May 2016

Output Interference and Testing Effects in Recognition Memory

Sean Allison Spangler
Syracuse University

Abstract

Output interference and testing effects are two common occurrences in episodic memory data. Output interference is the finding that as a recognition memory test increases in length, performance tends to decrease. This suggests that continued testing is detrimental to recognition memory. Testing effects are the general finding that repeatedly taking a test contributes to better long-term retention than repeated studying (re-reading) of the same material. This implies that practicing taking a test is beneficial to the retention and future retrieval of that information. However, there are no studies that examine both output interference and testing effects. The current study sought to test recognition memory under conditions that were conducive to both output interference and testing effects. All participants studied a list of words. Participants in the study condition then studied the list again and came back for a second appointment where they completed a recognition memory test on the words. Participants in the test condition took a recognition memory test after studying the words once. They then took the test again when they came back for the second appointment. Additionally, participants were either in the repeated or non-repeated targets condition. In the repeated target condition the words were repeated across all three list. In the non-repeated target condition, the words on the final test were only shown on the initial study list and not in the intermediary study list or test. The results showed that there were no significant differences in slope (performance on the final test over test block) between the study and test conditions and the repeated and non-repeated conditions. There was also not a significant interaction. Additionally there was not a significant difference in test accuracy between the study and test conditions. However, accuracy was significantly greater in the repeated condition than the non-repeated condition and there was a significant interaction. Post-hoc t-tests showed that, within the repeated targets condition, the test condition was significantly

more accurate than the study condition. There was not a significant difference in the non-repeated condition. These results suggest that repeated testing does not seem to affect output interference and seems to be beneficial when items are encountered more than once but not when they are encountered only once.

OUTPUT INTERFERENCE AND TESTING EFFECTS IN RECOGNITION MEMORY

by

SEAN A. SPANGLER

B.A., Franklin & Marshall College, 2013

Thesis
Submitted in partial fulfillment of the requirements for the degree of
Master of Science in *Experimental Psychology*.

Syracuse University
May 2016

Table of Contents

List of Illustrative Materials

Episodic memory is memory for autobiographical information (as encountered in different episodes). There are multiple ways to measure episodic memory, such as recognition. Recognition involves presenting study participants with a list of items (usually words) to study. They then complete a recognition test on items that were present on the study list (targets) or not present on the study list (foils). Performance on a recognition memory test is often measured according to signal detection theory (SDT; Green & Swets, 1966). SDT characterizes recognition as the ability to distinguish a signal, or an item that was actually present on the study list, from noise, or items that were not present. This means that, at test, an experiment participant can either respond "yes," meaning they remember seeing an item on the study list, or "no," meaning that they do no remember seeing it. Although there are two responses choices for each item, typically only the "yes" responses are analyzed as indications of a participant's recognition. The hit rate is the rate at which a participant responds "yes" when the item was actually present on the study list; thus, it is the rate of correct responses. The false alarm rate is the rate at which a participant responds "yes" when the item was not present on the study list; thus, it provides a measure of how often a participant recognized an item that was actually never shown before (within the context of that experiment). A third statistic, d', indicates the standardized difference between the hit and false alarm rate distributions of each participant. Essentially, d' provides an idea of how sensitive each participant is to detecting the "signal," which, in the case of recognition memory, is the ability to correctly recognize an item that was truly present on the study list as well as the ability to correctly reject an item that was not present on the study list (Green & Swets, 1966). Recognition memory researchers typically focus on different factors that help or harm the ability to successfully discriminate between targets and foils.

A type of commonly found memory effect known as a testing effect suggests that taking a memory test is beneficial to memory (Roediger & Karpicke, 2006). A testing effect is the result that repeated testing (essentially, repeated retrieval) of information contributes to better memory performance on a final test than does repeated studying of the same information. Experiments in the testing effect area of research typically include two conditions. In the study condition, participants study (reread) information multiple times. The studying is considered passive because it involves participants simply reading the information and trying to encode it but not actively retrieving or recalling it. In the test condition, participants initially study the information passively (as in the study condition) but then complete multiple tests such as free recall (Roediger & Karpicke, 2006). The tests are typically the same format and require that participants recall the same information each time. This testing is active in the sense that it requires participants to recall information they saw on the study list when given a cue. In both conditions, participants take a final test, typically after a delay, and performance on this test is taken to indicate how well they remembered the information. The test group usually performs significantly better on the final test than the study group, suggesting that something about the act of retrieving information contributes to better memory than passively studying information (Roediger & Karpicke, 2006).

