THE OPPOSITE EFFECTS OF SEMANTIC RETRIEVAL ON PRIOR AND FUTURE LEARNING

BY

KRISTIN M. DIVIS

THESIS
Submitted in partial fulfillment of the requirements for the degree of Master of Arts in Psychology in the Graduate College of the University of Illinois at Urbana-Champaign, 2012

Champaign, Illinois

Master’s Committee:

Professor Aaron S. Benjamin
Professor Brian H. Ross
ABSTRACT

Recently, interest in the effects of testing on memory has increased. In this set of experiments, I examined the effects of interleaved semantic retrieval on previous and future learning within a multi-list learning paradigm. Interleaved retrieval led to enhanced memory for lists learned following retrieval. In contrast, memory was impaired for lists learned prior to retrieval (Experiment 1). These results are consistent with recent work in multi-list learning and in the list-before-the-last paradigm, both of which reveal a crucial role for retrieval in enhancing list segregation. This pattern of results also follows clearly from a theoretical perspective in which retrieval drives internal contextual change, and contextual overlap between study and test promotes better memory. Consistent with that perspective, a 15-minute delay before the final test eliminated both effects (Experiment 2). Experiment 3 replicated the results of Experiment 1 with materials more appropriate for educational settings: interleaved semantic retrieval led learners to be more able to answer questions correctly about later texts but less likely to do so about earlier texts.
ACKNOWLEDGEMENTS

I would like to thank Professor Aaron S. Benjamin for his role as my adviser, and for his committed guidance and assistance during the research and preparation of my thesis. I also wish to thank Professor Brian H. Ross for serving on my committee and providing helpful comments and advice on this project. I would also like to thank Alison Trembacki for her assistance collecting data for this study and all the members of the Human Memory and Cognition Lab for their helpful comments and suggestions.
# TABLE OF CONTENTS

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

EXPERIMENT 1..........................................................................................................................7

EXPERIMENT 2..........................................................................................................................11

EXPERIMENT 3..........................................................................................................................13

GENERAL DISCUSSION............................................................................................................17

REFERENCES.............................................................................................................................21

APPENDIX A.............................................................................................................................24

APPENDIX B.............................................................................................................................25

FIGURES......................................................................................................................................29
Introduction

Testing has many beneficial effects on memory. It is well established that testing previously learned material enhances long-term memory for that tested material (e.g., the testing effect; for a review see Roediger & Karpicke, 2006). Retrieval events can also influence learning by affecting proactive interference (PI; Szpunar, McDermott, & Roediger, 2008; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011) and retroactive interference (RI; Jang & Huber, 2008).

The goals of the present set of experiments are to evaluate the effects of retrieval on both future learning and prior learning within a single experimental paradigm and to more precisely determine the origin of the costs and benefits of retrieval. The working hypothesis presented here, developed from prior work by Sahakyan and Kelley (2002) and Jang and Huber (2008), is that retrieval events lead to greater internal context change and thus greater contextual segregation between events prior to and after the retrieval event. This segregation has the potential to improve retention (by decreasing the interference between competing events) and to impair retention (by creating a greater disparity between contexts present at encoding and during the eventual criterion test). In Experiment 1, I examine the effects of interleaved semantic retrieval on both early and later learning when a final test immediately follows the study session. Experiment 2 extends the findings of Experiment 1 with a delayed final test. Experiment 3 replicates the results of Experiment 1 with more educationally relevant text materials.

The effects of testing on future learning

Though testing is mostly known for its large effects of enhancing memory for the tested material itself, testing also has an influence on later learning. In one of the first studies examining how testing influences future learning, Tulving and Watkins (1974) explored the
consequences of retrieval within an AB-AC interference paradigm. Having an immediate test following study of the AB list led to superior memory for the yet-to-be-learned AC items than when no test was given after the AB list. They posited that this effect came about because second list learning was impaired when the first list was not tested before studying the second list. Other research supported this claim by demonstrating that words from untested lists were also more likely to intrude into recall on later tests than words from tested lists (Darley & Murdock, 1971). These results indicate that testing may reduce the buildup of PI, a claim supported in later work by Szpunar et al. (2008): when subjects were given either extended study sessions or interpolated tests while studying lists of words, those in the interpolated test condition showed a marked reduction in PI across lists. Furthermore, the results of additional experiments indicated that PI builds over time, becoming greater as more studied lists remained untested before recall of the final list.

