DISSERTATION

MULTIPLE CHOICE TESTING AND THE RETRIEVAL HYPOTHESIS OF THE TESTING EFFECT

Submitted by

Amanda E. Sensenig

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2010

COLORADO STATE UNIVERSITY

March 29, 2010

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY AMANDA E. SENSENIG ENTITLED MULTIPLE CHOICE TESTING AND THE RETRIEVAL HYPOTHESIS OF THE TESTING EFFECT BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate work

Deana Davalos

David McCabe

Michael DeMiranda

Advisor: Edward DeLosh

Department Chair: Ernest Chavez

ABSTRACT OF DISSERTATION

MULTIPLE CHOICE TESTING AND THE RETRIEVAL HYPOTHESIS OF THE TESTING EFFECT

Taking a test often leads to enhanced later memory for the tested information, a phenomenon known as the "testing effect". This memory advantage has been reliably demonstrated with recall tests but not multiple choice tests. One potential explanation for this finding is that multiple choice tests do not rely on retrieval processes to the same extent as other types of tests. The set of experiments reported here examines the retrieval hypothesis of the testing effect in multiple choice testing. Experiment 1 is a replication and extension of previous research (Roediger & Marsh, 2005) with the addition of a re-study comparison condition. Experiments 2a and 2b encouraged participants to engage in retrieval processes during multiple choice tests. Experiment 3 implemented a version of the remember/know paradigm in order to assess retrieval of individual items on a multiple choice test. Overall, multiple choice testing did not produce a memory advantage over re-studying the material in the

experiments reported here. The results of these experiments are discussed in light of the retrieval hypothesis of the testing effect.

Amanda E. Sensenig
Department of Psychology
Colorado State University
Fort Collins, CO 80523
Spring 2010

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The *testing effect* refers to a robust, reliable phenomenon in the memory literature whereby previously tested information is better remembered than untested or re-studied information. This finding is of inherent interest to instructors who seek to help students retain course information. Studies conducted in a laboratory setting as well as in actual or simulated classroom settings have shown that testing can be useful not only for the assessment of knowledge, but also can be an effective learning tool (for a review see Roediger & Karpicke, 2006a).

However, not all types of tests contribute equally to long-term retention of information. McDaniel and Masson (1985) illustrated that cued recall tests led to better long-term retention than recognition tests. Glover (1989) and Carpenter and DeLosh (2006) showed that intervening free recall tests led to the best long-term retention of tested information, regardless of final test type (free recall, cued recall, or recognition). This research suggests that intervening tests requiring the retrieval or generation of information from memory, such as essay or short answer tests, lead to better long-term retention than recognition tests, such as multiple choice tests.

Such findings have direct implications for the classroom. For a variety of reasons, many instructors prefer to give multiple choice tests. Often class size,

time constraints, or other classroom variables are prohibitive to administering other types of tests. It is also easier to grade multiple choice tests quickly and with high reliability, unlike the subjective nature which can be involved in grading essay tests. An instructor can easily track performance on each question, in particular taking into account which response options or lures are best to include on a certain question. This advantage can potentially lead to a better understanding of how well the questions assess students' knowledge. Further, if instructors choose to incorporate more tests into their courses as a tool for student learning, as the testing effect literature suggests they should (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Roediger & Karpicke 2006b), multiple choice tests are potentially a quick and easy way to accomplish this. However, although the testing effect is robust and reliable with recall tests (e.g., Carpenter & DeLosh, 2006; Glover, 1989), studies of the testing effect with multiple choice tests have produced mixed results.

Some studies have shown a memory benefit stemming from multiple choice testing. Butler, Karpicke, and Roediger (2007) had participants read passages and take a multiple choice test over key concepts. Their results showed enhanced memory for tested information relative to information that was not subjected to an intervening memory test. Similar results were reported in another study testing memory for information from passages using multiple choice tests (Odegard & Koen, 2007). However, both of these studies employed a no-test condition as the comparison to multiple choice testing. In these cases, rather than having participants re-study some of the information as a comparison

condition, the information did not appear during the intervening phase at all. Thus, the reported testing advantage for multiple choice testing may simply be an artifact of exposure time, such that information that was processed a second time on the intervening test was better remembered than information that was not processed a second time.

Several studies have employed a re-study condition with different materials and report another result. McDaniel, Anderson, Derbish, and Morrisette (2007) conducted a study in a web-based course. They reported that although information initially tested with short answer or multiple choice tests led to a testing effect relative to untested information, only the intervening short answer test led to a memory benefit over re-studied information. Similarly, Kang, McDermott, and Roediger (2007) showed that when a re-study condition was included, memory for information tested in multiple choice format was no better than memory for the re-studied information.