For example, Roediger and Karpicke (2006) conducted an experiment that included two learning conditions (study or test) and three final test delay conditions (five minutes, two days, or one week). All participants read a prose passage. Participants were then given four 7-minute periods and were instructed to restudy (reread) the passage during each period if they were in the study condition or to write down everything they could remember from the passage during each period if they were in the test condition. Participants then took a final test on the passage (which was identical to the test the participants in the test condition had already completed four times). This final test was either given five minutes, two days, or one week after the last 7-minute period was completed. They found that there was a main effect of learning condition, such that the participants in the test condition recalled more information on the final test than participants in the study condition. There was also a main effect of final test delay such that participants in the one-week delay condition recalled the least information while participants in the five-minute delay condition recalled the most information. Importantly, there was also an interaction between learning and delay conditions such that participants in the study condition remembered more than participants in the test condition after the five-minute delay but this finding was reversed for the two day and one week delays. These results suggest that repeated studying may be more useful than repeated testing for immediate retention of information but repeated testing is more beneficial for long term retention (Roediger & Karpicke, 2006).

One suggestion as to why repeated testing contributes to better performance is that it provides better contextual information than repeated studying (Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014). The episodic context account of retrieval-based learning (or testing effects) suggests that context information (specifically, information about the temporal context of an item) is encoded when an item is encountered. When the item is encountered again, both the information about that item and its context are updated. During an item retrieval attempt, the items with updated context information should be easier to recall than items whose context information was not updated. Thus, items that are encountered multiple times should have updated context information and should be easier to recall than an item that was only encountered once. However, this does not explain why items on which subjects are repeatedly tested are better remembered than items that are repeatedly studied, since both options involve multiple encounters with the same item. The episodic context account of retrieval-based learning also suggests that there is a greater difference between the temporal contexts of information in a study list and a test list than there is between the temporal contexts of information in two study lists. Subsequently, when an item is encountered in a testing situation multiple times it becomes associated with more distinctive context features which contribute to a greater likelihood of retrieval than items whose contextual features are very similar (Karpicke, Lehman, & Aue, 2014).

Another account of testing effects is the elaborative retrieval account. This account says that when memory is searched for a target, items that are semantically related to that target become activated in memory. This activation, called semantic elaboration, is thought to occur when a person takes a test but not when a person restudies information, thus making testing more beneficial than studying for long-term retention. However, in experiments that compare the elaborative retrieval account to the episodic context account, participants in the episodic context conditions typically demonstrate better long-term retention than participants in the elaborative retrieval conditions. For example, Lehman et al. (2014) ran an experiment that included three conditions: a control condition, where participants simply studied five word lists; a retrieval practice condition, where participants recalled each list after they studied it; and an elaborative study condition, where participants said the first two words they thought of for each target word they saw. All participants then completed a free recall test of the final list. They found that the retrieval condition correctly recalled more words than the control and elaborative conditions while the elaborative condition had more intrusions than the control and retrieval conditions. (An intrusion is a word that the participant included as one he/she remembers seeing but was not actually present on the final list.) These results suggest that it is more likely that the episodic context account underlies why repeated testing is beneficial to memory (Lehman et al., 2014).

On one hand, testing effects imply that repeated testing is beneficial to memory. On the other hand, a different effect known as output interference suggests that, under certain conditions, testing is harmful. The more test items that are given, the more detrimental testing becomes to memory.

Output interference has not been thoroughly investigated in the recognition memory literature. Output interference is the finding that accuracy decreases as the number of test items increases (Tulving & Arbuckle, 1966). Output interference has been found to occur under conditions for which a variety of factors are controlled, such as study-test lag (or the distance between an item's appearance in the study list and its appearance in the test list), the presence or absence of accuracy feedback, and the temporal delay between study and test (Criss, Malmberg, & Shiffrin, 2011). These findings reinforce the idea that as a recognition memory test progresses, the test items contribute to interference, thus leading to a decline in accuracy (Malmberg et al., 2012).