Interestingly, the reduction in PI that is apparent following retrieval is not limited to conditions in which the retrieved material is from the preceding (and thus potentially interfering) list. Pastötter et al. (2011) examined the effects of alternate forms of retrieval. Subjects completed one of five tasks between studying five different lists of words: a distractor task (counting backward by 3s), restudying the immediately preceding list, free recall of the immediately preceding list, a 2-back short-term memory task, or a semantic retrieval task (generating examples from a category like “sports”). Immediate recall and final recall of the last list were enhanced for those in the three retrieval conditions (free recall, n-back, and semantic retrieval) compared to those in the distractor and restudy conditions, suggesting that the process of retrieval itself (and not just retrieval of potentially interfering items) enhances future learning. Such a result rules out a previously plausible hypothesis that the reduction in PI is driven by
better source memory (due to testing serving to enhance memory for the tested items). The fact that the benefit occurs (and is of approximately equal magnitude) when the interleaved retrieval does not actually test memory for (and therefore enhance) any of the previously studied material suggests a different basis for the reduction in PI.

The effects of retrieval on past learning

The effects of retrieval on future learning appear to derive from a reduction in PI following that retrieval event. The majority of studies addressing the effects of retrieval on prior learning are studies of the “testing effect,” in which the retrieved material is the prior list itself. In that case, it is quite clear that retrieval provides a substantial and long-lasting benefit to memory for that prior list (for a review of the testing effect, see Roediger & Karpicke, 2006). However, there are also hints in the less well known list-before-the-last paradigm (Shiffrin, 1970) that the act of retrieval fundamentally affects even previously learned material that is not retrieved.

Shiffrin (1970) asked subjects to recall words from the list immediately prior to the most recently studied list (e.g., if they studied List 1 followed by List 2, they were asked to recall words from List 1). Although the length of the to-be-recalled list influenced memory (the list length effect; Murdock, 1960; Roberts, 1972), the length of the intervening list did not. This result suggests that subjects were able to effectively exclude the most recent event from consideration and thereby avoid RI (which would be expected to increase with list length). However, later work showed that this effect only obtains when a retrieval event occurs between the two lists—in the absence of retrieval, shorter intervening lists lead to better performance on the tested list than longer intervening lists (Jang & Huber, 2008). Like the results considered in
the previous section, this finding suggests that retrieval decreases the degree to which the lists compete with one another at retrieval. Jang and Huber hypothesized that recall created a significant context change between the lists, and that this change caused the intervening list to interfere less with retrieval of the target list. Notably, overall performance for the prior list was reduced in the retrieval condition compared to the control. Further research utilizing the list-before-the-last paradigm also found that intervening retrieval events (as opposed to math problems) led to reduced intrusions but also overall reduced performance compared to a control condition (Sahakyan & Hendricks, 2012). This finding suggests that, although retrieval may reduce interference across lists, it also decreases access to material that was learned prior to the retrieval event. Taken together, the results here suggest that the consequences of retrieval are twofold: though retrieval often benefits memory by reducing interlist interference, it also reduces access to material learned earlier. The context-change hypothesis of the effects of retrieval provides a way of understanding both of these phenomena.

Context-change hypothesis

The context-change hypothesis posits that during encoding, fluctuating contextual cues are bound to an internal context (Estes, 1955; Mensink & Raaijmakers, 1988). Context is important because performance is enhanced when the context at time of retrieval resembles the study context. Here I assume that a retrieval event serves to enhance context fluctuation, thus binding subsequently encoded information to a more different internal context than would otherwise be the case. Similar hypotheses have been proposed to explain the underlying effects of retrieval in multi-list learning (e.g., Jang & Huber, 2008; Pastotter et al., 2011; Sahakyan &
Hendricks, 2012), as well as the *directed forgetting* effect. I briefly review those applications here.

In a typical *list-method directed forgetting* study, subjects study List 1 and are instructed to either remember or forget that list (usually between-subjects). They then study List 2 and are eventually given recall tests for one or both lists. Subjects in the “forget” condition recall fewer items from List 1 than subjects in the “remember” condition but more items from List 2 (Reitman, Malin, Bjork, & Higman, 1973), suggesting that intentional forgetting of List 1 reduces PI (Bjork & Bjork, 1996).

However, effects similar to those of directed forgetting can be achieved by means other than intentionally trying to forget. Manipulations that mimic directed forgetting effects include: imagining being invisible (Sahakyan & Kelley, 2002), imagining walking through one’s childhood home (Sahakyan & Kelley, 2002), or imagining a vacation (Delaney, Sahakyan, Kelley, & Zimmerman, 2010). In fact, effects of greater magnitude were seen when the imagined event was further away, in either time or space (Delaney et al., 2010). Tasks that do not mimic directed forgetting effects include solving math problems (Sahakyan & Kelley, 2002), number searches (Mulji & Bodner, 2010), speeded reading (Delaney et al., 2010), and counting tasks (Sahakyan, Delaney, & Goodmon, 2008; Pastötter & Bäuml, 2007; Pastötter et al., 2011). Most of the tasks that mimicked directed forgetting effects involved mental context change and retrieval of some sort (and imagining events may involve many of the same processes as retrieving them [e.g., Schacter, Addis, & Buckner, 2008]).