One explanation for the lack of an advantage for multiple choice testing over re-studying concerns the memory processes taking place at the time of the tests. The retrieval hypothesis of the testing effect says that the long-term retention benefit conferred by testing is contingent on the act of coming up with, or retrieving, the information (Cuddy & Jacoby, 1982; Dempster, 1996; Glover, 1989). Studies in which the degree of retrieval is manipulated have provided evidence in favor of this explanation. For example, Carpenter and DeLosh (2006) showed that when fewer cues were provided on an intervening test, subsequent memory for the tested information was better than when more cues

were provided. One explanation for this result is that providing fewer cues necessitated more complete retrieval processes at the time of the intervening test. The retrieval hypothesis may, therefore, help explain why a robust testing advantage is typically produced with recall tests but only sometimes with recognition tests. Free recall tests are commonly believed to require the most complete retrieval processes, due to the lack of external cues. In contrast, recognition tests can be completed based on the familiarity of the responses rather than retrieval of the correct answer (Chan & McDermott, 2007; Yonelinas, 2002).

The experiments described here were designed to address several key questions concerning the testing effect (or lack thereof) in multiple choice tests. Experiment 1 is a replication of a study conducted by Roediger and Marsh (2005) but includes a re-study comparison condition to address the question of whether multiple choice testing leads to a long-term retention advantage over re-studying, rather than over a no-test condition. Experiments 2a and 2b test the retrieval hypothesis by inducing retrieval of the answer on a multiple choice test. It was hypothesized that requiring more complete retrieval processes than is usually necessary with multiple choice tests may produce a reliable testing effect. Experiment 3 employed a version of the remember/know paradigm (Gardiner, 1988; Tulving, 1985) to investigate whether participants reported different states of awareness associated with the selected multiple choice responses. This procedure was implemented to allow for a comparison between the effects of familiarity and recollection on long-term retention in multiple choice tests. Taken

together, the findings from these experiments may inform the retrieval hypothesis of the testing effect.

CHAPTER 2

EXPERIMENT 1

Method

Participants. Seventy-eight students from the General Psychology and Research Methods courses at Colorado State University participated in this experiment. Data from 4 participants were excluded due to failure to follow instructions. This experiment was run in groups of 10-16 people using a within-subjects design and was completed within one hour.

Materials. Eighteen of the passages used by Roediger and Marsh (2005) and Odegard and Koen (2007) were chosen for use in the experiment. These passages originated in the reading comprehension sections of the Test of English as a Foreign Language and the Graduate Record Exam practice test books, and covered non-fiction topics such as "Laura Ingalls Wilder" and "sea otters". They were adapted for the current experiments and ranged from 225-300 words in length. Researchers selected three facts from each passage and these facts were assigned to one of three conditions in the intervening test phase. One fact from each passage was presented as a multiple choice question, with four possible responses. Another was presented as a true statement about the passage and served as a re-study condition. The third fact was not presented during the intervening phase, and therefore did not appear until the final test.

The assignment of the selected facts to the three conditions was counterbalanced such that each fact was presented as a multiple choice question, a re-study statement, and a non-tested (and therefore non-presented) item on different versions of the intervening test. Three 36-item intervening tests were designed, consisting of 18 multiple choice questions and 18 re-study items. The final test consisted of 79 short answer questions, which included all three of the original facts selected from each passage, transformed into question format. There were also 25 filler questions. The filler questions were included solely to replicate the methodology used by Roediger and Marsh (2005) and thus are not included in the analyses or discussion of the results in the current experiments.

Procedure. Participants read the passages one at a time as they were projected onto a screen at the front of a classroom. After completing each passage, participants placed a checkmark next to the corresponding number on a record sheet. Each passage was shown for 90 s, which was judged to be sufficient time for all participants to read through the passage once.

Upon completion of the passages, participants entered the intervening phase. Participants were instructed to answer the multiple choice questions by circling the most appropriate response, and to simply read and check off the statements, as these were true statements about the passages they had just read. They were given 6 min to complete this phase.

Following the intervening test, participants completed a 5 min distracter task in which they were asked to write down as many of the 50 states as they could. Participants were given up to 20 min to complete the final short answer

test (although most took no more than 15 min) and were instructed not to guess. If they did not know the answer to a particular question, they were instructed to draw a line through the answer space. These instructions were given to replicate previous studies as closely as possible. When everyone in the group had finished their test, participants were debriefed and dismissed.

Results and Discussion

The proportion of questions answered correctly on the intervening multiple choice test (M = .67, SD = .15) was comparable to other experiments conducted with the same materials (cf. Odegard & Koen, 2007, Experiment 1: M = .69, SD = .15; Experiment 2: M = .70, SD = .14). For this and subsequent experiments, the final short answer test results will be analyzed and reported in two ways: using unconditionalized data and data conditionalized on intervening test performance. In all cases, an alpha level of .05 will be used.

Unconditionalized Final Test Performance

In order to determine whether there was a memory advantage on the final short answer test for previously tested information, a repeated-measures analysis of variance (ANOVA) was conducted on the proportion of short answer questions answered correctly. There was a significant difference between conditions [$F(2,146)$= 107.74, MSE = 1.36, p< .00, η_p^2 = .60]. Post-hoc Tukey HSD (honestly significant difference) tests indicated that both multiple choice (M = .49, SD = .16) and re-study conditions (M = .56, SD = .18) led to significantly better memory than the previously untested information (M = .29, SD = .14), $ts(73)$ = 10.78 and 13.05, respectively, and that re-study resulted in significantly

better memory than the multiple choice condition, $t(73) = 3.78$ (see Figure 2.1). This result illustrates the lack of a testing advantage for the multiple choice condition over the re-study condition.