Output interference occurs in recognition memory tasks such as yes/no recognition tests and alternative forced choice tests (Criss, Malmberg, & Shiffrin, 2011). Although it is commonly observed in memory test performance, there is evidence that output interference can essentially be stopped and reversed under certain conditions. This change is called a release from output interference. Malmberg, Criss, Gangwani, & Shiffrin (2012) found that when the category of test items changed halfway through the test, the decline in performance ended and memory for the test items from the new category almost returned to initial performance levels. For example, participants completed a 150-trial two alternative forced choice recognition test. The first 75 trials included only animal names and participants had to choose which animal name they saw on the study list. After the first 75 trials, the test items changed to consist only of geological terms and participants now had to choose which geological term they saw on the study list. This category change was accompanied by a sudden increase in participants' test accuracy. However, output interference occurred again as the number of trials for that new category increased. In another condition in this experiment, the category of test items changed after every five trials. Thus, the test trials composed smaller blocks of different categories. In this condition there was not a release from output interference in the same way that there was when the test consisted of two large blocks of different categories. These results suggest that output interference is affected by item similarity in that items that are similar to each other contribute to more noise in memory. When a test item is similar to the test items that preceded it, it is more difficult to determine if that item was seen on the study list (Malmberg et al., 2012).

According to the retrieving effectively from memory (REM) model of recognition memory, the items that are seen on a study list are stored as traces in memory. These traces consist of information about the items themselves as well as information about the context surrounding the items. Memory probes (such as the test item in a recognition memory test) also contain this item and context information, and are matched to existing traces in memory to determine the likelihood that a test item is old or new (Shiffrin & Steyvers, 1997). In REM, the spurious matches that result from the match between a probe and incorrect or irrelevant memory trace cause interference. REM explains that errors occur by virtue of adding information to memory during the test, leading to the build-up of interference over the course of testing (Criss, Malmberg, & Shiffrin, 2010).

According to REM, item traces can either be updated or stored as new traces when they are encountered a second time in a context similar to that of the initial encounter (Shiffrin & Steyvers, 1997). If, at test, an item is remembered as being present on the study list, the best matching trace is updated to accommodate new information. If the item is not remembered then a new trace for the item is stored. Both of these situations lead to a decrease in hit rates and a relatively constant rate of false alarms over the test list, thus producing output interference (Criss, Malmberg, & Shiffrin, 2011).

These are two common memory effects, yet they offer conflicting accounts of how taking a memory test affects memory and retention. These effects seem contradictory on the surface, but the mechanisms underlying the two effects may not be the same. The current study seeks to understand conditions under which testing effects and output interference occur and explore whether they can be explained within the same theory of memory. Output interference can be explained as a disadvantage of adding item information to memory during the test whereas the testing effect can be explained as a benefit due to matching contextual features by virtue of the presence of a test. This study will help elucidate the relationship between output interference and testing effects. In addition to a classic testing effect manipulation, we include a condition where the targets on the final test should not benefit from matching context between the immediate and final tests because different target items are included in each test.

Method

Participants

This study included 145 undergraduate students enrolled in an introductory psychology course at Syracuse University. All participants received course credit for their participation.

Data from a total of 118 participants were included in the final analyses. Thirteen participants were excluded from the final analyses because they only participated in one appointment rather than both while 14 participants were excluded due to computer and program errors. The 118 participants were distributed across the four conditions as follows: n = 36 for the study/repeated targets groups, n = 36 for the test/repeated targets group, n = 25 for the study/non-repeated targets groups, and n = 21 for the test/non-repeated targets group.

Materials

The study was run through MATLAB. The stimuli consisted of a list of 424 words. The words were common nouns with an average length of 5.95 letters and an average KF frequency of 27.81 (Kucera & Francis, 1967). The program chose words at random and anew for each participant.