In the current experiments, I take the context change hypothesis as my basis for understanding the varied effects of retrieval on multi-list learning, specifically addressing both prior and future learning, as well as the effect of test delay. Having different contextual cues tied
to different lists promotes list isolation and reduces interference. Greater context change should lead to (a) enhanced performance on later lists because of a reduction in PI (Szpunar et al., 2008; Pastötter et al., 2011) and (b) overall reduced performance on earlier lists. This latter effect occurs because although RI is reduced, the context fluctuation renders the earlier list context more dissimilar than the test context (Jang & Huber, 2008; Sahakyan & Hendricks, 2012). In addition, a delay prior to test should reduce both of these effects because of the large difference in context between test and original study contexts.

Current Experiments

The goal of the present study is to further understand the effects of retrieval on multi-list learning. In order to achieve an uncontaminated look at both initial-list and last-list effects, I used a semantic retrieval task similar to that used by Pastötter et al. (2011) rather than the more commonly used episodic retrieval task (recalling either the immediately preceding list or the list-before-the-last; e.g., Szpunar et al., 2008; Shiffrin, 1970; Jang & Huber, 2008). Experiment 1 examined the effects of interleaved retrieval on initial-list and last-list performance with an immediate free recall test, while Experiment 2 introduced a 15-minute delay before the final test. Experiment 3 extended the results of Experiment 1 to more educationally relevant text materials.
Experiment 1

Methods

Subjects. Eighty-six undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment for course credit. Data from 7 subjects were dropped because they did not follow the instructions (they recalled words for the previous list instead of generating exemplars from the given category).

Design. Type of intervening task (unrelated semantic retrieval or control) and testing order of the studied lists were both manipulated between subjects. Memory performance was measured as the proportion of words correctly free recalled.

Materials. Fifty words (average word length = 5.14 letters, s.d. = 1.46) were drawn from the University of South Florida Free Association Norms (Nelson, McElvoy, & Schreiber, 1998). A random subset of ten words was used for each of five lists. Only words unrelated to each other were used. No words were repeated in the experiment.

Procedure. Subjects worked individually in a small room. PC-style computers programmed using Matlab with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997) were used to present stimuli and record responses. Prior to the initial study phase, subjects were presented with a set of instructions informing them they would be studying lists of words for a later free recall test (an example was also given). All words were presented for 4 seconds each with an interstimulus interval of 500 ms. The procedure was drawn from Pastötter et al. (2011). Subjects studied 5 lists of words, each separated by an intervening task. The distractor task consisted of counting backward by 3s from a randomly generated 3-digit number for 30 seconds. Subjects in the distractor condition were given 3 sets of these (for a total of 90 seconds spent on the task). Subjects in the retrieval condition first had one 30-second set of counting
backward by 3s (like the distractor condition) and then spent 60 seconds on a semantic retrieval task. They were given 60 seconds to type in as many exemplars from a given category (4-legged animals, sports, vegetables, or professions, randomly ordered across the intervening tasks). Instructions were given before each task. After studying the final list (List 5), all subjects were given 2 more sets of counting backward by 3s (for a total of 60 seconds). Half of the subjects were then given a free recall test on List 1 and half were tested on List 5. For completeness, and to avoid the appearance of deception in the initial study instructions, this test was followed by the test for the other list. However, because my theoretical stance makes no clear predictions about how this first test should affect performance on the second, I restrict my analysis here to performance on the first test. All responses were typed into the computer and no response time limit was imposed.

Results and Discussion

Significance levels for all statistical tests were set at an $\alpha < .05$ level. In order to account for error terms that were not distributed normally and be able to include both subject and item variability, mixed effects models were used rather than the standard analysis of variance statistical tests. All models were fitted via Laplace estimation with the lme4 package (Bates, Maechler, & Dai, 2011) in R software (R Development Core Team, 2008). As expected, testing order significantly influenced the results, so only the results of the first test are reported here (see Appendix A for results and discussion of the second test).

