8-2019

Mechanisms Underlying the Testing Effect

Mary Chapman
musicmg2@gmail.com

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

MECHANISMS UNDERLYING THE TESTING EFFECT

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Arts

Mary Chapman

College of Educational and Behavioral Sciences
School of Psychological Sciences
Educational Psychology

August 2019

This Thesis by: Mary Chapman

Entitled: *Mechanisms Underlying the Testing Effect*

has been approved as meeting the requirement for the Degree of Master of Arts in the College of Education and Behavioral Sciences in the School of Psychological Sciences, Educational Psychology

Accepted by the Thesis Committee:

_____

James Kole Ph.D, Chair

_____

Sue Hyeon Paek Ph.D, Committee Member

Accepted by the Graduate School

_____

Linda L. Black, Ed.D.
Associate Provost and Dean
Graduate School and International Admissions
Research and Sponsored Projects

ABSTRACT

Chapman, Mary. *Mechanisms Underlying the Testing Effect*. Unpublished Master of Arts thesis or creative project, University of Northern Colorado, 2019.

The current experimental study attempted to disentangle retrieval of target information from the context surrounding the information in the *testing effect*, or the finding that taking a practice test leads to better retention on a final test, for the purpose of discovering the mechanisms underlying the phenomenon. Twenty-three participants studied 30 cue-target pairs over three blocks and then either re-studied the pairs or practiced retrieving the target words for another three blocks. All participants completed a final test in which they recalled the target words for all 30 pairs once and performed a lexical decision task in which they had to indicate whether a string of letters was a word or a non-word. The words in the lexical decision task consisted of new words and old words, and response time and accuracy were recorded. None of the results were statistically significant, but the data tended to trend in specific directions. The practice test group had a higher proportion correct on the final test than the re-study group, trending toward a testing effect finding. For the lexical decision task, participants in the practice test group responded to old words slower but more accurately than those in the re-study group. The results support hypotheses that claim participants encode the context around the target words, taking more time to retrieve the context before they retrieve the

target words, but the context also aids in successfully retrieving the target words. In general terms, these results impact how students should learn material in educational settings. It is widely recommended that students test themselves to best learn information from class, but adding a context around the information to be learned, such as creating a story around the information, can be even more beneficial.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

CHAPTER I

INTRODUCTION

There is robust evidence that studying by taking tests improves retention over simply re-studying the information (for a meta-analysis see Chan, Meissner, & Davis, 2018). This phenomenon is known as the *testing effect*. Major hypotheses within the testing effect literature to explain this enhanced retention propose that the act of retrieving information modifies that target information in such a way that it creates additional pathways leading to the target information, making it easier to retrieve later. One hypothesis, the *elaborative retrieval hypothesis*, emphasizes the creation of elaborative semantic memory traces with retrieval practice (Carpenter & DeLosh, 2006). The other, the *episodic context hypothesis*, focuses on the temporal context around the target information, or how the internal and external context can change from one trial to the next and affect retrieval (Karpicke, Lehman, & Aue, 2014).

Many experiments conducted to support one hypothesis or the other manipulate the test format (Stenlund, Sundström, & Jonsson, 2016), the time between practice and final testing (Pansky, 2012), or other similar manipulations to uncover the mechanisms underlying retrieval and why practice testing produces better retention. However, the manipulations simply give further evidence that the testing effect exists and not why it exists. They do not answer why practice testing is better for long-term retention than re-

studying. The present study tested an alternate explanation for why retrieval practice is better for retention by proposing the *baseline activation* hypothesis. This hypothesis states that, in addition to creating more elaborative memory traces and increasing the number of contextual cues, retrieval of target information results in a strengthening of the target information itself. The present experiment attempted to separate the role of strengthening of target information from elaborative memory traces and contextual cues in the testing effect. The hypotheses were as follows:

H1 Participants in the retrieval practice group will have more proportion correct on the final test than in the re-study group, confirming the testing effect.

H2 Participants who practiced retrieving target words will respond to old words in the lexical decision task faster and more accurately than those who simply re-studied the word pairs in support of baseline activation.

CHAPTER II

REVIEW OF LITERATURE

Elaborative Retrieval Hypothesis

Various hypotheses have arisen to explain why retrieving information results in better retention than simply re-studying the information. One such hypothesis is the *elaborative retrieval hypothesis*, which states that during the process of retrieval, searching for the target information in long-term memory activates other information related to the target information. Successful retrieval of the target information creates a new memory that includes the target information plus the other information that was activated in long-term memory. This new memory is an elaborated memory trace that makes it easier to retrieve the target information later because of the added retrieval cues connected to the target information (Rawson, Vaughn, & Carpenter, 2015). For example, the cue-target word pair "floor" and "chair" could evoke a word that is related to both words such as "desk." When people are presented with the cue word "floor," they could retrieve the mediator word "desk" to aid in retrieving the target word "chair." Without the mediator, the cue word is the only pathway connected to the target word, thus making the target word less likely to be remembered. However, the mediator adds a second pathway connecting to the target word, making the target word more accessible.

The elaborative retrieval hypothesis has found some support and remains one of the most widely used hypotheses for explaining why there is a testing effect. One studythat compared the elaborative retrieval hypothesis to the transfer appropriate processing hypothesis found that the latter explanation was not completely adequate to explain the testing effect (Carpenter & DeLosh, 2006). The transfer appropriate processing hypothesis states that there is a testing advantage because the same processes used to retrieve and answer practice test questions are also used for the final test (Carpenter & DeLosh, 2006). If that were true, answering multiple choice practice questions should result in a testing advantage if the final test also has multiple choice questions. It would be the same result for short answer questions and any other conceivable test format. However, Carpenter and DeLosh (2006) found that when participants answered free recall practice questions, they performed better on the final test whether the final test format was free recall, cued recall, or recognition. As an alternative to the transfer appropriate processing hypothesis, the authors conducted more experiments either averaging the number of cues requested or directly manipulating the number of cues participants saw. The cues came in the form of presenting the first letters of the target word participants needed to retrieve. They found that those who saw fewer cues had better accuracy on the final test, hence supporting the elaborative retrieval hypothesis in that those who saw the least cues had to work harder to create the elaborated memory trace and allowed them to more easily retrieve the information later (Carpenter & DeLosh, 2006).