Procedure

This was a 2 (study or test) x 2 (repeated or not repeated targets) between-subjects design (see Figure 1). Participants were randomly assigned to both conditions. All participants studied a list of 100 words that were shown one at a time for 3 seconds with a 0.5 second interstimulus interval (ISI) that consisted of a blank screen. They were instructed to study each word. After completing the initial study list, all participants completed a 60 second math task that served as a distractor. For the second list, half of the targets were combined with 50 novel words for a total of 100 words. The second list differed according to the condition in which the participant was randomly sorted. In the *study* conditions, participants studied the second list. All words, timing, and instructions were the same as the first time they studied the list. There was no distractor task after this study list. In the *test* conditions, participants completed a yes/no recognition test. They were instructed to select "yes" if they remembered seeing that word on the list or select "no" if they did not remember seeing it. The test included 100 trials and was self-paced. There was no distractor task after the test. Participants were then finished with their first appointment. They returned approximately 48 hours later for the final test.

The final test was a two alternative forced choice test. All participants completed this test. Participants saw two words- one word was a target and the other was a foil. They were instructed to choose which word they remembered seeing by using the mouse to click on that word. The test included 50 trials (so participants saw a total of 100 words) and was self-paced. In addition to the study and test conditions, this study included repeated or non-repeated targets conditions. The *repeated* condition is a classic testing effect manipulation where all targets were on both the study list and second list. In this condition, the two alternative forced choice test consisted of targets that were previously encountered on both the initial study list and the second study list (for the study condition) or the yes/no recognition test (for the test condition). Thus, participants saw every target a total of three times throughout the experiment. New foils were randomly chosen for every list. In contrast, the *non-repeated* condition included targets from the study list but those targets were not the ones repeated on the second list. Half of the targets from the initial study list were used as targets on the second study list or the yes/no recognition test. However, on the forced choice test the targets consisted of the half that only appeared on the initial study list and not on the second study list or recognition test. Thus, participants in the non-repeated condition only saw the targets a total of two times throughout the experiment, and they never saw them in back-to-back lists as they did in the repeated condition. The foils in the non-repeated condition were randomly chosen and new for each list.

Results & Discussion

I conducted a 2 (study vs. test) x 2 (repeated vs. non-repeated targets) between subjects ANOVA. The dependent variable was the slope of each subject's test performance, or how test performance changed over the five test blocks. (There were 100 test trials, so each block was composed of 20 trials.) The slopes were analyzed to determine the extent to which output interference occurred in each condition. As shown in Table 1 and Figure 1, there was no main effect of study/test condition, $F(1, 114) = .233$, $p = .630$. There was also no main effect of target condition, $F(1, 114) = 1.012$, $p = .317$, and there was not a significant interaction between study/test condition and target condition, $F(1, 114) = 1.479$, $p = .226$. These results suggest that output interference was not affected by whether participants repeatedly studied or were tested on the items, nor was it affected by repetition or non-repetition of targets.

I conducted a second 2 (study vs. test) x 2 (repeated vs. non-repeated targets) between-subjects ANOVA with the dependent variable of accuracy. This was a measure of each participant's performance, measured by mean accuracy on the forced choice on the final test. It was hypothesized that accuracy would be greater in the test condition than in the study condition and in the repeated condition than in the non-repeated condition. As shown in Figure 2, there was a main effect of target condition such that accuracy was significantly greater in the repeated target condition than in the non-repeated target condition, $F(1, 114) = 41.090$, $p < .01$. There was no significant main effect of study/test condition, $F(1, 114) = 0.233$, $p = 0.630$. However, there was a significant interaction, $F(1, 114) = 5.893$, $p = 0.017$. The interaction suggests that within the repeated targets conditions, the test condition had greater accuracy than the study condition; however, this was not the case within the non-repeated targets conditions. A post hoc t-test showed that, within the repeated condition, the test condition was significantly more accurate than the study condition, $t(70) = 2.530$, $p = 0.014$. There were no differences between study and test condition accuracy in the non-repeated condition, $t(44) = -1.081$, $p = 0.286$.