The results of Experiment 1 are consistent with those of Roediger and Marsh (2005) and Odegard and Koen (2007) in that prior testing produced a memory advantage relative to a no-test condition. However, the present experiment also included a re-study condition, and in that case, there was no testing effect; when the comparison condition was not at a disadvantage with regard to exposure time, the testing advantage was eliminated. The finding that prior testing in multiple choice format does not enhance later memory performance any more than simply re-studying the information is consistent with other studies that have utilized a re-study comparison condition (e.g., Kang et al., 2007; McDaniel et al., 2007).

The results of Experiment 1 also showed that the re-study condition led, in fact, to better memory than the test condition, a finding that may be due to differences in exposure time to the specific facts from the passage. In the re-study condition, the key fact is presented again in isolation, whereas in the test condition, the fact in question is not subject to the same focused, exclusive re-presentation. The use of a re-study comparison condition is designed to *better* equate exposure time across conditions. However, it may put the test condition at a disadvantage, because although all re-studied items are presented and processed in the intervening phase, theoretically participants are only exposed to a tested item again if they are able to successfully answer the question. Thus,

those tested items that are not successfully completed may be at a disadvantage in terms of exposure time as compared to items in the re-studied condition. One way testing effect researchers often try to correct for this is to analyze final test performance conditionalized on successful intervening test performance. Limiting the analysis to only the successfully answered questions ensures that the fact of interest was attended to again in the test condition, as was the case in the re-study condition.

Conditionalized Final Test Performance

In this section, final test performance for previously tested items was conditionalized to include only those questions answered correctly during the intervening test phase. In order to examine whether there was a memory advantage on the final short answer test for previously tested information, a repeated-measures ANOVA was conducted on the proportion of short answer questions answered correctly. There was a significant difference between conditions [$F(2, 146) = 175.55$, $MSE = 3.41$, $p < .00$, $\eta_p^2 = .70$] and post-hoc Tukey HSD tests indicated that the multiple choice condition ($M = .72$, $SD = .21$) led to significantly better memory than the re-study condition ($M = .56$, $SD = .18$), $t(73) = 6.56$. In addition, both of these conditions led to better memory than the no-test condition ($M = .29$, $SD = .14$), $ts(73) = 17.71$ and 13.05 for the multiple choice and re-studied conditions, respectively. These results show that conditionalizing the data led to a reversal of the unconditionalized findings such that tested information was better retained than re-studied information (see Appendix A for a table of the results). Thus, when the analysis is limited to tested

items for which we can be confident that the factual information of interest has been attended to and processed, a testing effect emerges even with a multiple choice test.

Note, however, that this analysis has limited real-world validity. In a classroom setting, we are most interested in the question of whether taking tests is better for subsequent memory performance than re-studying information, regardless of how well students do on the initial tests. Experiments 2a and 2b therefore examine whether multiple choice tests can be modified to encourage retrieval processes, and in that way, might yield a net overall advantage for tested information over re-studied information, even in the unconditionalized data.

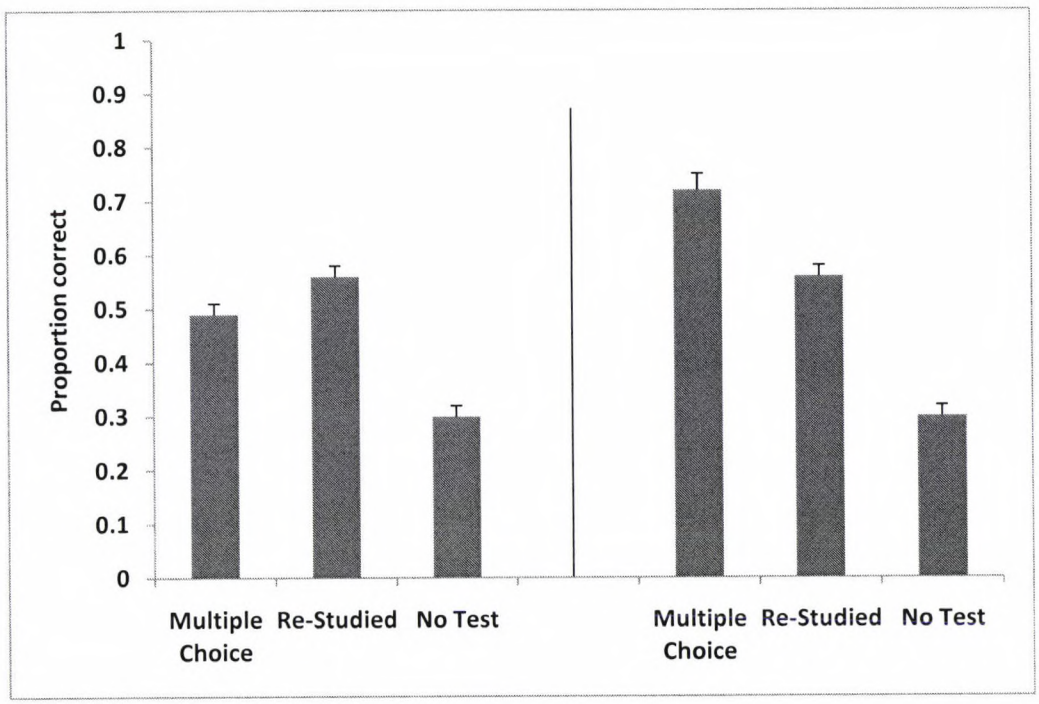

Figure 2.1: Unconditionalized final test performance (left panel) and conditionalized final test performance (right panel) for Experiment 1.

