THESIS

FORGETTING VERSUS FACILITATION: THE FATE OF NONTESTED INFORMATION IN THE TESTING EFFECT

Submitted by

Lauren Elizabeth Bates

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2015

Master's Committee:

Advisor: Edward DeLosh

Benjamin Clegg
Daniel Robinson

ABSTRACT

FORGETTING VERSUS FACILITATION: THE FATE OF NONTESTED INFORMATION IN THE TESTING EFFECT

The testing effect is an established memory phenomenon that demonstrates that retrieval enhances memory relative to restudying. Testing effects can be both direct and indirect. One example of an indirect effect of testing is retrieval-induced forgetting (RIFO), in which taking a test on a subset of information can actually impair recall of related, but nontested information. Recent research has also demonstrated retrieval-induced facilitation (RIFA), the opposite pattern, in which testing on a subset of information enhances memory for related but non-tested information. The present study sought to determine the key factors that determine whether the indirect testing effects on nontested information takes the form of forgetting versus facilitation. Both experiments examined memory for cue-target pairs and vary whether the final test is a cued recall test or a free recall test. Experiment 1 did so for category-exemplar pairs, in which each category cue was paired with several category exemplars, and varied the retention interval as well. Experiment 2 used a construction in which cue words were paired with multiple, unrelated targets. While the results of Experiment 1 supported the hypothesis that a free recall final test would elicit a facilitative indirect testing effect and a cued recall final test would elicit a forgetting effect, Experiment 2 did not follow this pattern. While we are able to draw some conclusions about how certain parameters (e.g. final test type) can aid the presence of either effect, additional avenues should be pursued.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

A large body of literature has demonstrated that testing, or retrieval practice, is a reliable way to enhance memory (Roediger & Karpicke, 2006b; Rowland, 2014). Taking an initial test facilitates recall on a final test, relative to simply restudying the information. This phenomenon, known as the testing effect, has been demonstrated both in the lab and in the classroom (Roediger & Karpicke, 2006a; McDaniel & Fisher, 1991). Due to the negative connotations that many individuals associate with the word “testing,” the term “retrieval practice” is preferred when discussing this phenomenon. For the purpose of this paper, both “testing” and “retrieval practice” will be used interchangeably and both refer to the act of taking a test. Having established the testing effect as a robust phenomenon, current research in this area has moved from simply demonstrated the effect to determining the parameters and underlying mechanisms involved in this phenomenon.

An important distinction pertains to the direct versus indirect testing effects. The finding that retrieval practice on an item boosts memory for that item is an example of a direct effect (Roediger, Putnam, & Smith, 2011). Indirect testing effects refer to potential effects on other aspects of learning. A few examples of these include learners’ metacognition and study strategies. The indirect effect of interest in the present study is the fate of nontested information when information from the same learning episode is subjected to a test.

Two theories have emerged to account for the fate of nontested information when related information is subjected to a practice test. The more positive outcome stems from work done by Chan, McDermott, and Roediger (2006). In their study, participants were asked to read prose passages on a certain topic (e.g., toucans). After their 25 min learning phase had passed, they were randomly assigned to either engage in retrieval practice on the information they had learned, or they were given an extra study opportunity. Participants in the retrieval practice condition were given questions about the article, one at a time with no feedback. Matching for time spent on either task, participants in the restudy condition were shown the same questions, but with the answers included. Therefore, they were not asked to come up with the answer themselves – they were simply allowed to see the information a second time in a slightly different form. On a final test, all participants were given 40 questions on the learning materials. Results showed that participants who engaged in retrieval practice not only had higher performance on questions that had been practiced earlier, but they also did better on questions related to the practiced questions, though these had not explicitly been shown during the retrieval practice phase. Therefore, a benefit emerged for the act of retrieval practice in enhancing memory for nontested information that was related to tested information.

In a well-known study by Anderson et al. (1994), participants studied a series of category – exemplar word pairs. Half of the categories were assigned to a retrieval practice (Rp) group, whereas the other half served as control categories that did not receive retrieval practice (Nrp). Of the Rp categories, half of the exemplars were included in the retrieval practice phase (referred to as Rp+ items) and the other half of the exemplars were not included in the retrieval practice phase (Rp-). This design allowed the researchers to compare memory for items that belonged to tested categories but were not themselves tested, to control items that were not tested, nor were they in the tested categories. Results showed that memory for the nontested items from practiced categories (Rp- items) was poorer than memory for the control (Nrp) items. This phenomenon has been termed retrieval-induced forgetting (RIFO). Based on this finding, Anderson et al.

posited that the act of engaging in retrieval practice for some category items inhibited memory for categorically-related nontested items.