Additionally I conducted two independent samples t-tests on the slope and accuracy measures from the first test in the test condition, where accuracy was measured by d-prime. There was no significant difference in slope of overall accuracy between the repeated and non-repeated conditions, $t(59) = -0.162$, $p = 0.872$. There was also no significant difference in accuracy between the repeated and non-repeated conditions, $t(59) = 0.608$, $p = 0.545$. Finally, there was not a significant difference between the hit rates of the repeated and non-repeated conditions, $F(1, 58) = 2.146$, $p = 0.149$, nor was there a significant difference between the false alarm rates of the repeated and non-repeated conditions, $F(1, 58) = 0.000$, $p = 0.997$. (See Figure 3.)

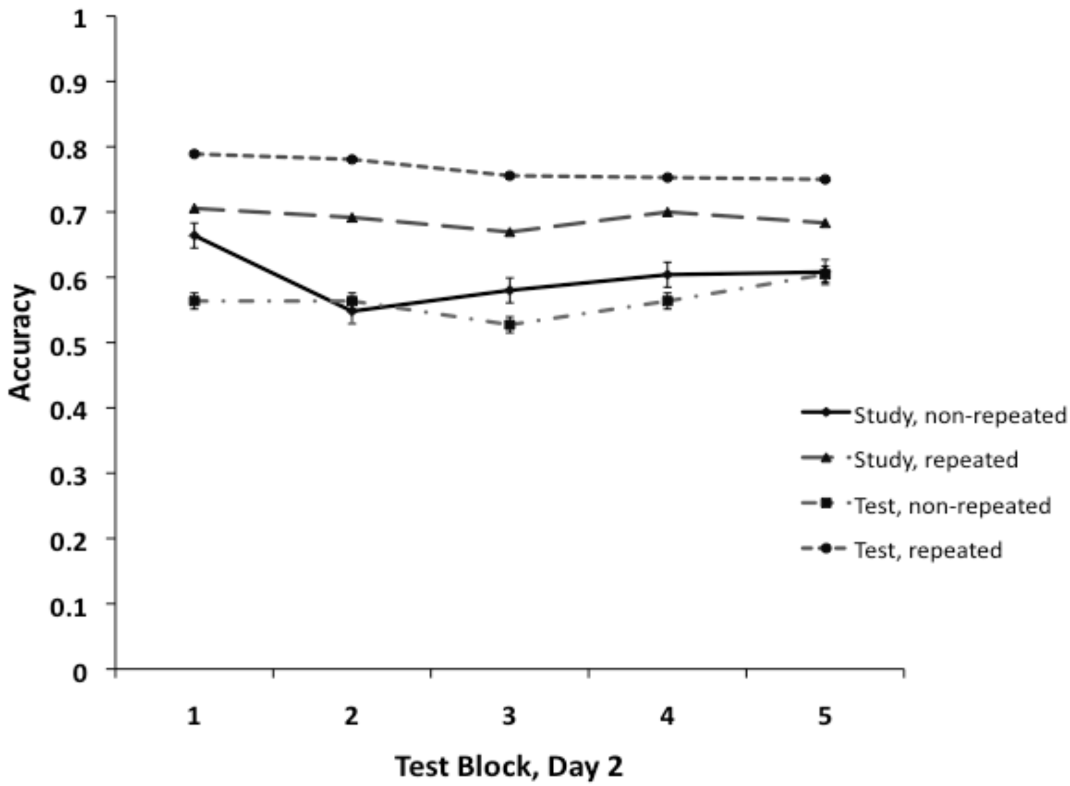

Figure 1. Mean accuracy per test block for four conditions on the Day 2 forced-choice test.