CHAPTER 3

EXPERIMENT 2a

Experiment 2a was designed to more directly assess the retrieval hypothesis of the testing effect in multiple choice testing. This hypothesis posits that a memory advantage emerges for tested items because an intervening test requires the use of retrieval processes, whereas re-studying does not elicit these same processes (Carpenter & DeLosh, 2006; Cuddy & Jacoby, 1982; Dempster, 1996; Glover, 1989). The rationale for Experiment 2 was that perhaps intervening multiple choice tests do not result in a testing advantage because they do not require retrieval of the information to the same extent as other types of tests. Thus, if participants are encouraged to retrieve the information on multiple choice tests, it may confer a testing advantage that appears even in the unconditionalized data.

Method

Participants. Eighty-eight General Psychology and Research Methods students at Colorado State University participated in this experiment. Data from six participants were excluded from the analyses due to failure to follow instructions. The experiment was conducted in groups of 1-8 people using a within-subjects design.

Procedure. The same materials as those described in Experiment 1 were used for Experiment 2a. Participants read 18 passages presented in a booklet

and placed a checkmark next to the corresponding number on a sheet when they finished each passage. Participants were allowed 90 s to read each passage, which was judged to be sufficient time to read through the passage once.

As in Experiment 1, three facts were selected from each passage. Because previous research suggests that retrieval of one fact can strengthen memory for related but untested facts (Chan, McDermott, & Roediger, 2006), all three facts from a passage were assigned to the same condition for the intervening test phase. Although all three facts from a particular passage were assigned to the same condition, the assignment of passages to conditions was counterbalanced such that all facts appeared equally often in each condition. During the intervening phase, items were presented one at a time on the screen in random order. Items in the *re-study condition* were presented as statements to be read and checked off on an answer sheet. In the *standard multiple choice condition*, the question stem and the four possible responses were presented together on the screen, and participants recorded the letter of the response they selected. In each of the first two conditions, the information was presented for 10 s followed by a 2 s inter-stimulus interval. In a *multiple choice plus retrieval condition*, the question stem appeared alone on the screen. Participants were instructed to covertly retrieve the answer when this occurred. After 5 s the four possible responses appeared and the participants had another 5 s to record the letter of the response they selected.

Following the intervening test phase, participants solved math problems for 5 min as a distracter task. Finally, a short answer test similar to the one in

Experiment 1 was administered. This test consisted of items previously re-studied, tested in the standard multiple choice condition, tested in the multiple choice plus retrieval condition, and filler items. Participants had 15 min to complete this test.

Results and Discussion

Performance on the Intervening Multiple Choice Tests

The analyses for Experiment 2a parallel those presented for Experiment 1. First, a comparison of performance on the two multiple choice tests was conducted using the proportion of correct responses on the multiple choice tests for the standard multiple choice condition (M = .71, SD = .15) and the multiple choice plus retrieval condition (M = .71, SD = .13). The analysis confirmed that participants answered the questions in these conditions at a similar rate [$F < 1$, $p > .05$].

Unconditionalized Final Test Performance

Final memory performance on the previously re-studied items presented with the standard multiple choice items (M = .57, SD = .22) and those presented with the multiple choice plus retrieval items (M = .52, SD = .21) did not differ, as indicated by a repeated-measures ANOVA [$F <1$, $p > .05$]. Thus, all re-studied items were combined into one single condition for the subsequent analyses.

A repeated-measures ANOVA was conducted on the proportion of correct responses on the final short answer test for the standard multiple choice (M = .52, SD = .19), the multiple choice plus retrieval (M = .51, SD = .16), and the re-studied (M = .55, SD = .19) conditions (See Figure 3.1). This analysis

demonstrated no significant differences in final memory performance between conditions [$F(2, 162) = 2.29$, $MSE = .03$, $p > .05$, $\eta_p^2 = .03$].

Conditionalized Final Test Performance

As in Experiment 1, final test performance for previously tested items was conditionalized to include only those questions answered correctly during the multiple choice test phase. A repeated-measures ANOVA was conducted on the proportion of short answer questions answered correctly in each of the 3 conditions. There was a significant difference between conditions [$F(2, 162) = 23.37$, $MSE = .47$, $p < .00$, $\eta_p^2 = .22$]. Post-hoc Tukey HSD tests indicated that the standard multiple choice condition ($M = .68$, $SD = .17$) led to similar memory performance on the final test when compared to the multiple choice plus retrieval condition ($M = .68$, $SD = .17$), $t < 1$.However, both previously tested conditions led to better memory than re-studying ($M = .55$, $SD = .19$), ts(81) = 6.00 and 5.42 in the standard multiple choice and multiple choice plus retrieval conditions, respectively. As in Experiment 1, these results show that conditionalizing the data led to a robust testing effect for both types of multiple choice tests.