The best-fit model included the fixed effects of list (1 or 5) and intervening task (retrieval or distractor) along with random intercepts for subjects. Adding items as a random effect did not reliably improve the fit of the model ($\chi^2 = .990, p = .320$), so it was not included in the final
model. Mean proportion correct are shown in Figure 1a. The model indicated no overall effect of task; however, there was a main effect of list ($z = 3.629, p < .001$), with higher overall performance in List 5 than List 1. Furthermore, a significant 2-way interaction ($z = 3.164, p = .0016$) revealed higher performance in List 1 for the distractor condition compared to the retrieval condition but higher performance in List 5 for the retrieval condition compared to the distractor condition. This pattern of results was supported by reliable simple effects of task in both List 1 ($z = 2.002, p = .0453$) and List 5 ($z = 2.500, p = .0124$).

Interleaved semantic retrieval events (compared to a nonmnemonic distractor task) led to overall reduced performance for List 1 recall but enhanced performance for List 5 recall in tests immediately following the study session. This pattern of results follows from the context-change hypothesis and is consistent with the idea that the interleaved retrieval events served to alter the context, leading each list to be contextually more distinct from the others.

The context-change hypothesis suggests that the advantage for List 5 should owe to reduced PI in that condition. Figure 2 shows the number of interlist intrusions during the final (free recall) test of Lists 1 and 5. While inferential analysis was underpowered due to the low occurrence of intrusions, a mixed effects model including the fixed factors of task and list along with random intercepts for subjects was fit to the data (including random intercepts for the list from which the intrusions came did not reliably improve the fit of the model [$\chi^2 = 1.554, p = .213$]). The model revealed a marginal main effect of list ($t = 1.951, p = .052$), with more intrusions in List 5; the main effect of task was not reliable ($t = 1.603, p = .1099$). The two-way interaction between task and list was also not reliable ($t = 1.404, p = .161$). However, analysis of the simple effects of task in each list revealed increased interlist intrusions on List 5 in the distractor condition compared to the retrieval condition ($t = 2.116, p = .035$) but no simple effect
of task for List 1 ($t=.141, p=.888$). While there were not enough intrusions to provide sufficient power for inferential analysis, the data reveal more interlist intrusions on the List 5 test in the distractor condition than in the retrieval condition. This result suggests that PI accrued in the distractor condition but did not in the retrieval condition. The effect of retrieval on List 5 performance was driven by the reduction in the buildup of PI (no retrieval event occurred between studying List 5 and final test, so the similarity of the test context and List 5 context remained consistent across both the retrieval and distractor conditions). The retrieval-driven reduction in PI from Lists 1-4 led to overall enhanced performance on List 5.

Notably, no effect of interlist intrusions is apparent in List 1, suggesting that retrieval did not significantly influence RI. According to the context-change hypothesis, reduced performance on List 1 in the retrieval condition was driven by the difference in contexts between List 1 and the final test. Because each interleaved retrieval event during the study session caused context to shift further than it normally would, the context at the time of the final test was further down the contextual stream (and thus more disparate than the List 1 context) in the retrieval condition than in the distractor condition. This greater mismatch of List 1 context and criterion test context led to overall reduced performance on List 1.
Experiment 2

Experiment 2 was designed to examine the effects of delaying the final test after the study session. If an extended retention interval functions similarly to a retrieval event, a delay between studying the lists and being tested on the lists should lead to either (a) the context at test being more fully segregated from any of the lists (including List 5) so that the benefits of reduced interference are seen but not the effects of the test context being most similar to List 5 context (predicting enhanced performance for *both* the earlier and later studied lists due to a reduction in RI and PI) or (b) the context at test being shifted dramatically away from the list contexts so that the differences between the list contexts are relatively small (in which case the benefits and costs of list isolation due to interleaved retrieval should be reduced or eliminated).

Methods

*Subjects.* One hundred seventeen undergraduate students at the University of Illinois at Urbana-Champaign participated for course credit. Data from 13 subjects were dropped for not properly following the instructions (e.g., recalling items from previous lists when they should be listing items from the given semantic category).

*Design, Materials, and Procedure.* Experiment 2 was identical to that of Experiment 1, with the following exceptions. Presentation time for each word was increased from 4 seconds to 6 seconds. A 15 minute delay was introduced between the study session (after the 60 seconds of counting following the final list) and the final tests. During the delay, subjects performed a spatial matching task.
Results and Discussion

The mixed effects models reported here were fitted via Laplace estimation in a similar fashion as in Experiment 1. Again, only the results of the first test are reported here (see Appendix A for results and discussion of the second test).