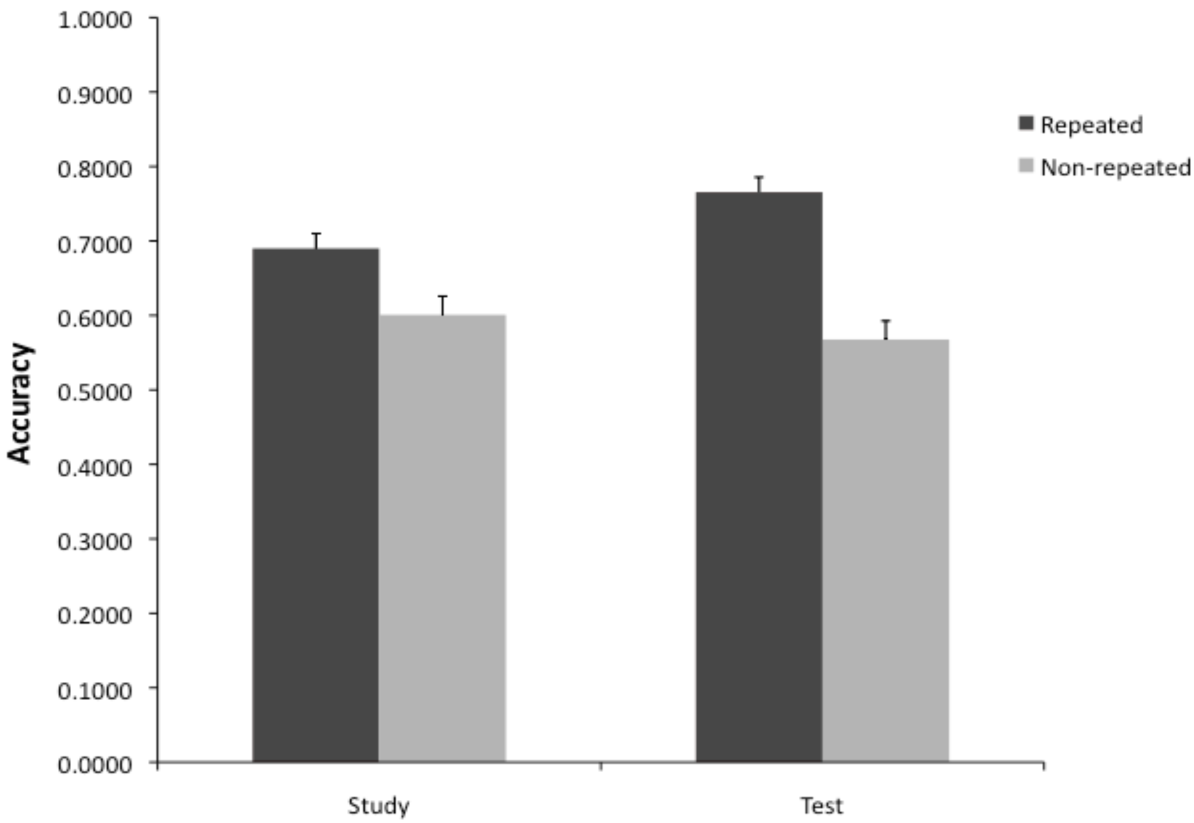

Figure 2. Mean accuracy on Day 2 forced-choice test for repeated study, repeated test, non-repeated study, and non-repeated test conditions.