Anderson (2003) stated that interference can be to blame when the retrieval of a memory is harmed by competing associations. A response may be inhibited by other, appropriate yet weaker responses. Essentially, this could explain the results from Anderson et al. (1994) in that the cued recall final test format could have led to competing associates during the final retrieval phase. This would support a cue-dependent model of interference. However, Anderson and Spellman (1995) have demonstrated that an inhibitory mechanism may be at work in RIFO, rather than strictly interference. They were able to demonstrate a RIFO effect using independent cues. This finding has not always been replicated (Rowland, Bates, & DeLosh, 2014), but it does allow for a different interpretation of these effects. Seeing interference as a mechanism for forgetting may lead us to postulate that final test format may be a way to elicit either a RIFO or RIFA effect. However, other research has sought to investigate other mechanisms behind these indirect testing effects.

Chan (2009) identified two possible factors that may determine whether RIFO or RIFA emerges. In a series of experiments examining both effects, he concluded that retention interval and integration of materials are two key factors in determining whether there is forgetting or facilitation of nontested information. Retention interval was manipulated so that 20 minutes between the intervening phase and final test phase was considered a short delay, and 24 hours was considered a long delay. Integration was manipulated in a variety of ways. In the first experiment, participants studied a prose passage. In the high-integration condition, sentences were presented one at a time in the order they were presented in the original paragraph, while in the low-integration condition, sentences were randomized and shown one at a time. Participants in the high integration were explicitly told they would be reading an article and should integrate the information given to them, while participants in the low integration condition were simply told they would be reading a series of sentences. The results from the first experiment showed a RIFO effect with a short delay (i.e. 20 minutes) and low integration, and a RIFA effect with a long delay (i.e. 24 hours) and high integration.

The second experiment in Chan's 2009 paper had participants study sentences that included an object and a location (e.g. the fork was in the nursery). Participants in the high integration condition were asked to integrate the materials while those in the low integration were not. Those in the high integration condition were also instructed to form a mental image of the object-location pair given in the sentence. During retrieval practice, participants were cued with the original sentence with only the first two letters of the object given (e.g. the fo_____ was in the nursery). The final test was formatted just as the retrieval practice phase (i.e. cued recall), as was done in the first experiment. Results demonstrated a similar pattern, in that the RIFO effect emerged when participants in the low integration condition experienced a short delay, and a RIFA effect emerged when participants in the high integration condition experienced a long delay. Therefore, Chan (2009) concluded that both integration of materials and retention interval were crucial when eliciting either effect.

These parameters come into question when we consider what more recent research has found. Rowland and DeLosh (2014) were able to produce a facilitation effect after a short delay with no integration of materials. That is, participants' retrieval of nontested information was facilitated when they studied word pairs that were not at all connected in terms of their meaning. The original argument put forward by Chan in 2009 stated that facilitation occurs when there is a long delay (i.e., 24 hours) between the initial test and final test, and when participants utilized integrated encoding of the to-be-tested materials. He also made the argument that forgetting occurs when there is a short delay (i.e., 20 minutes) and when conditions make it more difficult for participants to engage in integrated encoding. The study by Rowland and DeLosh (2014) refutes this claim, for facilitation was seen after a very short delay (i.e., 5 minutes), and there was no inclusion of an integration manipulation. Follow-up research from the same lab has also supported the replicability of the RIFO effect (Rowland et al., 2014). Given that both RIFA and RIFO have been demonstrated in the laboratory, the next step in investigating these effects is determining the conditions and factors that produce one versus the other.

Rowland and DeLosh (2014) identified another potentially important factor: the type of final test. One of the few major differences between the methods of these experiments was the nature of the final task. In both experiments, participants first completed the learning phase in which they were exposed to word pairs to be studied. Next, they were tested on a subset of that information (Rp). After a delay in which a distractor task was presented, the final test phase began. The first study by Rowland and DeLosh in 2014 had participants freely recall as many items as they could remember from the experiment during the final test phase, which resulted in facilitation for nontested items. On the other hand, the study done by Rowland et al. (2014) cued participants with category names in the final task. This simple manipulation could be the key to understanding under which conditions we can elicit either effect, respectively. Note, however, that Rowland and DeLosh (2014) observed these effects of final test in experiments that used lists of individual words. It is therefore not clear whether the type of final test is a critical factor for the paired-associate learning tasks used by Anderson and colleagues (1994).

In the present study, both experiments employed a paired-associate learning task in which learners studied cue-target pairs. Given the potential importance of the type of final test, half of the participants in all three experiments received a cued recall final test, and half received a free recall final test. Following from the Rowland and DeLosh (2014) study in which they examined both related and unrelated word lists, the current study examined the nature of the relationship between cue-target pairs in the learning set. In Experiment 1, category-exemplar pairs were used, with each category cue paired with multiple category exemplars (e.g. FRUIT – BANANA, FRUIT – STRAWBERRY). In Experiment 2, cue-target pairs consisted of unrelated words, but each cue was paired with six different targets (e.g. TABLE – ACHE, TABLE – DOOR). Finally, the first experiment will also manipulate retention interval in which five minutes will be considered a short delay and 20 minutes will be considered a long delay, given that delay was identified as an important factor in the Chan (2009) study.