Once again, the best-fit model included the fixed effects of list (1 or 5) and intervening task (retrieval or distractor) along with random intercepts for subjects. Adding random intercepts for item did not reliably improve the fit of the model ($\chi^2 = 1.845, p = .174$). Mean proportion correct are shown in Figure 1b. While there was a marginal main effect of list with higher performance in List 1 than List 5 ($z = 1.786, p = .074$), there was neither a reliable main effect of task ($z = .242, p = .809$) nor a reliable interaction between list and task ($z = .129, p = .897$).

The results from Experiment 2 indicate important boundary conditions on the results evident in Experiment 1. There was no effect of interleaved retrieval on List 1 and List 5 performance. According to the context-change hypothesis, the increased retention interval between the study session and the final test caused the context at test to shift sufficiently far away from the list contexts that the increased differentiation between list contexts (via interleaved retrieval) was now quite small relative to the difference between those contexts and the criterion test context. Thus, list isolation no longer benefited performance. In addition, List 5 no longer received the benefit of sharing a similar context with the final test, and List 1 performance in the retrieval condition (compared to the distractor condition) was not reduced because the difference between List 1 and final test contexts was not dramatically different across the retrieval and distractor conditions (relative to the new, very different test context). Taken together, these results indicate that the effects of retrieval depend on the relative placement within the contextual stream of the criterion test context.
Experiment 3

The benefits of reducing PI by using a retrieval task hold some promise of aiding human learning in more ecological valid situations than simple list learning. Experiment 3 extends the methods used in Experiment 1 from word lists to text materials. The most important change that follows from this shift in materials concerns the nature of the test. Whereas the tests of recall implemented in the first two experiments indicate the effects of retrieval on episodic memory, the test in Experiment 3 queries a more general understanding of the complex materials.

Methods

Subjects. Twenty-eight undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment for course credit.

Design. Type of intervening task (unrelated semantic retrieval or nonmnemonic distractor control) and testing order of the studied texts were both manipulated between subjects. The tested texts (1 and 4) were counterbalanced between subjects and the filler texts (2 and 3) were counterbalanced between subjects independently. Memory performance was measured via short answer and multiple choice questions.

Materials. Text materials were drawn from a standardized test (ACT) prep book (Dulan, 2010). The texts were related to animals (Coyotes, Porcupines, Seals, and Chronic Wasting Disease [CWD]) and averaged 608 words (s.d. = 55.62). The coyote and porcupine texts were in tested positions (Texts 1 or 4), while the seal and CWD texts were in filler positions (Texts 2 or 3).

Procedure. The procedure was similar to that of Experiment 1. The texts were presented one paragraph at a time. Reading time was self-paced and subjects could not go back to a
previous paragraph after advancing to the next one. Subjects studied a total of four texts. Each text was separated by an intervening task. The tasks were virtually identical to that in Experiments 1 and 2. Half of the subjects were in the distractor condition (3 sets of counting backward by 3s from a 3-digit number for 30 seconds each) and the other half were in the retrieval condition (counting backward for 30 seconds followed by 60 seconds of unrelated semantic retrieval). The semantic retrieval categories (listing types of sports, professions, or office supplies) were randomly ordered for each subject. After studying Text 4, subjects completed two more 30-second sets of the distractor counting task. The test was administered via pen and paper. For both Text 1 and Text 4, subjects completed short answer questions followed by ten multiple choice questions (see Appendix B). As in the previous experiments, subjects were tested on the alternative text following the first test. Subjects were instructed not to go back and change answers after moving to the next question, and no time limit was given to complete the test.

Results and Discussion

Once again the mixed effects models reported here were fitted via Laplace estimation in a manner similar to Experiments 1 and 2. Only the results from the first test are reported here (see Appendix A for results and discussion of the second test). The analyses were separated based on question type (multiple choice or short answer). Mean proportion correct are reported in Figure 3a (multiple choice) and Figure 3b (short answer).

Multiple Choice. The best-fit model for the multiple choice data included the fixed effects of text (1 or 4) and intervening task (retrieval or distractor). Including random intercepts for question in addition to random intercepts for subject significantly improved the fit of the model
($\chi^2 = 5.056, p = .0246$). While there was no reliable main effect of task ($z = .063, p = .971$), a main effect of text occurred where overall performance on Text 1 was higher than that of Text 4 ($z = 2.507, p = .0122$). Furthermore, a reliable 2-way interaction revealed enhanced performance for Text 4 but reduced performance for Text 1 in the retrieval task compared to the distractor task ($z = 2.894, p = .00381$). This pattern of results was supported by a marginal simple effect of task in Text 1 ($z = 1.893, p = .0584$) and a reliable simple effect of task in Text 4 ($z = 2.236, p = .0254$). Notably, the data followed similar trends, regardless of text type (e.g., coyote versus porcupine). Including the random effect of question also allowed the model to take into account variance introduced by the specific text.