Another study in support of the elaborative retrieval hypothesis manipulated the type of cues presented on the final test and the amount of time between study and practice testing or re-study (Rawson et al., 2015). Time was enforced through the number of word pairs presented, therefore, the group with more time between study and practice had longer lists with more items while the group with a shorter time lag had shorter lists. On the final test, all participants saw either the cue they were familiar with or a different mediator cue that could also be related to the target word with the instruction to think of the target word they practiced that best goes with the mediator cue. In Experiment 2, the authors found that participants who had a longer time lag, performed the practice tests, and saw the mediator cues on the final test had the best accuracy. Rawson et al. (2015) explained the results in terms of the elaborative retrieval hypothesis, claiming that those who saw the mediator cues had to search their long-term memory more to retrieve the correct target word which created an even more elaborate memory trace than those who saw the familiar cue. This effect was compounded by practice testing and a longer lag, meaning that the memory trace was not as strong with a longer lag, thus participants had to search their memories more when they had to retrieve what they studied (Rawson et al., 2015).

Carpenter (2009) completed a study in which cue strength was manipulated. Participants either saw cues that were highly related to the target word or cues that were not highly related to the target word. In support of the elaborative retrieval hypothesis, the target words that were paired with a strong cue initially performed better on the practice trials, but target words paired with weak cues had better accuracy on the final

test in Experiment 1. The weak cues forced participants to perform a more thorough search through long-term memory to retrieve that target word, and any associations that were also remembered became part of the elaborated memory trace surrounding the target word, creating more pathways to retrieve the target word.

Although the elaborative retrieval hypothesis has received ample support, there has been evidence that contradicts it. Lehman and Karpicke (2016) broke down the elaborative retrieval hypothesis into two assumptions: the act of retrieving information activates mediators related to that information in long-term memory and that these extra mediators aid in later retrieval of that information. Their experiments proceeded to test these assumptions. Experiments 1a, 1b, and 2 tested the first assumption by adding a lexical decision task to the typical testing effect paradigm. Lexical decision tasks ask participants to indicate whether a string of letters is a word or non-word and measure the response time to make that decision, the idea being that they will respond faster to words they had seen during practice or to words that are related to these familiar words. The elaborative retrieval hypothesis would predict that words related to words seen in practice (mediators) would have a faster response time for the retrieval trials over the study trials, but Lehman and Karpicke (2016) found that there was no interaction between type of practice trial (re-study vs. test) and word type in the lexical decision task. In fact, there was a larger priming effect for mediators in the re-study trials over the retrieval trials (Lehman & Karpicke, 2016). Experiments 3a, 3b, 4, and 5 manipulated the number of mediators in various ways and found that a larger number of mediators was negatively correlated with proportion recalled if the test had cues, and no relationship if the test was

free recall (Lehman & Karpicke, 2016). According to the elaborative retrieval hypothesis, more mediators provide more pathways to retrieve the target word, but Lehman and Karpicke (2016) found no effects of mediators or found the opposite effect.

Episodic Context Hypothesis

The other major hypothesis relevant to this study is the *episodic context hypothesis* proposed by Karpicke et al. (2014). The authors define context as incomplete information that is encoded at the same time that a specific item is encountered, and this incomplete information can include aspects of the external environment and internal mental state (Karpicke et al., 2014). An example of incomplete information coming from the external environment would be the color of the words in a cue-target pair scenario or even something bigger such as the layout of the workspace in which a participant is sitting. Participants focus on the word they have to learn, but their memories absorb their surroundings along with the word, but because the surroundings are not the focus, the information stored in memory about the surroundings is incomplete.

In addition to absorbing elements of the external environment, participants can also maintain elements of their internal environment in the memory for the target information. Klein, Shiffrin, and Criss (2007) proposed that context can include bodily functions and cognitive strategies that can be observed or not easily observed. For example, participants may create a story or mnemonic device when trying to learn word pairs to make it easier to retrieve the target word, even if they were not explicitly asked to do so by the experimenter and the experiment does not directly manipulate internal or

external context. As with incorporating elements of the external environment into the memory for the target information, participants can also incorporate elements of their internal environment.

At the core of the episodic context hypothesis is the idea that retrieving an item revises the memory of the context surrounding that item, which makes it easier to retrieve the item later (Karpicke et al., 2014). Time passes with each retrieval attempt and the context surrounding the item also changes, therefore the representation of the context is updated with each retrieval attempt, and items in memory that have not been updated as much are harder to retrieve because the context cues associated with the item are not up to date (Karpicke et al., 2014). As an example, after successfully retrieving an item once, it should be easier to retrieve the item again on the next trial than on the tenth trial. The context around the first trial is more similar to the context around the second trial than the context around the tenth trial unless the memory for the context is allowed to update on each subsequent trial. If the item is successfully retrieved on the tenth trial, the context is updated. Thus, the episodic context hypothesis focuses on temporal context and how easy it is to retrieve an item based on how much time has passed and how much the context has changed since the last time the context was updated (Karpicke et al., 2014).

There has been some support for the episodic context hypothesis. Lehman, Smith, and Karpicke (2014) sought to disentangle the benefit seen from retrieval practice and the elaboration that occurs under the elaborative retrieval hypothesis. Participants studied five different lists of words. In the re-study group, participants studied a list and did a distractor task after each list for the first four lists. The retrieval practice group had

participants study a list and retrieve the words from that list for the first four lists, and the elaboration group studied a list and created two mediators for each word in the list for the first four lists. For the fifth list, participants studied it, recalled the list, and then had to remember as many words as possible from all of the lists (Lehman et al., 2014). The authors predicted that elaboration would expand the search set and create interference with retrieving the target information rather than forming multiple pathways to that information. They also predicted that simple retrieval practice would reduce the size of the search set and make it easier to retrieve the target information in support of the episodic context hypothesis (Lehman et al., 2014).

The results were that participants who practiced retrieving the lists were able to recall more items from the fifth list while those in the elaboration group recalled the least from list five (Lehman et al., 2014). The retrieval practice group also had less interference in recalling list five from previous lists and the elaboration group had the most intrusions (Lehman et al., 2014). For the final recall of all lists, the retrieval practice group remembered the most words overall, and the re-study group remembered the least (Lehman et al., 2014). The authors claim that these results support the episodic context hypothesis because interference from all the lists was reduced for the retrieval practice group, meaning the participants created a context from the words in each original list and reinstated that context when they had to retrieve a particular list.

Whiffen and Karpicke (2017) also found evidence for the episodic context hypothesis. In the first experiment, participants studied three blocks containing two lists of six words each. They studied the first list of six words and did a 30 second distractor

task, then studied the second list of six words and completed the distractor task again. This was the format for all three blocks. After the study blocks, the re-study group studied the three blocks in the same way while the list discrimination group was shown a word from the lists and had to indicate whether the word was from the first or second list in each respective block. At the end all participants had to recall as many words as possible from all of the lists. Experiment 2 added a group that had to study the words in each list and provide pleasantness ratings to contrast episodic recall (list discrimination) with semantic encoding (pleasantness rating; Whiffen & Karpicke, 2017). Experiment 3 had participants study lists of related words instead of unrelated words, and the groups consisted of the re-study group, list discrimination group, and a group that studied the words and identified the taxonomic category for each list, or the general item that the words in each list described (Whiffen & Karpicke, 2017).