As the nature of the practice test appears to be the factor with the most influence in the forgetting versus facilitation issue, it is hypothesized that forgetting will emerge in the cued recall conditions while facilitation will emerge in the free recall conditions. The role of delay is less clear, as facilitation has been found at both long and short delays. This leads me to believe that the effect of delay is moderated by other variables, such as final test type. I hypothesize that short delay will show a facilitative effect when final test type is free recall, while long delay will show a forgetting effect with cued recall. For the unrelated words, I believe that a similar pattern will appear, where we see a main effect of final test type. However, should the results show that facilitation consistently emerges when stimuli are unrelated, then this could support the presence of competition between semantically related items as a possible mechanism underlying the forgetting effect.

CHAPTER 2: EXPERIMENT 1

Method

Participants

Sample size was determined using a power analysis with G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009), using an effect size of 0.4 and power of 0.95. This effect size was chosen due to insufficient data from Anderson et al. (1994). Since it was not possible to compute an effect size from their forgetting effect, 0.4 was selected in an attempt to harness a small to medium effect. The software determined the optimal sample size to be 112 participants to detect a small to medium-sized effect. The sample size for experiment one ended up being 262, due to an abundance of participants and highly motivated undergraduate research assistants. The sample consisted of undergraduates from Colorado State University participating for course credit in lower-division psychology courses.

Materials

This experiment utilized the E Prime software package with participants tested one at a time on a personal computer. A total of eighteen word lists with six items each were used, with two of the lists serving as fillers to account for primacy and recency effects, and the other sixteen serving as experimental lists. For each participant, eight experimental lists were randomly selected as ones that receive instances of retrieval practice (Rp), and the other eight did not have a retrieval practice phase (Nrp). Of the Rp lists, half the items were randomly selected as Rp+ items (i.e., subject to retrieval practice), and the other half as Rp- items (i.e., not subject to retrieval practice).

Category selection. Eighteen categories, two of which served as fillers, were drawn from published norms (Marshall & Cofer, 1970). Categories were selected to be as dissimilar as possible to avoid stimulus confusion across categories. For example, with *Fruit* chosen as one category, *Vegetable* was not chosen as another category. Using Marshall and Cofer's norms, interrelatedness between categories was minimized. Categories were also selected to minimize the phonemic similarity of category labels. Category labels were constrained to a single word, which was unambiguous in nature. Lastly, the word frequencies of category labels (Kucera & Francis, 1967) were kept in the low to moderate range, falling between 25 and 100 occurrences per million.

Exemplar selection. For the purposes of this experiment, six strong exemplars were chosen for each of the categories. According to Battig and Montague (1969) category norms, these strong exemplars had an average rank order of 8, ordered by frequency of report. Exemplars were low frequency, non-compound, unambiguous words with an average word frequency of 12 occurrences per million (Kucera & Francis, 1967). No two exemplars within a category began with the same first two letters, ensuring that each cue (i.e., the first two letters of each exemplar) in the practice phase was unique. In addition, the effectiveness of each cue was assessed by measuring versatility (Solso & Juel, 1980), yielding a mean versatility value of 244. Versatility of a set of letters refers to the number of words with that set of letters in that position. For example, the letter versatility of the combination of BA as the first two letters of a given word is 413, because there are approximately 413 words within the Kucera and Francis (1967) norms that begin with BA.

Design

Experiment 1 used a mixed design with three independent variables, the first of which was item type. This variable was within-subjects and contained three levels, following from Anderson et al’s (1994) study on the RIFO effect. Rp+ items are those that belong to a category list that is included in the retrieval practice phase, and also are subjected to retrieval practice. Rp- items are those that belong to the same list as the retrieval practice items, but do not get practiced. Nrp items are control items that belong to the lists in which no retrieval practice occurs. The primary comparison of interest is the comparison between Rp- items and Nrp items, as this indicates whether memory for nontested information from practice lists is significantly above or below baseline (as specified by the control, Nrp condition). The order in which items belonged to each condition was counterbalanced.

The second independent variable was the format of the final test. This was a between-subjects variable with two levels: free and cued. Approximately half of the participants (n = 146) were randomly assigned to (a) a cued recall condition in which the category labels are shown as cues for the exemplars shown in the experiment, and the other half (n = 116) to (b) a free recall condition in which they were asked to simply recall as many items from the experiment as they could. Randomization was implemented by randomly assigning a participant to a version of the experiment when they showed up to the lab.

The third independent variable was also between-subjects, and it referred to the delay between retrieval practice and the final test phase. Approximately half of the participants in the cued recall condition and half of the participants in the free recall condition were randomly assigned to a short delay condition (n = 133), and the other half to a long delay condition (n =

129). Short delay was defined as a five minute retention interval, while long delay was defined as 20 minutes.

The dependent variable of interest was the proportion of information recalled. This was measured by counting the number of words participants were able to recall for each of the three levels of item type, and dividing them by the total number of items in each category. It was particularly important to use proportions, due to the fact that the Nrp category will have twice as many items as the Rp+ and Rp- categories.