*Short Answer.* Two independent raters scored the short answer questions (intraclass correlation, $r = .789$). The best-fit model included the fixed effects of intervening task (retrieval or distractor) and text (1 or 4). It also included random intercepts for subject, question, and rater, along with random slopes for question (the fit of the model was significantly improved by including random slopes instead of only random intercepts for question; $\chi^2 = 21.961, p = .009$). Because the usual Markov Chain Monte Carlo (MCMC) method of determining reliability has not been implemented in the lme4 package, I instead calculated significance using “worst case scenario” estimates of the degrees of freedom ($df = 200$) for the given $t$-values.

The results revealed no main effect of task ($t = .634, p = .527$) or text ($t = 1.337, p = .183$). However, a reliable interaction ($t = 2.035, p = .043$) indicates that subjects in the retrieval task had lower performance on Text 1 but higher performance on Text 4 than those in the distractor task. Analyses of the simple effects revealed that this trend is marginally reliable in
Text 4 ($t = 1.881, p = .0614$) but not Text 1 ($t = .982, p = .327$).\(^1\) While the increased noise in the short answer analysis led to overall weaker results, it is important to note that these data numerically replicate the multiple choice question data.

Experiment 3 replicated the results of Experiment 1, extending the effect to materials more relevant for educational settings. Having an unrelated retrieval event between studying texts led to reduced performance on earlier material but enhanced performance on later material. Once again these results suggest that the interleaved retrieval events have the beneficial effect of segregating texts (and thus reducing intertext interference) but the harmful effect of reduced access to earlier material at time of the critical test of that material (due to a greater shift down the contextual stream).

---

\(^1\) Analyses using transformed (empirical logit) data revealed similar critical effects. The two-way interaction between text and task was reliable ($t = 2.304, p = .0222$). The simple effect of task in Text 4 was reliable ($t = 2.156, p = .0323$) but not in Text 1 ($t = .859, p = .391$).
General Discussion

The current study utilized a multi-list paradigm to explore the beneficial and harmful effects of unrelated semantic retrieval on memory for both earlier and later items. Interleaved retrieval events led to overall reduced performance on earlier items but enhanced performance on later items when compared to a nonmnemonic control condition. This effect held for both word lists and text materials when the criterion test immediately followed the study session; however, the effect disappeared when the criterion test was delayed.

These results follow directly from the context-change hypothesis. The more similar the study context and the test context, the more likely one will be to remember items from the study session. Similarly, the more similar two study contexts are, the more likely they will be to interfere with each other. The interleaved retrieval events served to alter subjects’ internal context, causing it to shift more rapidly than it normally would. This led to the potentially beneficial effect of each list context in the study session being more differentiated from one another (and thus less likely to interfere with each other). However, the interleaved retrieval events (and subsequent context shifts) also led to the potentially harmful effect of the criterion test context being shifted much further away from the study contexts than would have normally occurred with no interleaved retrieval events.

The dual nature of retrieval events—beneficially segregating the study contexts but decreasing the match between study context and criterion test context—led to improved access on later learning but impaired access on earlier learning. Because the last study session (e.g., List 5) was not affected by an additional shift in context (via a retrieval event) between it and the criterion test, the driving factor for performance was the reduction in PI. Therefore, interleaved retrieval led to enhanced performance on later items (e.g., List 5). While interleaved retrieval
events also led to better list segregation (and potentially reduced interlist interference), the driving factor was the greater mismatch in context between the studied items (e.g., List 1) and the criterion test when compared to the nonmnemonic control condition. Furthermore, RI did not appear to play a significant role in either condition. With the context dissimilarity driving performance, interleaved retrieval led to overall reduced performance for earlier items (e.g., List 1).

Introducing a delay between the study session and criterion test allowed the criterion test context to shift dramatically further away from that of the study session. The context shifts (via interleaved retrieval) between the lists in the study session were then small relative to the new criterion context (which was much further down the contextual stream compared to the no-delay condition). Because the shifted context at the criterion test reduced the context effects of interleaved retrieval within the study session to relatively little, it follows that interleaved retrieval did not influence performance.