The results in Experiment 1 were that participants recalled more words in the final free recall test when they had to reinstate the context around the original encoding or study session by clicking on which list they remembered the words originally coming from (Whiffen & Karpicke, 2017). The authors also did an analysis on how well free recall responses clustered around each study block and found that the list discrimination group clustered the individual lists better than the re-study group. Experiment 2 results mirrored those in Experiment 1 and the pleasantness rating group also recalled more words than the re-study group. There was not a significant difference in recalled words between the list discrimination and pleasantness rating groups. Clustering was most prevalent in the list discrimination group and least prevalent in the pleasantness ratings

group. In Experiment 3, the list discrimination and semantic category groups recalled more words over the re-study group and those groups did not differ from each other. Clustering was most prevalent in the list discrimination group and least prevalent in the semantic category group, but an analysis of category clustering showed that the semantic category group recalled the most words by semantic category than the other groups. These results provide evidence that participants' memories absorb the context around the items they focus on (words in a list) according to how they are asked to respond (list discrimination, pleasantness ratings, or semantic category judgments), which supports the episodic context hypothesis.

The Current Study

The elaborative retrieval and episodic context hypotheses have found support and may explain some of the mechanisms than can underlie retrieval. For example, they may help to clarify phenomenon such as tip-of-the-tongue in which people know they have knowledge of an item but cannot immediately retrieve it. In this case, it is useful to remember the original context in which people last remember encountering the item so they can retrieve the actual item, which exemplifies the episodic context hypothesis. The elaborative retrieval hypothesis may provide one strategy for purposely trying to remember specific items, such as students elaborating on a piece of information they want to remember for a test and then testing themselves as a way to study. However, these hypotheses do not explain all instances of attempted memory retrieval. If people are confident that they have a memory of an item and they retrieve it successfully right away, they do not necessarily have to search their memories for the item before they can

retrieve it, and they do not have to reinstate the context around the item before retrieving it. As such, the elaboration- and context- driven hypotheses potentially do not account for retrieval instances in testing effect research in which participants immediately successfully retrieve the target information.

Since the discussed hypotheses do not explain all retrieval instances, evidence is needed to establish whether retrieval works differently in different instances or whether retrieval works the same in all instances but the elaboration and context hypotheses pick up on encoding strategies or other memory processes in addition to the retrieval process. Testing effect experiments conducted so far have not successfully disentangled context and elaboration effects from the retrieval process, or the baseline activation of retrieval itself. *Baseline activation* can be described as a strengthening of the target memory itself. Whereas the elaboration and context accounts explain successful retrieval in terms of multiple pathways to the target information through the use of context cues or mediators, baseline activation describes retrieval as one pathway to the target information that becomes strengthened the more it is used. It is possible that context and elaboration are part of the fundamental process of retrieval, but it is also possible that there is a baseline activation for retrieval, and elaboration and context are separate processes that affect retrieval.

The testing effect finding is robust, therefore researchers make a recommendation to educators and students to conduct practice tests on any information they want to learn instead of re-studying the information. However, researchers do not know with certainty how retrieval works and why it is beneficial for long-term retention. This study attempted

to clarify the mechanisms underlying retrieval by disentangling elaboration and context from the retrieval process to provide support either for the elaboration and context hypotheses or for baseline activation. The experiment consisted of a study phase, followed by a practice phase in which participants either re-studied cue-target pairs or practiced recalling the target words. Afterwards, they completed the final test and a lexical decision task where participants had to make a discrimination judgment on whether strings of letter were words or non-words. This task measured accuracy and response time. If participants responded slower to old words, or words that they encountered in the study and practice phases, it provided support for the elaboration and context hypotheses in that they had to use extra time to retrieve the context around the target words before they could retrieve them. However, if they responded to old words faster, they were able to retrieve the target words immediately without also recalling the context or elaborations around the target words, which supported baseline activation.

CHAPTER III

METHODS

Subjects

Twenty-four participants were recruited from the University of Northern Colorado subject pool, and they participated in exchange for class credit. The data for one participant were not used because of formatting issues with the output file from the computer program. As such, the data from 23 participants were used in the analyses.

Materials

A computer program was used to present the experiment on Windows computers provided by UNC in a computer lab. Up to three participants were tested at a time. While the computers were not in individual cubicles, the experimenter separated each participant by at least one work space. Participants were assigned to condition based on a fixed rotation and sat in an office chair at whatever distance from the computer screen was comfortable for them.

Two lists of word pairs were created, and each participant practiced using one of the lists. List A consisted of 30 pairs of words and was taken from Carpenter and Yeung (2017). These authors collected properties of the cues, including concreteness, familiarity, imageability, and how frequently each word is used in everyday language

based on the MRC Psycholinguistic Database. Concreteness of cues, which is a measure of how real versus abstract a word is, fell between 496 and 670 out of a range of 100 to 700 (Carpenter & Yeung, 2017). A higher rating indicates a word that is more concrete, and a lower rating indicates a more abstract word. Familiarity of cues fell between 421 and 636 out of a range of 100 to 700, meaning the words were more familiar as opposed to less familiar (Carpenter & Yeung, 2017). Imageability of cues, which is a measure of how easy it is to conjure a mental image of each word, fell between 600 and 652 out of a range of 100 to 700, which means the words were easy to mentally represent (Carpenter & Yeung, 2017). The cue-target pairs chosen from Carpenter and Yeung (2017) had probabilities of retrieving the target when the cue is presented between .011 and .019 and represents the strength between each cue and its target. Examples of cue-target pairs include Broom-Floor, Kite-Paper, and Shore-Waves (see Appendix B for a list of all List A cue-target pairs).

List B also had 30 pairs of words taken from Carpenter (2009), in which the author took words from Wilson's database with concreteness between 500 and 700. Cue-target strength of the pairs fell between .011 and .017 and were matched as much as possible with the pairs in List A. Any pairs that shared a word with List A were excluded as well as any words between the lists that were similar, such as Bread from List A and Bead from Carpenter (2009) and Grass from List A and Glass from Carpenter (2009). Examples from List B include Building-Stone, Manners-Dinner, and Virus-Doctor. Accuracy was recorded for the participants who were in the practice test group and for all participants in the final test. (see Appendix B for a list of all List B cue-target pairs).

The lexical decision task consisted of 120 words and scrambled words (non-words). The non-words were scrambled in a way that they could still be somewhat pronounceable like a word but were not actual words. Examples of non-words include aroch, edrolfnu, osesh, and veirsl. In List A, there were 60 non-words, 58 new words or words that participants did not see during practice, and two old words or words that participants did see during practice. List B had 60 non-words, 59 new words, and one old word. Examples of new words include orange, ale, elm, and bluegill. The two old words in List A and the one in List B were all cues, not targets. Accuracy and response time were recorded.