The findings from Experiment 2a do not provide support for the hypothesis that inducing retrieval will lead to a memory advantage on the final short answer test relative to taking a standard multiple choice test. However, participants were asked to covertly retrieve the information in the multiple choice plus retrieval condition. Thus, it is not possible to know whether participants attempted to retrieve or successfully retrieved the information in this condition. They may have simply waited passively until the response options appeared on the screen, or

may have tried to retrieve the answer and failed. In order to more directly assess the effect of inducing retrieval in multiple choice testing, an additional experiment was conducted. This additional experiment followed the same general procedure as Experiment 2a, except rather than asking participants to read the question stem and covertly retrieve the answer prior to seeing the four response options during the intervening phase, participants were asked to retrieve the information and write it down on their response sheets before the response options appeared. In this way, participants' retrieval of the correct response could be assessed more directly.

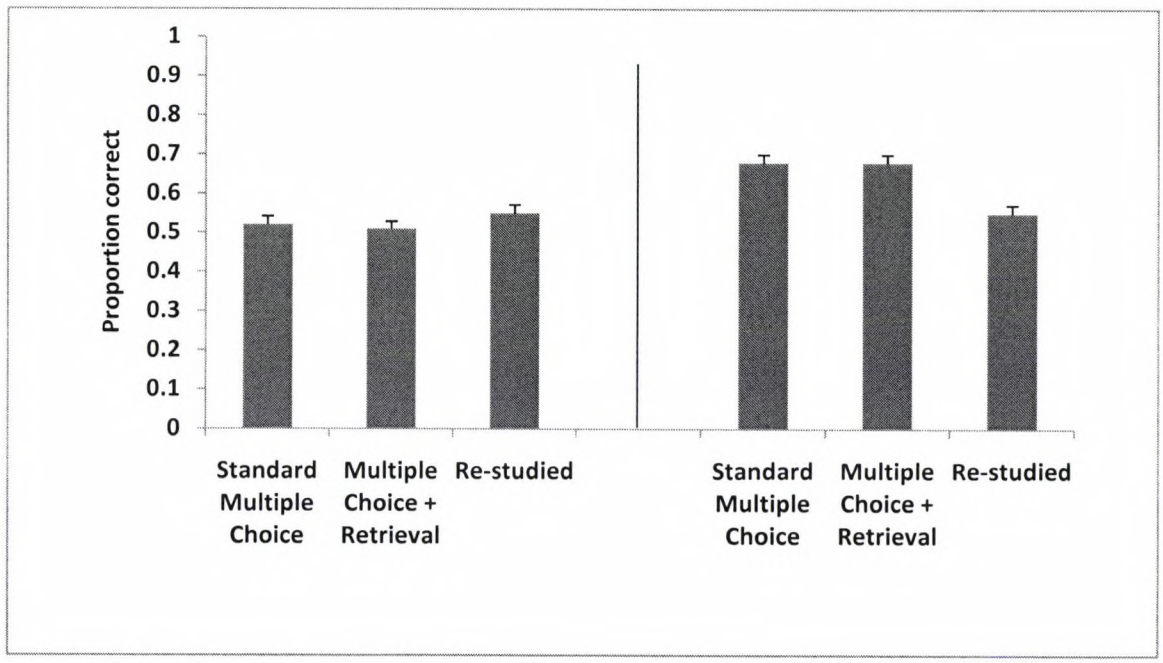

Figure 3.1: Unconditionalized final test performance (left panel) and performance conditionalized on intervening test performance (right panel) for Experiment 2a.

CHAPTER 4

EXPERIMENT 2b

Method

Participants. Sixty-five General Psychology and Research Methods students at Colorado State University participated in this experiment. Data from 7 participants were excluded due to failure to follow instructions. The experiment was conducted in groups of 2-8 people and followed a within-subjects design.

Procedure. The materials and procedure for Experiment 2b were the same as those described for Experiment 2a, with the following minor changes. During the intervening phase, items in the multiple choice plus retrieval condition were presented the same way as in the previous experiment, but participants were asked to view the question stem, retrieve the answer, and write it down on the answer sheet. Because they were asked to complete this extra step, the information was presented for 7 s rather than 5 s.

After 7 s passed, the four response options appeared on the screen and participants were instructed to record the letter of the response they selected on the line next to their written answer. Five seconds were allotted to complete this part of the question. During the multiple choice plus retrieval condition participants were encouraged to write down a retrieved response if at all possible, and only to leave it blank if they had no idea whatsoever as to the answer. They were also instructed to always write down a letter response during

the second phase of each question. Further, participants were instructed not to go back and change their answers after the response options appeared on the screen. Because of the additional time given for the multiple choice plus retrieval questions, the standard multiple choice questions and the re-studied items were adjusted to 12 s in the current experiment, so as to equate for overall exposure time. All other aspects of the experiment were the same as described in Experiment 2a.

Results and Discussion

Performance on the Intervening Multiple Choice Tests

The analyses for Experiment 2b parallel those presented for Experiment 2a. A comparison of performance on the two multiple choice tests was conducted using the proportion of correct responses on the multiple choice tests for the standard multiple choice condition (M = .67, SD = .14) and the multiple choice plus retrieval condition (M = .67, SD = .13). A repeated-measures ANOVA confirmed no significant difference in initial test performance between the two test conditions ($F < 1$, $p > .05$).