While interleaved retrieval led to strong effects when the criterion test immediately followed the study session, these effects disappeared once the criterion test was delayed. These results suggest that a contextual boundary exists for the influence of interleaved retrieval within a study session. When the context at criterion test is sufficiently different from that of the study contexts, the retrieval-driven differences in context become relatively small. The current study indicated that a significant shift in context at criterion test can occur in as little as fifteen minutes. Events (other than time delays) that significantly change context are also likely to render the effects of retrieval null. It is essential that the context at criterion test holds a certain degree of resemblance to that of the study session in order for interleaved retrieval to be an effective tool during study. Notably, once the study session criterion is lost, that does not mean it
cannot be regained. Results of the second test in the delay experiment (Experiment 2) suggest that reinstating the study context once again leads to different effects of task based on list (see Appendix A for results of the second test). Further exploration of the boundaries of the influence of retrieval is still needed. Specifically, it would be beneficial to further examine the effects of retrieval once the study session context has been lost and then reinstated at criterion test.

While further research is still needed to determine the boundaries of the effects of retrieval during study (both in terms of how quickly the effect fades away after the study session and how strongly it reemerges with the reinstatement of the study context), many potential applications exist. Most prominent is the application to educational settings. While introducing context change (via unrelated retrieval or another method) during study enhances later learning, it also impairs retention for material learned earlier. Moreover, this effect extends from rote episodic list-learning to more general understanding and memory for complex text materials. If the final test will induce a context similar to the study session and educators find that their students are having difficulty retaining information learned at the end of the study session, interleaved retrieval might be a simple tool to add to their curriculum; however, if they find that their students are having difficulty remembering information learned earlier in the study session, it would be best to avoid tasks that induce major context change. Awareness of how simple tasks might induce context change and how that in turn affects learning is the first step to enhancing classroom learning.

The current set of experiments examined the negative and positive consequences of interleaved unrelated semantic retrieval within a study session. When the criterion test immediately followed the study session (and thus held some resemblance to the study contexts), retrieval led to enhanced performance for later material via a reduction in the buildup of PI but
overall reduced retention for earlier material. While this effect held for both word lists and more complex text materials, it disappeared when the final criterion test was delayed. Further research is still necessary to deepen our understanding of the potentially beneficial and harmful effects of retrieval and their applications.
References


Appendix A

The first test of the critical material (e.g., List 1 or List 5) influenced performance on the second test. The results of the second test across experiments are reported here. The same methods were used as in the main analysis of the first test. Except where noted, the same models created the best fit. See Figure A1 for mean proportion correct for all three experiments.

Experiment 1. (Note that the best fit model included random intercepts for item in addition to subject). There was no reliable main effect of task ($z = .914, p = .360$) or list ($z = .693, p = .488$). The two-way interaction between task and list also was not reliable ($z = 1.539, p = .124$).

Experiment 2. The analysis revealed a main effect of list with overall higher performance in List 1 ($z = 5.200, p < .001$) but no main effect of task ($z = .191, p = .848$). A significant two-way interaction between task and list ($z = 3.612, p < .001$) revealed enhanced performance on List 1 but reduced performance in List 5 for subjects in the retrieval task compared to the distractor task. This reinstatement of retrieval effects suggests that the first test of the material shifted the current context back to the study context (therefore alleviating the effect of the delay).

Experiment 3. Analysis of the multiple choice responses revealed a marginal main effect of task with overall higher performance in the distractor condition ($z = 1.801, p = .0717$) but no main effect of text ($z = .630, p = .529$). The two-way interaction between task and text was not reliable ($z = 1.547, p = .122$). Analysis of the short answer responses revealed no main effects of task ($t = .230, p = .818$) or text ($t = .919, p = .359$) and no two-way interaction between task and text ($t = .960, p = .338$).
Appendix B

Short Answer Questions (Porcupine Text):

1. Describe how a porcupine uses its quills for defense AND the consequences for its victim.
2. Describe the negative AND positive opinions held about porcupines in different regions of the world.
3. Describe a porcupine’s typical habitat and daily routine.
4. Describe a porcupine’s defense system AND how it might be overcome by a predator.
5. Describe the porcupine’s cycle of reproduction AND characteristics of the young.