Design

This study used a 2 (practice condition) x 3 (word type) experimental design. The independent variables were practice condition (re-study vs. practice test) and lexical decision task word type (non-word, new word, old word). Practice condition was between subjects and word type was within subjects. Twelve participants re-studied the material and 11 did practice tests, and each participant in the lexical decision task saw 60 non-words and either 58 new words and two old words (List A) or 59 new words and one old word (List B). Additionally, there were two variables treated as covariates: List (A or B) and Task Order (Recall-Lexical Decision or Lexical Decision-Recall). Recall-Lexical Decision (LD) signified that participants completed the final test first and the lexical decision task second and LD-Recall signified that they completed the lexical decision task first. The dependent variables were accuracy for the final test and accuracy and response time for the lexical decision task.

Procedure

A protocol was written and sent to the IRB in September 2018. It included a description of this experiment and a possible follow-up experiment that would manipulate context. In October 2018, the IRB sent its approval of the study and gave it an exempt status because of minimal risk to participants.

No more than three participants arrived at the computer lab at a time. After signing the consent form, they were asked to read the instructions on the computer screen and then were asked to summarize what they had to do. The experimenter corrected any misconceptions and verbally summarized the instructions. This was done at the beginning of each of the four phases. Phase one consisted of participants studying the cue-target pairs one pair at a time for six seconds each. The pairs appeared in the upper left quadrant of the screen and were in 16-point font. After six seconds the next pair appeared with no break between them. When all 30 pairs had been displayed three times each over three blocks, the instructions for the second phase appeared. The order that the pairs were presented was random for each block.

In the second phase, participants either re-studied the 30 pairs or practiced recalling the target words. The formatting of the words did not change at all for those who re-studied the pairs. Participants in the practice test group instead saw the cue word followed by a blank where the target word would be with a text box underneath the blank. They were asked to type in the word that they remember being associated with the cue word. The order of the pairs did not change for each block for the practice test group. Both groups saw (or were tested over) the word pairs for another three blocks each. The

participants who re-studied again had six seconds to study each pair, and those who practice tested had six seconds to type the target word. No feedback was provided for the practice test group.

The third and fourth phases consisted of participants either taking the final test first or completing the lexical decision task first. Those who took the final test in phase three completed the lexical decision task in phase four and vice versa. This was done as a counterbalancing measure in case seeing the word pairs again in the final test gave participants an advantage in the lexical decision task. The final test looked exactly the same as the practice test with the cue being displayed next to a blank and a text box underneath the blank where participants could type in the word they remembered being associated with the cue word. They typed in their response for each pair once with no blocks contrary to the study and practice phases, and they had six seconds to type. The responses to the final test were later scored by hand to give credit to responses that were misspelled but otherwise correct. The lexical decision task displayed a string of letters for 10 seconds, and the participants were instructed to indicate as quickly and accurately as possible whether the string of letters was a valid English word or not. There were 120 non-words and words altogether, and each stimulus appeared immediately after a response was given to the previous one.

CHAPTER IV

RESULTS

This experiment attempted to replicate the testing effect and discover the mechanism underlying memory retrieval that results in superior performance in accuracy on a final test. Measuring not only accuracy but also response time on a task that presents familiar and unfamiliar words can help clarify whether a context is created around information to be retrieved and whether that context is necessary for retrieval. If context is needed for retrieval, participants should respond slower to familiar words in the lexical decision task, which implies that they had to take extra time to retrieve the context first in order to retrieve the target information. The results are split into two major analyses. The first analysis tests the replicability of the testing effect by measuring the final test proportion correct, and the second analysis tests for differences in proportion correct and response time between the different word types in the lexical decision task. All analyses were done using SPSS and declared significant at the .05 significance level.

Final Test Proportion Correct

A One-way ANOVA was performed for final test proportion correct with Practice condition (re-study or test) as the single factor. Test order (Recall-LD or LD-Recall) and List (A or B) were added in as covariates to ensure that the lists were equal in difficulty and that those who completed the final test first did not receive an advantage in the lexical decision task. Levene's Test of Equality of Error Variances was not significant,

F(1, 21) = .000, p = .996, signifying that the error variance for final test proportion correct is equal across the two groups. The overall analysis for final test proportion correct after including the covariates was not significant, F(3, 19) = 2.886, p = .063, Sum of Squares (SS) = .363, partial r^2 = .313, as such the main conclusion is that there was no difference between re-study and practice test groups (see Figure 1). This study did not find that practice testing led to better results on the final test over re-studying.

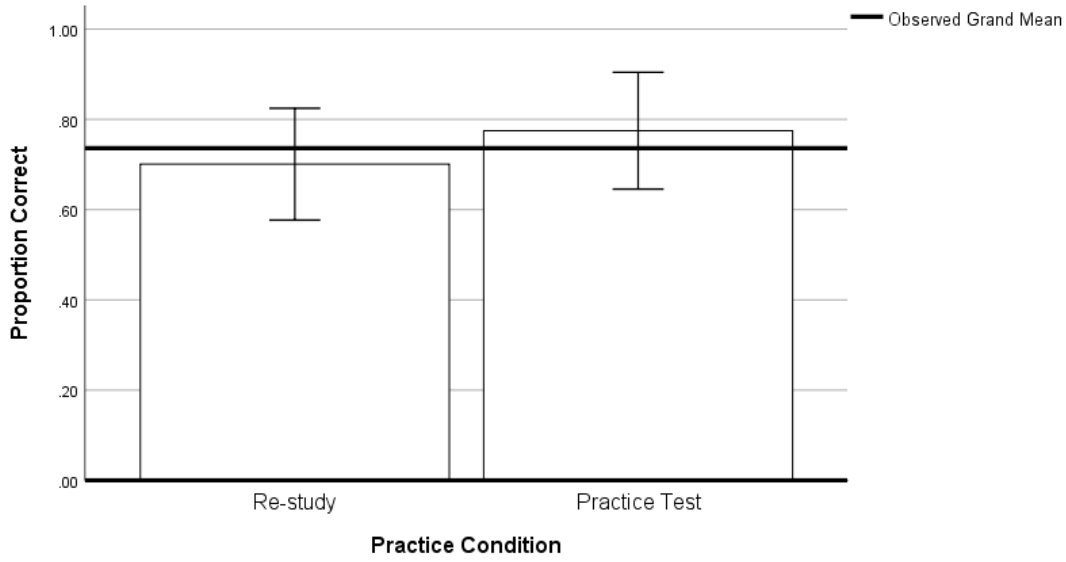

Figure 1. Proportion correct for the final test as a function of Practice Condition.
Note. Error bars represent the 95% confidence interval.