Procedure

Participants begin the experiment by reading a set of instructions, indicating that they were first to study a series of word pairs. Each category-exemplar pair within a list was presented one at a time for 3 seconds each. The presentation order of items within a list was randomized. Participants studied all possible word pairs during the initial study phase.

During the subsequent retrieval practice phase, participants were asked to enter information using the keyboard to complete missing letters in the word pairs (e.g., if FRUIT – ST_____ was shown, they should have responded with STRAWBERRY). The items included in this phase were the Rp+ items. Each word pair was presented a total of three times in a completely randomized order. Each word pair was on the screen for 7s.

During the following distractor phase, participants were asked to complete a series of math problems. Depending on the delay condition, they either worked on math problems for 5 minutes or twenty minutes. Math problems were shown for one minute at a time, and consisted of a long string of operations.

After completing the distractor task, participants entered into the final test phase. In the cued recall final test condition, participants were shown category labels one at a time for 30 seconds each, and were asked to respond with as many words from that category as they could remember. Category presentation on the final test was completely randomized. In the free recall final test condition, participants had 9 minutes to try to remember as many of the exemplars from the 18 categories as possible, with their only instruction being to recall items that were shown during the experiment. The time was chosen to match the 30 seconds spent on each of the 18 lists during the cued recall final test phase.

Results

Testing Effect

To assess the presence or absence of the testing effect in each factorial condition, planned comparisons were performed on Rp+ and Nrp items. A robust testing advantage was observed across all combinations of delay and final test (*d*s = 1.44 to 3.03). For free recall at a short delay, *t*(49) = 17.52, *p* < .001, *d* = 3.03. For free recall at a long delay, *t*(65) = 14.32, *p* < .001, *d* = 1.97. For cued recall at a short delay, *t*(83) = 29.61, *p* < .001, *d* = 3.12. For cued recall at a long delay, *t*(61) = 12.14, *p* < .001, *d* = 1.44.

A three-way ANOVA was then conducted to look at the magnitude of the testing as a function of delay and final test type. Item type (Rp+ vs. Nrp) served as the within-subjects variable, with delay (5 min vs. 20 min) and final test type (cued recall vs. free recall) as between-subjects variables. There was a significant main effect of item type, with Rp+ items better remembered than Nrp items, *F*(1, 255) = 1297.5, *p* < .001. There was also a significant main of delay, with memory better after a 5 min delay than a 20 min delay, *F*(1, 255) = 10.3, *p* < .01, and a significant main effect of final test type, with performance better on the cued recall test than the free recall test, *F*(1, 255) = 102.8, *p* < .001. These main effects were qualified by significant two-way interactions, showing that the magnitude of the testing effect was moderated by final test type and item type. There was a significant interaction between test type and item type, *F*(1, 255) = 9.04, *p* < .01, such that the testing effect was of larger magnitude in free recall than cued recall (see Figure 1). There was also a significant interaction between delay and item type, *F*(1, 255) = 95.4, *p* < .001, showing a larger testing effect at the 5 min delay (Rp+=.69, Nrp=.22) than the 20 min delay (Rp+=.53, Nrp=.27). Of less importance, there was also a significant two-way interaction between test type and delay, *F*(1, 255) = 3.91. The three way interaction between these variables was nonsignificant, *F*(1, 255) = 0.98, *p* = .32. Refer to Table 1 for descriptive statistics.

RIFO/RIFA

Planned comparisons were performed on all Rp- and Nrp conditions, as those were the main comparisons of interest in this experiment. The analyses showed that free recall with a short delay was statistically significant *t*(49) = 4.44, *p* < .001. Free recall with a long delay was not statistically significant, *t*(65) = 1.02, *p* = .312, and cued recall with a short delay was marginally significant, *t*(83) = -1.92, *p* = .059. Cued recall with a long delay was also statistically significant, *t*(61) = -11.03, *p* < .001. Refer to Table 2 for the results of each planned comparison. Cohen's d was computed to determine the effect sizes of each of the four conditions; free recall with a short delay (*d* = .66), free recall with a long delay (*d* = .09), cued recall with a short delay (*d* = -.20), and cued recall with a long delay (*d* = -1.04). The positive effect sizes in the free recall conditions illustrate a facilitative effect, while the negative effect sizes in the cued recall conditions illustrate a forgetting effect.

A three-way ANOVA was run to assess the influence of final test type, delay, and item type on memory performance. Final test type has appeared as a potential moderating variable and should be tested. A significant main effect of final test type was revealed, *F*(1, 255) = 86.9, *p* <

.001, as well as a significant main effect of item type, *F*(1, 255) = 6.5, *p* = .01. There was not a significant main effect of delay, *p* > .05. A significant interaction between final test type and delay emerged, *F*(1, 255) = 6.10, *p* = .014, as well as a significant interaction between final test type and item type, *F*(1, 255) = 78.15, *p* < .001, and a significant interaction between delay and item type, *F*(1, 255) = 37.12, *p* < .001. The three-way interaction between these variables was marginally significant, *p* = .07. Refer to Figure 1 for a graphical representation of the data.