Multiple Choice Questions (Porcupine Text):

1. How many quills does a typical porcupine have?
   a. 10,000
   b. 20,000
   c. 30,000
   d. 40,000

2. How large is a typical adult porcupine (in length)?
   a. 1-2 ft
   b. 2-2 ½ ft
   c. 1 ½ - 3 ft
   d. 2-3 ½ ft

3. How long is the porcupine’s gestation period?
   a. 4 months
   b. 5 months
   c. 6 months
   d. 7 months

4. Which of the following were NOT identified as successful predators of porcupine?
   a. Dogs
   b. Bobcats
   c. Cougars
   d. Coyotes

5. What regions do porcupines generally inhabit?
   a. Northern
   b. Eastern
c. Southern  
d. Western

6. How do porcupines USUALLY sleep?  
a. On their backs  
b. In a tree  
c. Burrowed underground  
d. Under dense foliage on the ground

7. What use for porcupine quills was NOT listed in the passage?  
a. Jewelry  
b. Hair accessory  
c. Clothing  
d. Shoes

8. How many offspring does a female typically give birth to in a year?  
a. 1  
b. 2  
c. 3  
d. 4

9. In terms of size, where does the porcupine rank in the rodent family?  
a. 1\textsuperscript{st}  
b. 2\textsuperscript{nd}  
c. 3\textsuperscript{rd}  
d. 4\textsuperscript{th}

10. What do porcupines like to eat?  
a. Beaver  
b. Small rodents  
c. Bark  
d. Needles

Short Answer Questions (Coyote Text):  

1. Describe the characteristics commonly attributed to coyotes in stories AND where they originated.\footnote{This question was not used in the analyses due to multiple subjects misunderstanding to what it referred.}  
2. Compare AND contrast characteristics of the coyote and collie.
3. Describe the coyote’s eating habits in terms of both sustenance and environment. Give examples.
4. Describe the coyote’s hunting patterns and how and when they might vary.
5. Describe and give examples of the environments coyotes might inhabit.
6. Describe and give examples of coyote’s physical abilities.

Multiple Choice Questions (Coyote Text):

1. What types of prey will coyotes hunt when working in a team?
   a. Small pets
   b. Deer
   c. Sheep
   d. Young livestock

2. What specific animal was mentioned as a hunting partner?
   a. Beaver
   b. Badger
   c. Cougar
   d. Collie

3. How fast can coyotes run?
   a. 10 mph
   b. 20 mph
   c. 30 mph
   d. 40 mph

4. What’s the HIGHEST a coyote can leap?
   a. 8 ft
   b. 11 ft
   c. 14 ft
   d. 17 ft

5. What feature was NOT mentioned as something coyotes are willing to overcome?
   a. Swimming
   b. Urban environments
   c. Cyclone fences
   d. Other large predators

6. What are the coyotes yips used for?
   a. Frighten prey
b. Mating calls  
c. Warning signals  
d. General communication

7. Where did the coyote originate?  
a. Northern US  
b. Eastern US  
c. Southern US  
d. Western US

8. What was NOT mentioned as being done to keep the coyote population in check?  
a. Trap  
b. Shoot  
c. Increase natural predators  
d. Poison

9. Where are coyotes NOT found  
a. North America  
b. Central America  
c. South America  
d. Arctic

10. Which of the following were not mentioned as natural predators to the coyote?  
a. Wolves  
b. Mountain lions  
c. Bears  
d. Badgers
Figures

Figure 1. Proportion recalled as a function of list (1 or 5) and task (retrieval or distractor) for Experiment 1 (a) and Experiment 2 (b). The error bars represent the standard error of the mean.

(a)
(Figure 1 continued)

(b)

![Bar chart showing proportion correct for List 1 and List 5 for Distract and Retrieve conditions.](image-url)
Figure 2. Number of interlist intrusions as a function of list (1 or 5) and task (retrieval or distractor) for Experiment 1. The error bars represent the standard error of the mean.
Figure 3. Proportion correct as a function of list (1 or 5) and task (retrieval or distractor) for Experiment 3 multiple choice questions (a) and short answer questions (b). The error bars represent the standard error of the mean.

(a)
(Figure 3 continued)

(b)
**Figure A1.** Proportion correct as a function of study position (List 1 or List 5 for Experiments 1 and 2; Text 1 or Text 4 for Experiment 3) and task (retrieval or distractor) for the second criterion test in Experiment 1 (a), Experiment 2 (b), Experiment 3 multiple choice questions (c), and Experiment 3 short answer questions (d). The error bars represent the standard error of the mean.

(a)
(Figure A1 continued)

(b)
(Figure A1 continued)

(c)

![Bar chart showing the proportion correct for different conditions. The x-axis represents Text 1 and Text 4, and the y-axis represents the proportion correct. The chart compares Distract and Retrieve conditions.]

- Distract
- Retrieve
(Figure A1 continued)

(d)

![Bar graph](image-url)

- Y-axis: Proportion Correct
- X-axis: Text 1 and Text 4
- Bars represent Distract and Retrieve conditions

- Text 1: Distract = 0.45, Retrieve = 0.35
- Text 4: Distract = 0.60, Retrieve = 0.40

Error bars indicate variability.