If the model had been significant, Test order would have been significant, $F(1, 19) = 7.296$, $p = .014$, SS = .306, partial $r^2 = .277$, suggesting that there was an appreciable difference between whether participants saw the final test first or the lexical decision task first. List would not have been significant, $F(1, 19) = .477$, $p = .498$, meaning the lists were essentially equal in difficulty, and Practice condition would not be significant either, $F(1, 19) = .750$, $p = .397$, SS = .031, partial $r^2 = .038$. The mean for the re-study group was .701 and standard error was .059, while the practice test group had a mean of .775 with a standard error of .062. There was a trend toward a testing effect, but it did not reach significance (see Figure 1).

Lexical Decision Task

The analysis for the lexical decision task was a mixed-factorial repeated measures ANOVA with Word type (Non-word, New word, Old word) as the within-subjects factor and Practice condition (re-study or test) as the between-subjects factor with List (A or B) and Test order (Recall-LD or LD-Recall) as covariates. The dependent variables were proportion correct and response time measured in seconds. Mauchly's Test of Sphericity was conducted for both dependent variables. Proportion correct had a chi-square value of 1.542 and a p value of .462, while the chi-square value for response time was 11.421 with a p value of .003. As such, the sphericity assumption was violated for response time so the Huynh-Feldt adjustment was used to report the statistics for response time. Sphericity is the assumption for repeated measures tests that the variances between pairs of trials are equal. Mauchly's Test of Sphericity is a test that compares the variances between trials and states whether the differences are statistically significant. If the differences are

significant, sphericity is violated and the degrees of freedom have to be adjusted to a more conservative level because the violation inflates the Type 1 error rate, which is rejecting the null hypothesis when it is true in actuality.

Proportion Correct

The main effect for Word type was not significant, $F(2, 38) = .432$, $p = .652$, $SS = .010$, partial $r^2 = .022$, which means there was not a difference in proportion correct between nonwords, old words, or new words (see Figure 2). The mean for Non-words was .858 and the standard error was .026, for New words the mean was .916 with a standard error of .018, and the mean for Old words was .979 with a standard error of .022. Since the overall model was not significant, no definitive conclusions can be drawn about tests of significance for Word type pairs. However, if the model had been significant, the comparison between Non-words and Old words, $p = .005$, would have been significant based on Bonferroni's test of multiple comparisons. The reason that the repeated measures ANOVA was not significant but the pairwise comparisons were significant could be because of the low sample size for the Old words, and comparing a group with high variability (the Non-words and New words) to a group with low variability (the Old words) can cause it to be significant. The other main effect of Practice condition was also not significant, $F(1, 19) = .820$, $p = .376$, $SS = .009$, partial $r^2 = .041$, and the re-study group had a mean of .906 with a standard error of .017 while the practice test group had a mean of .929 with a standard error of .018 (see Figure 3). The main conclusion is that overall proportion correct was not significant for Word type or Practice

condition, but according to the means for each of those groups, there was a trend for old words to be more accurate than nonwords and new words, and for the practice test group to be more accurate.

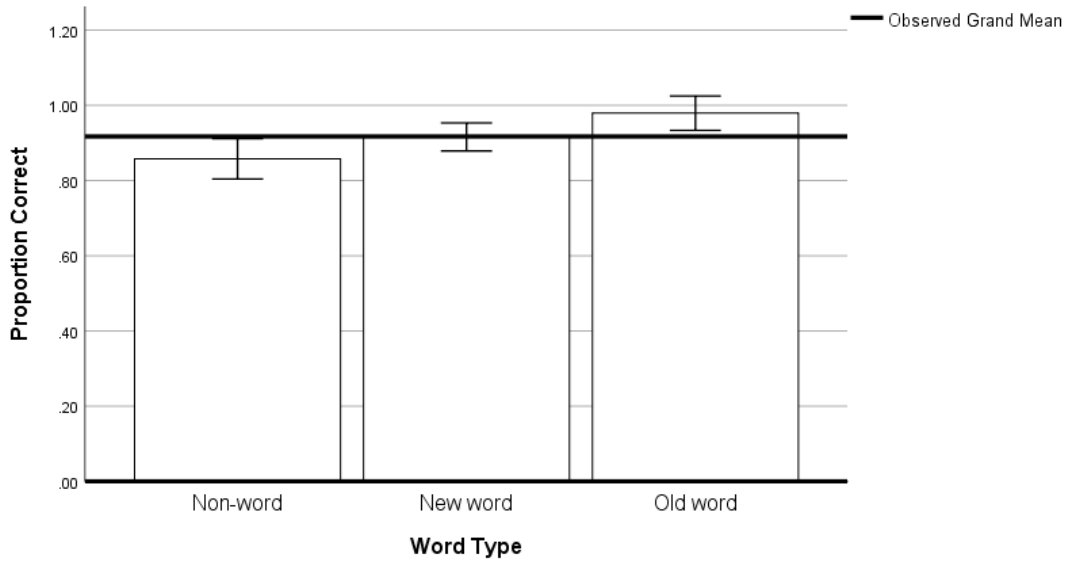

Figure 2. Proportion correct for the lexical decision task as a function of Word Type. *Note*. Error bars represent the 95% confidence interval.

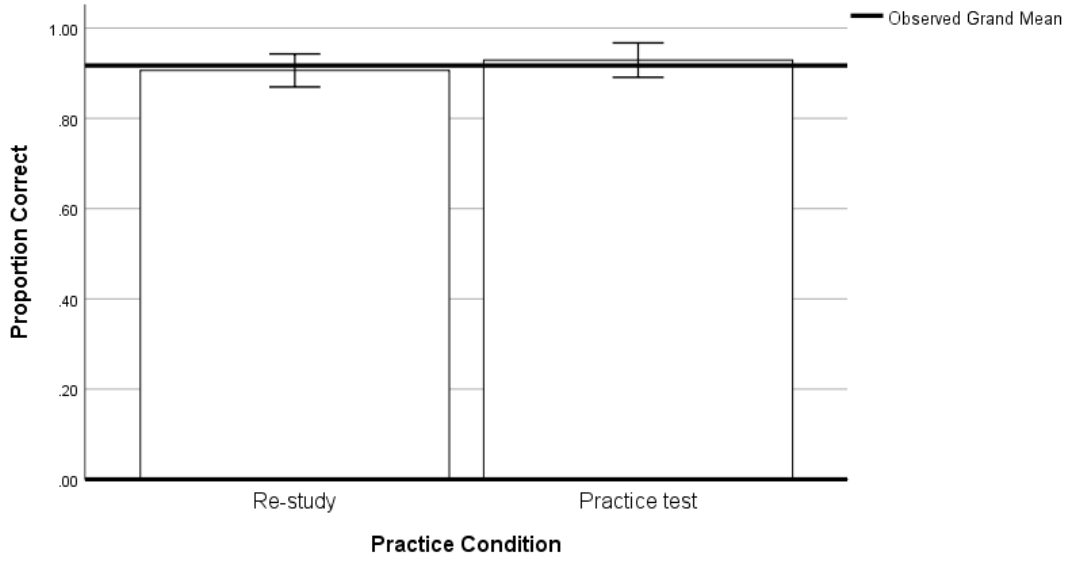

Figure 3. Proportion correct for the lexical decision task as a function of Practice Condition.
Note. Error bars represent the 95% confidence interval.