Unconditionalized Final Test Performance

As in Experiment 2a, there was no difference in performance on re-studied items appearing with the standard multiple choice condition (M = .52, SD = .19) and those intermixed with the multiple choice plus retrieval condition (M = .50, SD = .18), thus all re-studied items were combined into a single condition for the following analyses. A repeated-measures ANOVA was conducted on the proportion of correct responses on the final short answer test for the standard

multiple choice (M = .49 SD = .15), the multiple choice plus retrieval (M = .49, SD = .14), and the re-studied (M = .51, SD = .16) conditions (see Figure 4.1). This analysis demonstrated no significant differences across conditions [$F(2,114)$ = .75, MSE = .01, $p > .05$, $\eta_p^2 = .01$].

Conditionalized Final Test Performance

As in previous experiments, final test performance for previously tested items was conditionalized to include only those questions answered correctly during the multiple choice test phase. This included performance in the standard multiple choice condition (M = .69, SD = .16), as well as the multiple choice plus retrieval condition, which was broken down into two sub-categories: those questions for which an answer was correctly retrieved and recorded [multiple choice plus retrieval (retrieved); M = .93, SD = .11] and those questions for which the correct response letter (A-D) was recorded on the initial test [multiple choice plus retrieval (letter); M = .71, SD = .14]. A repeated-measures ANOVA was conducted on the proportion of short answer questions answered correctly in each of the 3 testing conditions as well as the re-studied condition. Overall, there was a significant difference in performance across conditions [$F(3,171)$ = 97.04, MSE = 1.72, $p < .00$, η_p^2 = .63]. Final test performance was significantly better in all three test conditions than in the re-study condition, when examining the conditionalized data, ts(57) = 7.54, 16.15, and 8.13 for the standard multiple choice, multiple choice plus retrieval (retrieved), and multiple choice plus retrieval (letter) conditions, respectively.. In addition, performance for items in the multiple choice plus retrieval (retrieved) condition was significantly better than the other

test conditions *t*s(57) = 9.19 and 8.13 in the standard multiple choice and multiple choice plus retrieval (letter) conditions, respectively.There was no difference in final test performance for the standard multiple choice and multiple choice plus retrieval (letter) conditions,$t < 1$.

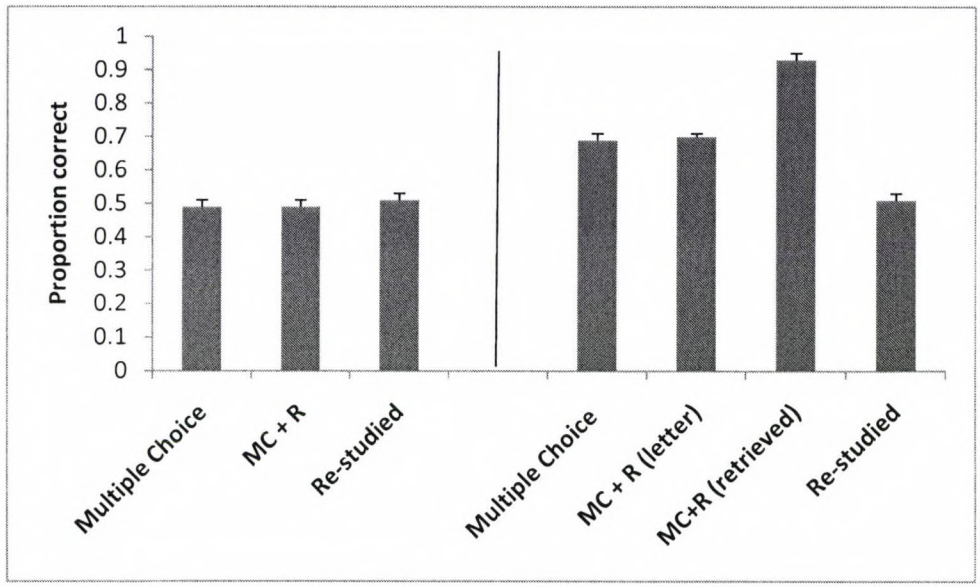

Figure 4.1: Unconditionalized final short answer test performance (left panel) and performance conditionalized on intervening test performance (right panel) for Experiment 2b.
MC+R = Multiple Choice plus Retrieval

The pattern of results shown in the conditionalized data indicates that simply instructing participants to attempt to retrieve the response [multiple choice plus retrieval (letter)] is not enough to improve performance above the level of standard multiple choice testing. However, when the analysis is limited to cases in which the correct response was retrieved [multiple choice plus retrieval (retrieved)] final test performance was significantly better than in all other conditions. This may potentially be interpreted as providing support for the retrieval hypothesis, however, it is important to note that item difficulty could have

played a role in this result. It is possible that some items were inherently easier than others, and the easy items may be the ones that participants retrieved on the intervening test. Thus, the conditionalized data must be interpreted as providing only tentative support for the retrieval hypothesis of the testing effect in multiple choice testing.