CHAPTER 3: EXPERIMENT 2

Method

Participants

As was done for Experiment 1, sample size was determined using a power analysis with G*Power 3.1 (Faul et al, 2009), using an effect size of 0.4 and power of 0.95. Total sample size was estimated to be 140 participants. The current experiment ended up with a sample size of 157. As in Experiment 1, the subjects were undergraduate Psychology students participating for partial fulfillment of course credit in lower division courses.

Materials

Stimuli consisted of cue – target word pairs that had no measurable association between the cue and target. There were 20 cue words, with each cue word paired with 4 targets, yielding 80 different cue – target pairs. The words selected as cues and targets were generated using the English Lexicon Project database (Balota et al, 2007). Only non-abstract nouns were chosen, with a length of four to seven letters. Words were constrained to average frequency of occurring in the English language, and had average orthographic and phonological neighborhood metrics. In constructing the cue – target pairs, no two targets sharing the same cue began with the same first two letters (e.g., if TABLE – KEY is shown as a word pair, TABLE – KEG could not be used). Nelson, McEloy, and Schreiber’s (1998) word association norms were also used to ensure that cue – target pairs did not have any measurable association with one another. The 80 resulting cue – target pairs were used to construct 10 lists of 8 items each. The first and last list served as primacy and recency fillers, and the other eight lists served as experimental lists. Each list consisted of two cue words and the four targets associated with each cue.

Design

This experiment utilized a 2 x 3 mixed design, in which final test type served as the between-subjects variable with two levels (i.e., free and cued recall), and item type served as the within-subjects variable with three levels (i.e., Rp+, Rp-, Nrp). The dependent variable was the proportion of words recalled.

Procedure

The procedure followed Experiment 1 very closely, with the major difference being the use of unrelated cue-target pairs instead of category-exemplar pairs, displayed in the form TABLE - KEY. To account for the increase in task difficulty that is expected when using these unrelated word pairs, the number of targets was reduced from six to four, and the full set of lists were shown three times during the study phase, with each pair appearing individually for three seconds each. The retrieval practice phase operated similarly to Experiment 1, in that the Rp+ items were shown three times each for seven seconds in the format of TABLE – KE_____. Delay was not manipulated in this experiment; rather, all participants got a five minute distractor task between the retrieval practice and final test phase. This was due to the fact that the primary measure of interest was final test type, and not delay (i.e. retention interval).

For the cued recall final test condition, participants saw each cue word for 20 seconds and were asked to recall all words they could remember that were paired with that cue. Cues were given in random order. For the free recall final test condition, participants were given 400 seconds (i.e. just under seven minutes) to recall as many targets as they can, with the instructions clarifying that they should recall as many target words as they could from the entire experiment. In all other respects, the procedure was identical to that of Experiment 1.

Results

Testing Effect

As was done in Experiment 1, planned comparisons were first run to determine if a testing effect emerged among the data. The comparison of interest for this result was between the Rp+ and Nrp items. Robust testing effects were found in both conditions. For the free recall final test condition, $t(78) = 8.85$, $p < .001$, $d = .84$. For the cued recall final test condition, $t(77) = 12.10$, $p < .001$, $d = .78$. Both conditions demonstrate a large testing effect. Due to the fact that the task in Experiment 2 was made more difficult due to the lack of semantic association between stimuli, it is unsurprising that performance is lower overall in this case, when comparing to Experiment 1.

RIFO/RIFA

Refer to Table 3 for descriptive statistics. Once again, planned comparisons were performed on the Rp- and Nrp items to see if there was either a forgetting or facilitative effect of retrieval practice. For the free recall condition, a significant forgetting effect was found, $t(78) = -2.51$, $p = .014$, $d = -.17$. For the cued recall condition, a significant facilitative effect was found, $t(77) = 2.46$, $p = .016$, $d = .15$. Therefore, the pattern that emerged from Experiment 2 was shown to be the reverse of what was found in Experiment 1, and the opposite of what was originally hypothesized according to research by Chan (2009) and Rowland and DeLosh (2014). Refer to Figure 2 for a graphical representation of these data.

CHAPTER 4: GENERAL DISCUSSION

For Experiment 1, the hypothesis that there would be an effect of final test type was confirmed. More accurately, it was confirmed that a free recall final test elicited a RIFA effect while a cued recall final test elicited a RIFO effect, regardless of retention interval (i.e. delay). Robust testing effects emerged regardless of condition in Experiment 1, but the presence of either RIFO or RIFA did seem to depend on final test type. However, this should be interpreted with caution considering the fact that a difference in baseline was observed in Experiment 1 across conditions. This led to an investigation of whether the pattern would persist when items were unrelated, as was done in Experiment 2. However, in Experiment 2 when item relatedness was manipulated so that word pairs no longer shared a semantic association, the pattern became less clear. In this instance, a RIFO effect emerged when using a free recall final test, and neither a RIFA effect emerged in the cued recall condition. This leads to a puzzling view of these indirect effects of retrieval practice. The crossover interaction from Experiment 1 did seem to highlight that Chan (2009) left out an important variable when stating which conditions must be present to elicit either RIFO or RIFA – final test type.