None of the interactions between Word type and Practice condition, $F(2, 38) = 1.601$, $p = .215$, $SS = .036$, partial $r^2 = .078$; see Figure 4), Word type and Test order, $F(2, 38) = .470$, $p = .629$, and Word type and List, $F(2, 38) = .224$, $p = .801$, were significant, however, there was a trend for the practice test group to be more accurate in identifying old words and new words than the re-study group (see Table 1). If this trend were to become significant, it would lend support to the prediction that practice testing does lead to better recognition of practiced words.

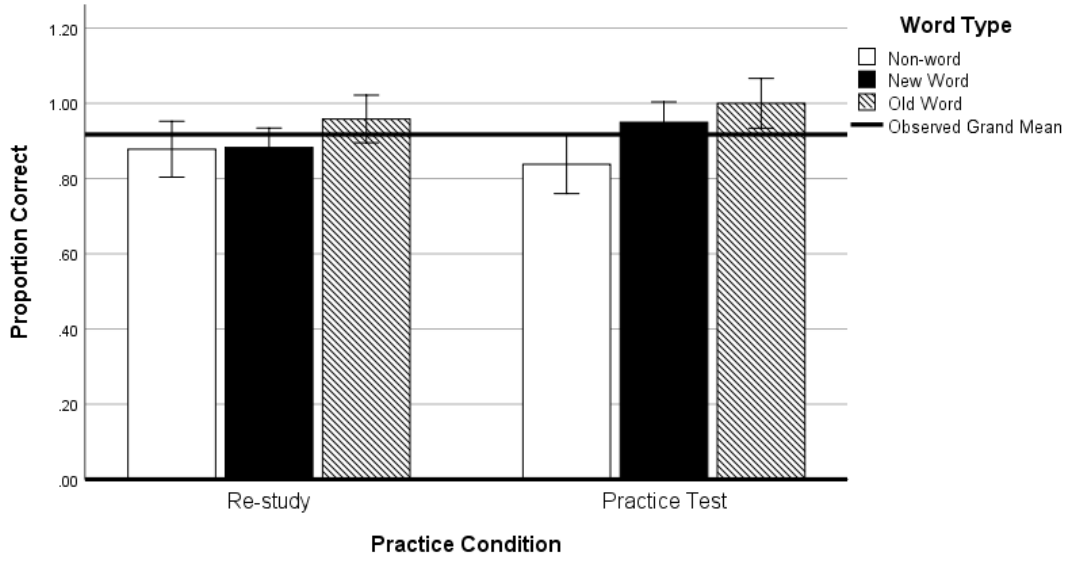

Figure 4. Proportion correct for the lexical decision task as a function of Practice Condition and Word Type.
Note. Error bars represent the 95% confidence interval.

Table 1

Proportion Correct of Re-study and Practice Test Groups

	Re-study		Practice Test	
	Mean (SE)	95% CI	Mean (SE)	95% CI
Non-word	.878 (.035)	[.804, .952]	.838 (.037)	[.760, .915]
New word	.882 (.025)	[.831, .934]	.950 (.026)	[.896, 1.004]
Old word	.958 (.030)	[.895, 1.022]	1.00 (.032)	[.934, 1.066]

Note: SE = Standard Error; CI = Confidence Interval

Response Time

The main effect of Word type on response time was not significant, F(1.661, 31.557) = 2.992, p = .073, SS = .132, partial r^2 = .136, and the mean for nonwords was .847 with a standard error of .030, for new words the mean was .765 with a standard error of .019, and the mean for old words was .669 with a standard error of .038 (see Figure 5). Since the omnibus analysis was not significant, significant pairwise comparisons are not considered, but if the model had been significant, the comparisons between nonwords and new words, p = .031, nonwords and old words, p = .014, would have been significant using Bonferroni's test of multiple comparisons. No definitive conclusion can be drawn, but there was a trend for nonwords to be the slowest, followed by new words, and old words were the fastest. The other main effect of Practice condition was also not significant, F(1, 19) = .046, p = .833, SS = .001, partial r^2 = .002. The mean for the re-study group was .757 with a standard error of .023 and the mean for the practice test group was .764 with a standard deviation of .024 (see Figure 6). Based on the means, there was a trend for the practice test group to be slower in distinguishing between words and nonwords than the re-study group in the lexical decision task. The overall conclusion is that response time was not significant for Word type or Practice condition, but there are possible trends that show old words being responded to faster than new words and nonwords and the practice test group responding slower than the re-study group in general if the model had been significant. List would have been significant, F(1, 19) = 5.241, p = .034, suggesting that the list that participants practiced using (A or B) may have had an effect on their response time in the lexical decision task.

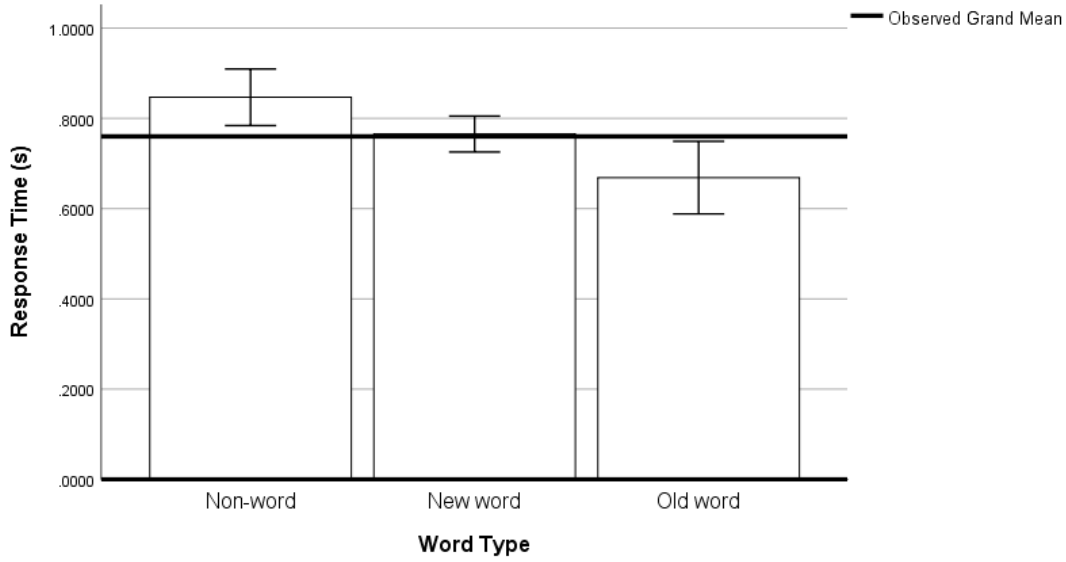

Figure 5. Response time in seconds for the lexical decision task as a function of Word Type.
Note. Error bars represent the 95% confidence interval.