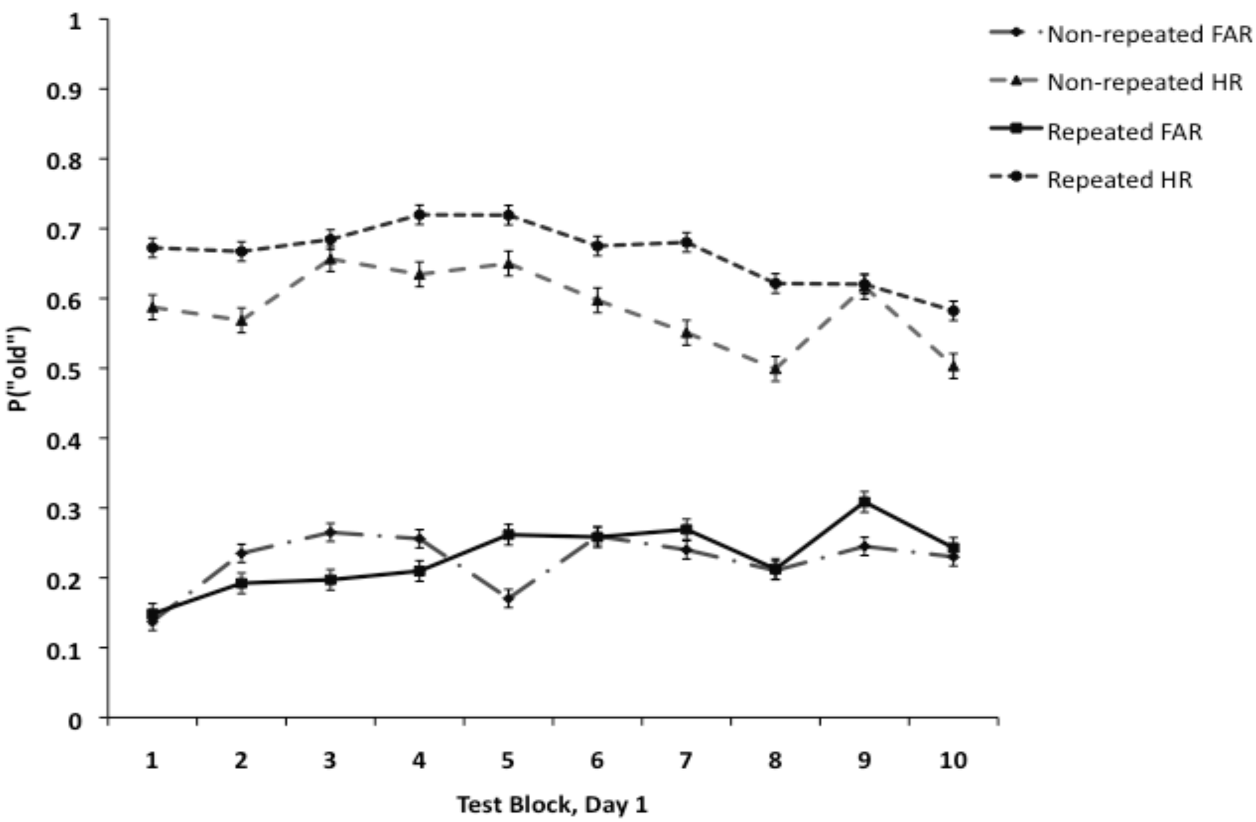

Figure 3. Mean hit (HR) and false alarm (FAR) rates per test block for test condition during Day 1 test.

	Mean (Standard Error)	
	Study	Day 2 Test
Repeated	-0.004 (0.008)	-0.011 (0.008)
Non-repeated	-0.006 (0.010)	0.010 (0.011)

Table 1. Mean slope and standard error values for study, Day 2 final test, repeated, and non-repeated conditions.

The present study sought to better understand how testing affects episodic memory. Specifically, it aimed to explore the relationship between output interference, or the decrease in performance as the number of test items increases, and testing effects, or the finding that repeated test taking contributes to better long-term retention than repeated studying does.

The results show that there was a testing effect in the repeated targets condition but not in the non-repeated targets condition. These results support the episodic context account of testing effects because they demonstrate an advantage for repeated exposure to the same words in the same context over repeated exposure to the same words in different contexts (Karpicke et al., 2014; Lehman et al., 2014). According to the episodic context account, the items in the repeated condition should have had updated context information and more distinctive context features since they were encountered on both the study list and the first test list. The items in the non-repeated condition were only encountered on the study list and should not have had updated context information or as distinctive context features as the items in the repeated condition, thus making them more difficult to retrieve. It is not quite clear what these results suggest about the elaborative retrieval account of testing effects. The results could be viewed as support for this account as well since there was a testing effect in the repeated targets condition. However, there should not have been a difference in semantic elaboration between the repeated and non-repeated conditions. Therefore, this account cannot explain why there was not a testing effect in the non-repeated condition.

The results suggest that output interference was not differentially affected by the study vs. test conditions nor by the repeated vs. non-repeated target conditions. Not only were there no significant differences in output interference between these conditions, but also we did not observe output interference to the extent that it usually occurs in recognition memory tests. It is possible that we did not observe output interference in the non-repeated condition because accuracy was close to chance in both the study and test conditions. However, accuracy in the repeated target conditions was above chance. One possibility is that the items were not as easy to confuse as the test continued because they had been previously encountered in a testing context. This could potentially eliminate output interference because participants could be more likely to correctly identify items they had already seen or correctly reject items that they know are new.

