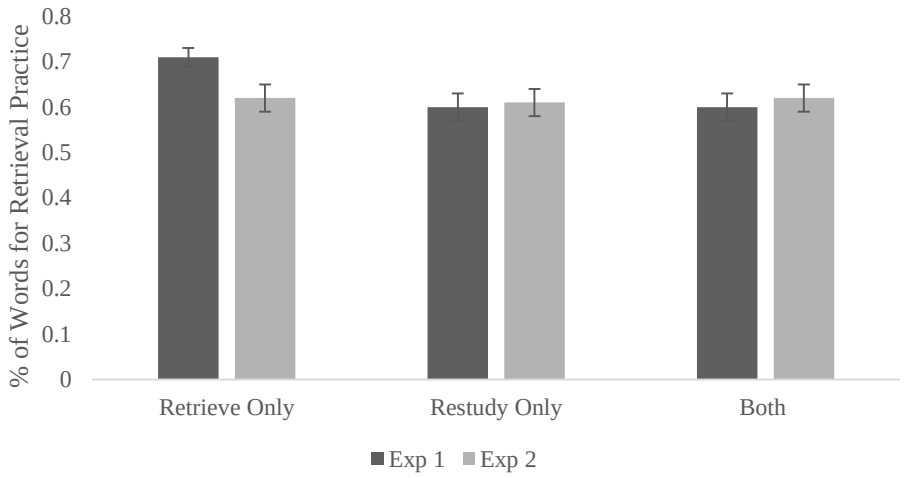


Figure 9. Study Decisions Across Experiments



References

[Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008).

Examining the testing effect with open- and closed-book tests. Applied Cognitive Psychology,22, 861_876.

Ariel, R. (2013) learning to learn: The effects of task experience on strategy shifts in the allocation of study time. Journal of Experimental Psychology, 39(6), 1697-1711

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation:

When agendas override item-based monitoring. *Journal of Experimental Psychology:*

General, 138, 432– 447.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings.

In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name

learning. Applied Cognitive Psychology, 19, 619_636.

Carpenter, S.K., Lund, T.J., Coffman, C.R., Armstrong, P.L., Hamm, M.H., & Reason, R.D.

(2016). A Classroom study on the relationship between student achievement and retrieval enhanced learning. Educational Psyhcology Review, 28(3), 353-375.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval induced facilitation:

Initially nontested material can benefit from prior testing of related material. Journal of Experimental Psychology: General, 135, 553–571

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing

for cued recall. Applied Cognitive Psychology, 14, 215_235.

deWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect:

Implications for making a better reader. Memory & Cognition, 32, 945–955.

Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role

of strategy production? *Psychology and Aging, 13,* 597-607.

Dunlowsky,J. & Hertzog, C. (2000) Updating knowledge about encoding strategies: a compenential anaylsis of learning about strategy effectiveness from task experience.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition,* 20, 374-380.

Einstien, G.O, Mullet, H.G, & Harrison, T.L, (2012) The testing effect: Illustrating a fundamental concepnt and changing study strategies.

Finley, J.R., & Benjamin, A.S. (2012) Adaptive and qualitative changes in encoding strategy with experience: evidence from the test-expectancy paradigm. Journal of Experimental Psychology, 38(3), 632 -652.

Izawa, C. (1992). Test trials contributions to optimization of learning processes: Study/test trials interactions. In A. F. Healy & S. M. Kosslyn (Eds.), Essays in honor of William K. Estes: Vol. 1. From learning theory to connectionist theory (pp. 1_33). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning

of Swahili–English translation equivalents. *Memory, 2,* 325–335.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student

learning: Do students practice retrieval when they study on their own? Memory, 17, 471–479.

Karpicke, J.D., Blunt, J.R., Smith, M.A., & Karpicke, S.S. (2014). Retrieval-based learning: the

need for guided retrieval in elementary-school children. Journal of Applied Research in Memory & Cognition, 3, 198-206.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge

during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 31, 187-194.

Kornell, N. & Bjork, R.A. (2007). The promise and perils of self-regulated study. Psychonomic

Bulletin and Review, 14 (2), 219 – 224.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the Enemy of

induction? Psychological Science, 19, 585–592

Kornell, N., Castel, A.D., Eich, T.S., Bjork, R.A. (2010) Spacing as the friend of both memory

and induction in young and older adults, Psychology and Aging, 25(2), 498 – 503)

Kornell, N. & Son, L.K. (2009) Learners' choices and beliefs about self testing, Memory, 17:5,

493-501

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011).Retrieval during learning facilitates subsequent memory encoding. Journal of Experimental Psychology: Learning, Memory, andCognition, 37, 287–297. doi:10.1037/a0021801

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P.

R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). New York, NY: Academic Press.

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key

concepts in textbook materials. *European Journal of Cognitive Psychology*, *19*(4-5), 559-579.

Roediger, H. L. III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-

enhanced learning in the classroom: long-term improvements from quizzing. Journal of Experimental Psychology: Applied, 17, 382–395.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testingmemory: Basic research and implications for educational practice.Perspectives on Psychological Science, 1, 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science,17,* 249–255.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1-36). Oxford: Elsevier

Rouder, J.N., Speckman, P.L., Sun, D.& Morey, R.D. (2009) Bayesian t tests for accepating and

rejecting the null hypothesis, 16 (2), 225 – 237.

Rowland & Edward L. DeLosh (2015) Mnemonic benefits of retrieval practice at short retention

intervals, Memory, 23:3, 403-419

Storm, B.C., Hickman, M.L., Bjork, E.L. (2016), Improving encoding strategies as a function of

test and knowledge experience. Memory Cognition, 44, 660 – 670.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation.

Journalof Experimental Psychology: Learning, Memory and Cognition, 26, 204_221

Tullis, J.G., Finley, J.R., & Benjamin, A.S. (2013). Metacognition of the testing effect: guiding

learners to predict the benefits of retrieval. Memory Cognition, 41, 429-522