Figure 6. Response time in seconds for the lexical decision task as a function of Practice Condition.
Note. Error bars represent the 95% confidence interval.

None of the interactions between Word type and Practice condition, F(1.661, 31.557) = .626, p = .513, SS = .028, partial r^2 = .032; see Figure 7), Word type and Test order, F(1.661, 31.557 = 1.827, p = .182), and Word type and List, F(1.661, 31.557) = .249, p = .740, were significant. There was a trend for the practice test group to respond to old words and nonwords more slowly than the re-study group, supporting the hypotheses based on elaboration rather than the baseline activation hypothesis. The data suggest those who did retrieval practice needed extra time to retrieve the elaborated memories surrounding the presented old word to eventually retrieve the familiar word itself (see Table 2).

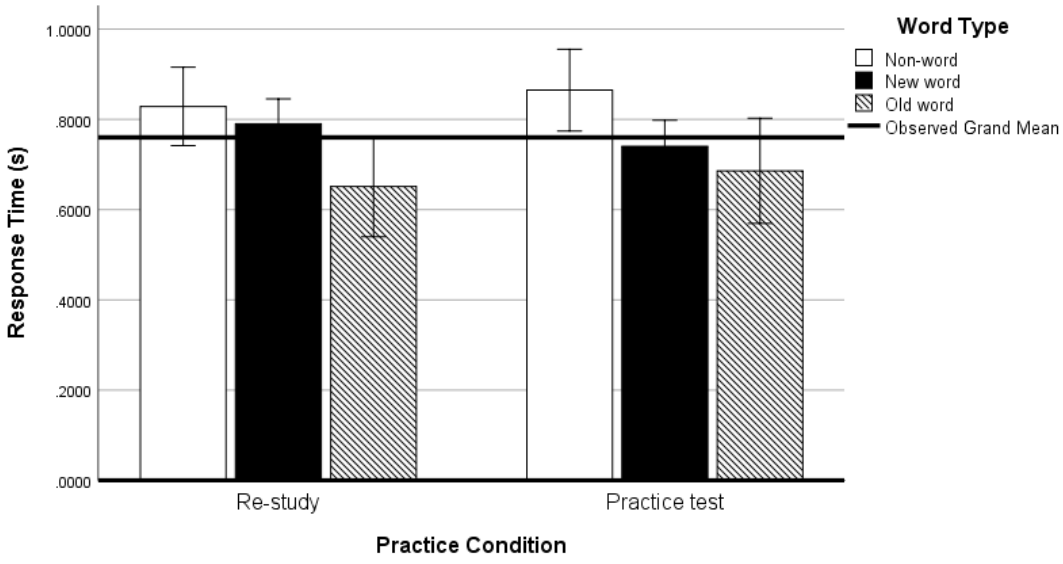

Figure 7. Response time in seconds for the lexical decision task as a function of Practice Condition and Word Type.
Note. Error bars represent the 95% confidence interval.

Table 2

Mean Response Time in (s) of Re-study and Practice Test Groups

	Re-study		Practice Test	
Word Type	Mean (SE)	95% CI	Mean (SE)	95% CI
Non-word	.829 (.041)	[.742, .915]	.865 (.043)	[.774, .956]
New word	.790 (.026)	[.735, .845]	.740 (.028)	[.683, .798]
Old word	.651 (.053)	[.540, .763]	.686 (.056)	[.570, .802]

Note: SE = Standard Error; CI = Confidence Interval

CHAPTER V

GENERAL DISCUSSION

Summary

This experiment attempted to provide evidence in support of either the elaboration and context accounts of the testing effect or baseline activation. None of the results were significant, but there was a trend for participants in the retrieval practice group to have a higher proportion correct on the final recall test than those who re-studied the word pairs. A trend that supported the elaboration and context hypotheses was that participants in the practice test group were slower but more accurate to classify old words as words in the lexical decision task over the re-study group. If this result had been significant, it would provide evidence that participants reinstated the context around the target words when asked to retrieve them, or they created elaborations around the target words that they had to retrieve before they could remember the actual target words on the final test.

Many theories that explain how retrieval works make the assumption that any item in memory has a type of context surrounding it (Atkinson & Shiffrin, 1968), whether the context is purposeful as with the elaborative retrieval hypothesis or is involuntary and incomplete as with the episodic context hypothesis. When the item is retrieved from memory, the context connected to it can also be recalled. However, people can also

sometimes retrieve the needed information without needing to search the context around it. In this case, it is possible that the only pathway in memory that is activated is the one leading to the target information because that pathway is strong enough by itself. The results of this experiment do not support baseline activation and instead support the theories that claim context is fundamental to the retrieval process.

An alternative explanation for the results in this study is embodied in the transfer appropriate processing hypothesis, which states that the testing effect exists because the methodology gives an advantage to the practice test group. Participants who have to retrieve the target information during practice, and retrieve it in the same manner during the final test, have seen how the final test is formatted (Duchastel & Nungester, 1982). Completing multiple-choice practice tests leads to better retention as measured by the final test because the final test is also multiple-choice, whereas the re-study group was not exposed to the format of the final test. Applied to the current experiment, the practice test group had to recall the target word during the practice phase, and the final test had the same requirement. There may have been a trend supporting the testing effect simply because those in the practice test group used the same retrieval mechanism in the practice phase and the final test.

The trend found in the lexical decision task could be explained by typical testing behavior by the participants such as demand characteristics. Students spend a large portion of their lives in school and taking tests. They are trained to try their best to not respond with an incorrect answer. The participants in this experiment were college students and the experiment focused on taking tests. During the lexical decision task, they

could have responded slower to old words because they were double checking to make sure they were about to respond correctly that the old word was in fact a word. They could have been thinking that because the old word was immediately familiar, but could have been altered slightly such that the correct answer would be that it was in fact a non-word, that they slowed down to make sure they were about to respond correctly.