These results have several important implications. First, they suggest that repeated testing does not necessarily affect output interference. Additionally, seeing the targets twice prior to the final test does not seem to decrease the extent of output interference relative to only seeing the targets once prior to the final test. These results do not align with the initial hypothesis regarding output interference, nor do they completely align with the testing effect literature. A testing effect is typically observed by comparing final test accuracy, and not change in performance over the test. However, the current study results suggest that taking a test multiple times does not seem to benefit recognition memory as a test's length increases. Furthermore, it appears that greater exposure to final test targets, and the assumed benefits of this exposure, such as those suggested by the episodic context account of testing effects, also did not improve recognition memory over the duration of the final test. These results do not provide much insight as to the types of conditions or aspects of testing contribute to output interference. If output interference is not lessened by taking a test multiple times or by seeing the test items multiple times, what aspect of taking a test contributes to the decline in performance? This seems like a critical point for future research regarding output interference.

References

Criss, A.H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 316-326.

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: John Wiley & Sons, Inc.

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

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

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

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science, 23,* 115-119.

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

Shiffrin, R. M. & Steyvers, M. (1997). A model for recognition memory: REM- retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145-166.

Vita

Sean A. Spangler
454 Huntington Hall
Cognition, Brain, and Behavior Program
Syracuse University
484-269-2120
saspangl@syr.edu

EDUCATION

- **B.A. in Psychology, *cum laude*, 2013, Franklin & Marshall College,** Lancaster, Pennsylvania
  - HONORS THESIS: *Individual Differences in the Working Memory of Concurrent Information in Multiple Modalities*

RESEARCH EXPERIENCE

- **Research Interests**
  - Episodic memory, individual differences in memory, relationship between memory and intelligence, relationship between episodic memory and working memory
- **Research Positions**
  - Graduate student in Cognition, Brain, and Behavior Program, Syracuse University, 2013-present
    - Advisor: Dr. Amy Criss
  - Research assistant and honors thesis student, Psychology Department, Franklin & Marshall College, 2012-2013
    - Advisor: Dr. James Geer
- **Posters**
  - Spangler, S. A., Criss, A. H., Hewett, S., Shefner, J., & Wang, D. (2014, November). *The (null) effect of acute naproxen administration on memory.* Poster presented at 55th annual meeting of the Psychonomic Society, Long Beach, CA.
  - Spangler, S. A., Criss, A. H., Hewett, S., Shefner, J., & Wang, D. (2014, May). *The effect of acute naproxen administration on memory.* Poster presented at 2014 Context and Episodic Memory Symposium, Philadelphia, PA.
  - Geer, J. H., Spangler, S. A., & Cohn, M. F. (2013, March). *Emotional words and working memory.* Poster presented at annual meeting of Eastern Psychological Association, New York City, NY.

ACADEMIC AWARDS AND HONORS

- Syracuse University Fellowship, 2013-2014
- Kenneth A. Brookshire Memorial Prize for excellence in independent research, Psychology Department, Franklin & Marshall College, 2013
- Departmental honors in psychology, Franklin & Marshall College, 2013
- Honors List (3.70 GPA or higher), four consecutive semesters, 2011-2013
- Dean's List (3.25 GPA or higher), four consecutive semesters, 2009-2011
- The CIRI Foundation General Semester Scholarship, eight consecutive semesters, 2009-2013
- Franklin & Marshall College Dean's Scholarship, 2009-2013

ACADEMIC/TEACHING EXPERIENCE

- **Teaching Assistant, Syracuse University, Syracuse, NY, 2014**
    - PSY 205 Foundations of Human Behavior: lead four recitation sections per week; responsible for reviewing material covered in lecture and teaching new information, creating activities for class involvement, and grading student assessments and assignments
- **Teaching Assistant, Franklin & Marshall College, Lancaster, PA, 2012**
    - Introduction to Psychology: assisted professor in laboratory explanations and demonstrations, answered students' questions, assisted students in lab activities

MEMBERSHIPS

- **Psi Chi**, 2012