According to Chan in 2009, the most important elements for eliciting either RIFO or RIFA are integration of materials and delay. Chan demonstrated that high integration during retrieval practice led to a facilitative effect, while low integration led to a forgetting effect. In addition to this, a long delay (i.e. 24 hours) was shown to elicit RIFA while a short delay (i.e. 20 minutes) produced RIFO. Rowland and DeLosh (2014) altered the notion that these two aspects were of optimal importance when they demonstrated that RIFA could be elicited after a very short delay (i.e. five minutes) with items that were unrelated, and therefore lacked innate integration. There are indications that delay may interact with other variables (e.g., item type), but as was demonstrated by the lack of a significant main effect of delay, this factor alone does not seem to be enough to tease apart these effects. Experiment 2 of the present study could be evidence for integration as an important moderating variable. This experiment manipulated integration differently than Chan (2009), but both studies share a common understanding that integration may be important. This parallel exists, while the findings are not identical. This may suggest that across experiments the variable of integration could matter, though perhaps more as an interacting variable with final test type.

The shared theme among these findings and the current paper is that we have established a few phenomena that seem to affect the presence of RIFO or RIFA, but it is still unclear as to what the optimal conditions must be to ensure we see one effect emerge over the other. The current project was able to demonstrate in Experiment 1 that it is possible to control whether RIFA or RIFO emerges when using categorized lists. Again, it is important to be cautious with the interpretation since there was also a change in baseline in Experiment 1. However, the inclusion of unrelated words in Experiment 2 demonstrated that the absence of semantic relatedness altered whether or not these same parameters were effective. Still, the results from Experiment 2 can be interpreted in terms of Anderson (2003), which may allow for a slightly more optimistic take on these findings.

Concerning Experiment 1, it is interesting to see how the change in final test type can have such an impact on the pattern of results when looking at proportion of items recalled. One theory that has emerged that could potentially explain this is that of output interference (Jonker, Seli, & MacLeod, 2013). This theory states that the recall of Rp+ items could potentially be harming the recall of Rp- items, in that the output of the former items interferes with the output of the latter. Nrp items would not experience this deficit as the targets associated with each cue would be equally prioritized. In an experiment with a cued recall final test format, it is possible that the effect of output interference are made stronger by the fact that participants are cued with a category label, which could potentially restrict their ability to access specific targets. Considering the current study’s results, this could be the case when items are semantically related. However, Experiment 2 demonstrated a RIFO effect for free recall using stimuli that were episodically related rather than semantically related. According to Rowland and DeLosh (2014), it should not be the case that semantic relatedness is required to elicit a RIFA effect, as they were able to do this using unrelated word lists that were only episodically related. However, findings from Anderson et al. (1994) demonstrated that a RIFO effect can vanish when the relatedness between items is weak. This could be a contributing factor to the findings from Experiment 2. The output interference theory is more related to the findings from Experiment 1, but the message is mixed when we try to look at item relatedness as a factor involved in these effects. Relatedness does seem to be important when considering whether a forgetting or facilitative effect emerges, but there has yet to be a consistent set of circumstances to allow us to draw more specific conclusions as to how it plays a role here. However, both Anderson et al. (1994) and Rowland and DeLosh (2014) seem to demonstrate that when items are weakly related we may no longer see a forgetting effect. Chan (2009) is the one who seems to differ on this point, due to his emphasis on integration of materials as a contributing factor to RIFO or RIFA. More work needs to be done to better understand how item relatedness plays a role.

Another key distinction between the results of the present study and those from Rowland and DeLosh (2014) has to do with the blocking of the lists. Rowland and DeLosh (2014) had participants go through a study phase, followed by a brief distractor and retrieval practice phase before going on to another list. In the current study, as well as others in the RIFO literature (Anderson et al., 1994; Rowland et al., 2014) the lists were blocked so that all items were presented during one large study phase, followed by one retrieval practice phase that included all of the Rp+ items. It is possible that blocking the lists may be one way that a forgetting effect may emerge. There is some evidence to support that interleaving tests among study sessions can produce an enhanced benefit over and beyond simply having one test at the end of the study episode (Szpunar, McDermott, & Roediger, 2008). Future research will seek to investigate this concept.

A better understanding of RIFO and RIFA could lead to better understanding of the underlying mechanisms of the testing effect. Currently, one of the main theories behind why it is responsible for increased retention has to do with the spreading of activation from the tested item to other related items. The elaborative retrieval hypothesis states that the event of a test may activate elaborative information related to the target response (Carpenter, 2009). What this means is that engaging in retrieval may activate information that is related to the tested information. The fact that this additional information is activated may be part of what is giving tested information such a significant memory boost. When we consider the effects of RIFO and RIFA and how they impact information that is related to tested information, yet not tested itself, it leads us to consider how the act of testing is impacted by the spreading of activation. The elaborative rehearsal hypothesis frames this as a positive effect of retrieval, as it potentially enhances retention of tested information. However we have evidence of instances in which RIFO emerges (Anderson et al., 2014; Rowland et al., 2014), therefore there are certain cases in which the spreading of activation may not be enough to elicit a benefit for nontested information.