Limitations

There were some limitations with the experiment. Out of 60 words in the lexical decision task, only two in List A and one in List B were old words, therefore the power is low for those words in the lexical decision task analyses. Since there were so few old words, they were not adequately represented in the analyses. If they were not adequately represented, the variability of responses to the old words was also not representative compared to if participants had been allowed to vary their responses with more old words. A way to address this issue is to either have 40 of each word type (non-words, old words, new words) or 60 non-words, 30 old words, and 30 new words. Another issue was that, while cue-target pairs were presented in a random order for each block during the practice phase for the re-study group, the order of the pairs was the same for each block in the practice test group. This may have prevented the participants who completed the practice tests from fully benefiting from the retrieval process. Another issue was that participants in the practice test group did not receive feedback for wrong answers, thus if they responded with an incorrect target word they tended to continue getting that same target word wrong for the other two practice blocks. There is evidence that receiving corrective feedback is important for finding a testing effect (Kang, McDermott, & Roediger, 2007).

Future Directions

A possible follow-up study to explore the effectiveness of using context as a cue to aid retrieval is to directly manipulate the context by manipulating font or background color or the workspace in which participants sit. The number of context cues could also be a possible manipulation, with the idea that more cues could result in more successful retrieval attempts. Another possible follow-up experiment is to suppress the ability to encode context by asking participants to study words while they are engaged in articulatory suppression to prevent them from encoding any mental context that might aid in later retrieval. Evidence has been found that articulatory suppression, which is verbally reciting something such as the word "the" over and over, can impair accuracy for typing numbers that are spelled out as words (Kole, Healy, & Buck-Gengler, 2003). As such, it is possible that articulatory suppression can also impair the ability to encode the context in a given situation.

Conclusion

In conclusion, much evidence has been found for the testing effect, but why retrieval practice leads to better retention has not been elucidated. The trends found in this experiment support hypotheses that claim purposeful elaboration or incidental context are a fundamental part of the process of retrieval based on the result that retrieval practice led to a slower but more accurate response to discriminating familiar words from non-words and unfamiliar words. The implications of this research are that it can refine the recommendations made to educators and students by clarifying how testing should be used for maximal learning. Rather than simply recommending that students should

practice test as a way to study, the recommendation can also include whether students should elaborate on the information before they practice test or if they should try to test themselves in the same, similar, or different settings.

REFERENCES

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.), *The psychology of learning and motivation* (pp. 89–195). New York, NY: Academic Press.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1563-1569. Doi: 10.1037/a0017021

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & cognition*, *34*, 268-276. Doi: https://doi.org/10.3758/BF03193405

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language, 92*, 128-141. Doi: http://dx.doi.org/10.1016/j.jml.2016.06.008

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin, 144*, 1111-1146. Doi: 10.1037/bul0000166

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research*, *75*, 309–313. Doi: 10.1080/00220671.1982.10885400

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528-558. Doi: 10.1080/09541440601056620

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *Psychology of learning and motivation* (Vol. 61, pp. 237-284). Academic Press.

Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III*. New York: Psychology Press.

Kole, J. A., Healy, A. F., & Buck-Gengler, C. J. (2003): Does number data entry rely on the phonological loop? *Memory, 13*, 388-394. Doi: 10.1080/09658210344000224

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 1573-1591. Doi: http://dx.doi.org/10.1037/xlm0000267

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1787-1794. Doi: 10.1037/xlm0000012

Pansky, A. (2012). Inoculation against forgetting: Advantages of immediate versus delayed initial testing due to superior verbatim accessibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1792-1800. Doi: 10.1037/a0028460

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

Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology*, *36*, 1710-1727. Doi: 10.1080/01443410.2014.953037

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1036-1046. Doi: 10.1037/xlm00003

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL LETTER

Institutional Review Board Approval Letter

DATE:	October 22, 2018
TO:	James Kole, PhD
FROM:	University of Northern Colorado (UNCO) IRB
PROJECT TITLE:	[1325997-1] Mechanisms Underlying the Testing Effect
SUBMISSION TYPE:	New Project
ACTION:	APPROVAL/VERIFICATION OF EXEMPT STATUS
DECISION DATE:	October 22, 2018
EXPIRATION DATE:	October 21, 2022

Thank you for your submission of New Project materials for this project. The University of Northern Colorado (UNCO) IRB approves this project and verifies its status as EXEMPT according to federal IRB regulations.

Dr. Kole -

Thank you for your patience with the UNC IRB process. Your protocols and materials are verified/ approved exempt.

Please update the logo in the letterhead and the contact information for mistreatment as a research participant to Nicole Morse (Sherry May retired in summer 2018) in the consent form before use in the study.

Best wishes with this study.

Sincerely,

Dr. Megan Stellino, UNC IRB Co-Chair

We will retain a copy of this correspondence within our records for a duration of 4 years.

If you have any questions, please contact Nicole Morse at 970-351-1910 or nicole.morse@unco.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Northern Colorado (UNCO) IRB's records.

APPENDIX B

LIST A AND LIST B CUE-TARGET PAIRS

List A

	Cue	Target	C-T strength
1	Ball	Boy	.011
2	Blood	Skin	.011
3	Bomb	Fire	.013
4	Bread	Meat	.013
5	Broom	Floor	.016
6	Chair	Bed	.013
7	Clock	Radio	.012
8	Coin	Bill	.012
9	Deer	Woods	.014
10	Feast	Party	.018
11	Golf	Grass	.014
12	Harp	Flute	.012
13	Heart	Body	.016
14	Horse	Dog	.012
15	Jail	Thief	.013
16	Kite	Paper	.017
17	Knife	Gun	.013
18	Lamb	Wolf	.016
19	Lunch	Pail	.013
20	Moose	Bull	.018
21	Neck	Bone	.018
22	Night	Train	.019
23	Nurse	Needle	.014
24	Rock	Mountain	.019
25	Roof	Rain	.016
26	Shore	Waves	.014
27	Skunk	Stripe	.016
28	Snake	Spider	.012
29	Soap	Cloth	.011
30	Truck	Bus	.014

List B

	Cue	Target	C-T strength
1	Barn	House	.016
2	Barrier	Fence	.014
3	Bay	River	.013
4	Building	Stone	.017
5	Comet	Planet	.014
6	Contest	Money	.015
7	Desert	Island	.015
8	Directions	Street	.013
9	Flick	Brush	.013
10	Hole	Circle	.016
11	Ice	Drink	.016
12	Jacket	Shirt	.013
13	Leaf	Flower	.012
14	Lips	Teeth	.014
15	Lounge	Hotel	.014
16	Maid	Dress	.011
17	Manners	Dinner	.014
18	Mist	Water	.013
19	Mitten	Child	.011
20	Morning	Light	.014
21	Print	Letter	.014
22	Pupil	School	.016
23	Raft	Beach	.011
24	Rights	Court	.015
25	Seed	Fruit	.011
26	Speak	Mouth	.017
27	Spin	Dance	.013
28	Steam	Coffee	.014
29	Theater	Music	.014
30	Virus	Doctor	.013