Overall, the key finding of interest in Experiments 2a and 2b is the lack of a difference between the multiple choice plus retrieval, multiple choice, and re-study conditions, even when overt retrieval was required. One interpretation of this result is that, counter to the retrieval hypothesis, the act of generating or retrieving information from memory does not necessary convey a memory advantage, at least for the types of materials and procedure examined here. An alternative possibility is that students already engage in retrieval to some extent on standard multiple choice tests, hence the lack of an overall advantage for the multiple choice plus retrieval condition. For example, on certain questions students may truly be remembering the material from the encoding episode, whereas on other questions they may simply be selecting the most familiar response out of the four possible. Experiment 3 was designed to parse these instances apart in order to compare cases where participants rely on familiarity to those in which a more complete form of retrieval is taking place.

CHAPTER 5

EXPERIMENT 3

Experiments 2a and 2b revealed a similar pattern of final test performance for items in the standard multiple choice condition and the multiple choice plus retrieval condition. Because this result emerged even when overt retrieval of the responses was encouraged (Experiment 2b), it suggests that participants may already be engaging in retrieval to some extent on standard multiple choice tests. Experiment 3 was designed to assess whether retrieval occurs during multiple choice testing without direct instruction from the experimenter, and if so, whether those items that are identified as retrieved are better remembered on a final test.

Method

Participants. Fifty-eight General Psychology and Research Methods students at Colorado State University participated in this experiment. Data from four participants were excluded due to failure to follow instructions. The experiment was conducted in groups of 7-8 people and followed a within-subjects design.

Procedure. As in the previous experiments, using the same materials, participants read 18 passages presented in booklets and placed a checkmark next to the corresponding number on a sheet when they finished each passage. Participants were allowed to view each passage for 90 s, which was judged to be sufficient time to read through the passage once.

Three facts were selected from each passage, and all three facts from a passage were assigned to one of the conditions presented in the intervening test phase. Although all three facts from a particular passage were assigned to the same condition, the assignment of passages to conditions was counterbalanced such that facts from each passage appeared equally often in each condition. Some items were part of a re-study condition, presented as a statement to be read and checked off on a line provided. Other items were presented as multiple choice questions with four possible responses. The multiple-choice questions and re-study statements were randomly intermixed.

Next, participants were instructed to go back through the items a second time. For re-studied statements, participants were instructed to read the statement a second time and place a second checkmark on the line. For multiple choice questions, participants were asked to think about why they had selected a particular response option and label each question as a Type A or B memory, or a Guess (see Appendix B for the full set of instructions). This procedure is modeled after the remember/know procedure (Gardiner, 1988; Tulving, 1985) and is designed to elicit reflection from participants regarding the memory processes they engaged in at the time of the test. A Type A rating was given for a question if participants could recollect specific details from the passage at the time they were answering the question (a typical "remember" response). A Type B rating indicated that the response they chose was familiar but they could not recollect details about the fact from the passage (a typical "know" response). A Guess response meant that they had simply guessed as to the answer. The

terminology selected for the current study reflects the finding that more accurate reports of remembering and knowing are elicited through the use of neutral terms such as "Type A" or "Type B" memory than the traditional "Remember" or "Know" (McCabe & Geraci, 2009).

Following the intervening test phase, participants solved math problems for 5 min as a distracter task. Finally, a short answer test identical to the one used in Experiments 2a and 2b was administered, and participants were given up to 15 min to complete this test.

Results and Discussion

Performance on the Intervening Multiple Choice Test

The data for this experiment were analyzed in a similar fashion to the previous experiments. First, a repeated-measures ANOVA was conducted to examine the proportion of correct responses on the multiple choice test for items rated Type A (M = .92, SD = .11), Type B (M = .69, SD = .22), and Guess (M = .39, SD = .18). This analysis indicated that participants correctly answered these questions at significantly different rates [$F(2,106) = 132.70$, $MSE = 3.79$, $p < .05$, $\eta_p^2 = .72$]. One might expect this result, given that Type A answers were produced based on recollection of specific details, Type B answers were given based on a feeling of familiarity, and Guess answers were presumably accompanied by neither of these processes.

Unconditionalized Final Test Performance

In order to examine whether there was an overall testing advantage for multiple choice testing, a repeated-measures ANOVA was conducted on the

proportion of correctly answered final test questions. When all previously tested items were combined into one condition for analysis, as in the previous experiments, no testing advantage emerged for multiple choice testing (M = .46, SD= .16) when compared to re-studying (M = .45, SD = .19) [$F(1,53) = 1.69$, MSE = .02, $p > .05$, $\eta_p^2 = .03$].

A separate analysis was then conducted to assess final test performance for tested items previously rated Type A, Type B, and Guess. A repeated-measures ANOVA was conducted on the proportion of correct short answer responses for the three previously tested conditions and the re-study condition, and this analysis showed that there were significant differences in final short answer test performance between conditions [$F(3, 159) = 148.12$, MSE = 3.83, $p < .05$, $\eta_p^2 = .74$]. Post-hoc Tukey HSD tests revealed better memory for Type A questions (M = .76, SD = .18) than for all other conditions, ts(53) = 9.76, 21.73, and 13.25 for the Type B, Guess, and re-studied conditions, respectively.. Type B questions (M = .43, SD = .24) were correctly answered at a significantly lower rate than Type A questions, and at a significantly higher rate than Guess questions (M = .11, SD = .17, $t(53) = 9.70$). Performance on Type B questions was not different from that of re-studied items (M = .45, SD = .19, $t < 1$). Additionally, performance on re-studied items was significantly better than Guess items, $t(53) = 10.77$ (see Figure 5.1).