Aside from the important theoretical implications associated with the current study, there are also considerable practical implications as well. From an educator's perspective, structuring a course so that students' learning is maximized is of the highest importance. Knowing the circumstances that elicit either a RIFO or RIFA effect can aid in this endeavor. If we know the parameters under which RIFO is observed, then we can take measures to ensure that we do not implement these parameters in a classroom setting. Along those same lines, knowing the circumstances that elicit RIFA could benefit class structure to ensure that we always make an effort to provide conditions ripe for a facilitative effect.

We have learned from Experiment 1 in the present study that a free recall final test can elicit RIFA when using semantically related items. Since most lecture content is semantically related, this could point us towards a course of action that would promote learning. We know from the testing effect literature that free recall tests (e.g., essay exams) are the most desirable test format in that we see the greatest benefit from testing (Roediger & Karpicke, 2006b). Therefore, the knowledge that RIFA can emerge under these conditions is an additional piece of support for this style of assessment. While it may be more work on the instructor's part, we have data to suggest that this benefits students' learning beyond other types of testing such as recognition or short answer questions.

Due to the interesting pattern of data that emerged from Experiment 1, a natural future direction would be to investigate the role of delay. While the pattern was unclear as to what role delay played, that does not mean it is worth looking into. It is also important to consider that many studies in the RIFO and RIFA literature have used a variety of delays with different effects (Anderson et al., 1994; Chan, 2009; Rowland et al., 2014; Rowland & DeLosh, 2014). As Chan (2009) stated, delay could be a large player when it comes to eliciting either RIFO or RIFA.

Potentially look into longer delays (e.g. 24 hours). Then we can make stronger claims about the potency of these effects, as it is possible that comparing a delay of five minutes and 20 minutes is not sufficient to make strong claims about the role of delay. Another potential factor to look into in future research is the presence of intrusions. It may be worthwhile to look at the data from Experiment 1 and investigate whether participants were simply guessing items that were not part of the experiment (e.g., FRUIT – APPLE) since stimuli were categorically related. It could be the case that participants are worse at discriminating present versus absent items at the long delay (i.e., 20 minutes) due to the fact that their memories have faded some. Future research will seek to investigate whether there is a higher intrusion rate at the 20 minutes compared to five minutes. This would allow us to better assess whether participants are accessing the category.

As it stands, we seem to have some evidence that final test type plays a role in the occurrence of the indirect effects of testing, though this seems to be heavily related to stimulus type, and also the potential blocking of test episodes. While the picture remains unclear, we do seem to have isolated specific situations in which we can effectively control for the indirect effects of testing.

CHAPTER 5: TABLES AND FIGURES

Table 1

Experiment 1 Descriptive Statistics by Item Type

	Free Recall		Cued Recall	
	Short Delay	Long Delay	Short Delay	Long Delay
	M (SD)	M (SD)	M (SD)	M (SD)
Rp+	.60 (.21)	.49 (.19)	.78 (.16)	.58 (.18)
Rp-	.18 (.12)	.19 (.11)	.30 (.17)	.21 (.13)
Nrp	.11 (.09)	.18 (.12)	.33 (.12)	.35 (.14)

Table 2

Experiment 1 Planned Comparisons for Rp- and Nrp Items

					95% Confidence Interval of the Difference	
Test Type	Delay	t	df	p	Lower	Upper
Free Recall	Short Delay	4.44	49	0.000	0.04	0.11
	Long Delay	1.02	65	0.312	-0.01	0.04
Cued Recall	Short Delay	-1.92	83	0.059	-0.05	0.00
	Long Delay	-11.03	61	0.000	-0.16	-0.11

Table 3

Experiment 2 Descriptive Statistics by Item Type

	Free Recall M (SD)	Cued Recall M (SD)
Rp+	.28 (.19)	.34 (.22)
Rp-	.12 (.12)	.21 (.21)
Nrp	.14 (.12)	.18 (.18)

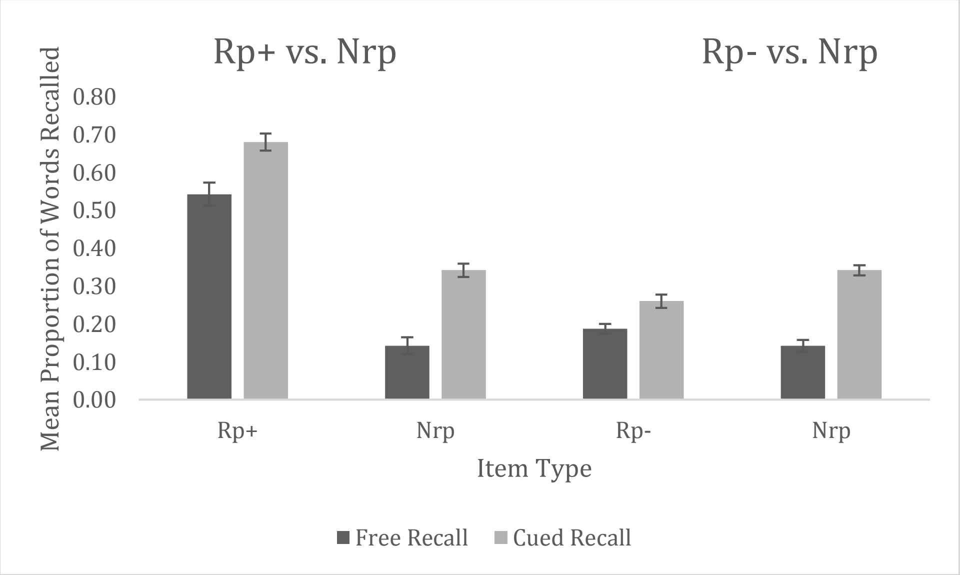

Figure 1: Data presented from Experiment 1. Mean proportion of words recalled as a function of final test type. Error bars represent standard error of the mean.

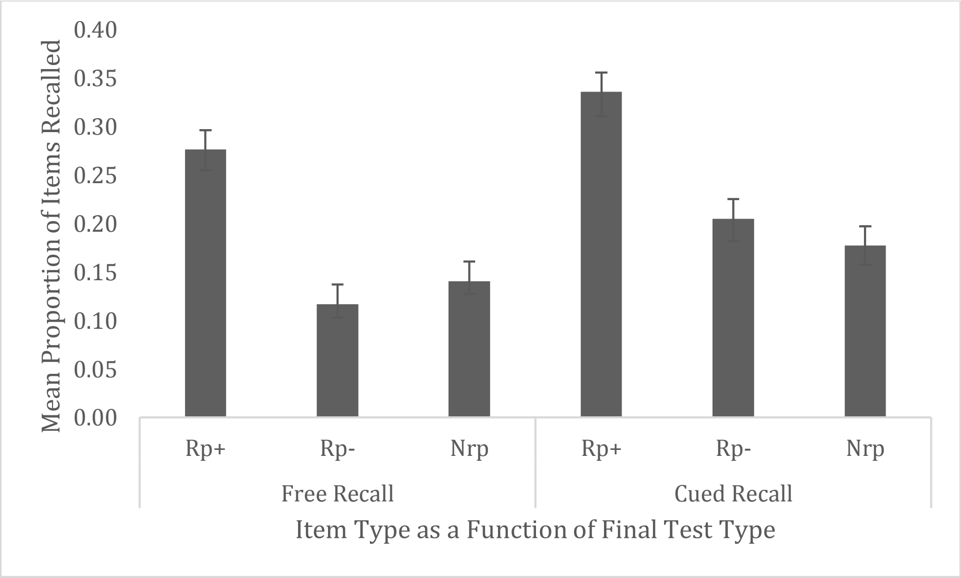

Figure 2: Data presented from Experiment 2. Mean proportion of items recalled as a function of both item type and final test format. Error bars represent standard error of the mean.

REFERENCES

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49,* 415-445.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063-1087.

Anderson, M.C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review*, *102*(1), 68-100.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445-459.

Battig, W. F. & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology, 80,* 1-46.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569. doi:10.1037/a0017021

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*, 153-170.

Chan, J. C. K., McDermott, K. B. & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553-571.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Jonker, T., Seli, P., & MacLeod, C. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review, 120*(4), 852-872. doi:10.1037/a0034246

Kucera, H. & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Marshall, G. R. & Cofer, C. N. (1970). Single-word free association norms for 328 responses from the Connecticut cultural norms for verbal items in categories. In L. Postman & G. Keppel (Eds.), *Norms of word association* (pp. 321-360). New York: Academic Press.

McDaniel, M. A. & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*(2), 192-201. doi:10.1016/0361-476X(91)90037-L

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/Raven, J. C. (1995). *Advanced progressive matrices set II*. Oxford: Oxford Psychologists Press.

Roediger, H. L. & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255.

Roediger, H. L. & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181-210.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1-36. doi: 10.1016/B978-0-12-387691-1.00001-6

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432-1463. doi:10.1037/a0037559

Rowland, C. A., Bates, L. E., & DeLosh, E. L. (2014). On the reliability of retrieval-induced forgetting. *Frontiers in Psychology – Cognition.* doi: 10.3389/fpsyg.2014.01343

Rowland, C. A. & DeLosh, E. L. (2014). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review, 5*. doi: 10.3758/s13423-014-0625-2

Solso, R. L. & Juel, C. L. (1980). Positional frequency and versatility of bigrams for two-through nine-letter English words. *Behavior Research Methods and Instrumentation, 12,* 297-343.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392-1399. doi:10.1037/a0013082