The key finding of interest is that there was a significant testing effect with multiple choice testing for the specific items for which participants reported a recollective experience (items rated Type A). There was not, however, a

significant testing advantage for those items that were not accompanied by a recollective experience.

Conditionalized Final Test Performance

Because performance on the initial test was not equivalent across conditions, and for consistency across experiments, another analysis was conducted to examine final test performance for previously tested items, conditionalized to include only those questions answered correctly during the intervening test phase. A repeated-measures ANOVA indicated that there were significant differences between conditions in the conditionalized data as well [$F(3, 159) = 70.69$, $MSE = 3.13$, $p < .05$, $\eta_p^2 = .57$]. Post-hoc Tukey HSD tests revealed that all conditions were significantly different from one another, with Type A items ($M = .82$, $SD = .16$) and Type B items ($M = .58$, $SD = .29$) but not Guess items ($M = .24$, $SD = .31$) producing a testing advantage over re-studying [$t(53) = 5.71$ (Type A vs. Type B); $t(53) = 12.73$ (Type A vs. Guess); $t(53) = 13.74$ (Type A vs. re-studied); $t(53) = 8.07$ (Type B vs. Guess); $t(53) = 3.61$ (Type B vs. re-studied); $t(53) = 4.30$ (Guess vs. re-studied)].

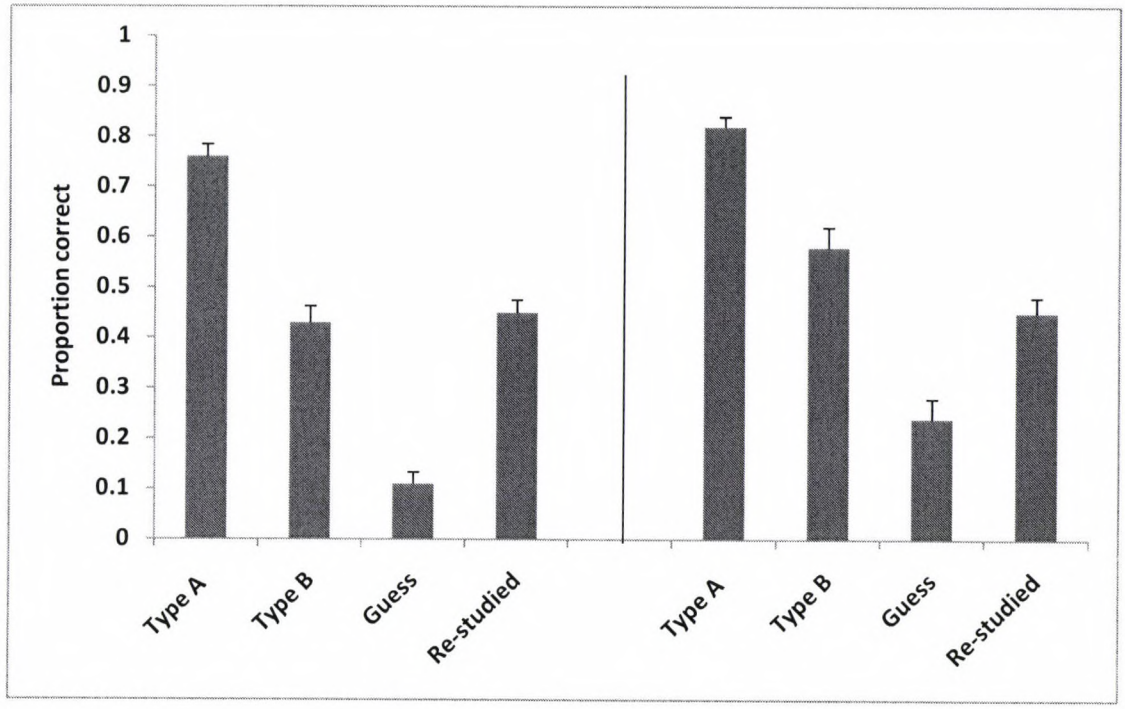

Figure 5.1: Unconditionalized final test performance (left panel) and final test performance conditionalized on intervening test performance (right panel) for Experiment 3.

CHAPTER 6

GENERAL DISCUSSION

The current study examined the testing effect for multiple choice intervening tests, drawing on ideas from the retrieval hypothesis (Carpenter & DeLosh, 2006; Cuddy & Jacoby, 1982; Dempster, 1996; Glover, 1989). Experiment 1 was a replication and extension of previous studies (Odegard & Koen, 2007; Roediger & Marsh, 2005) which compared memory for information from passages that was previously tested with multiple choice questions, for information that was previously re-studied, and for information that was not re-studied or tested. The results replicated past research in that taking a multiple choice test led to improved memory performance when compared to a no-test condition. However, the inclusion of a re-study comparison condition in Experiment 1 showed that, in the unconditionalized data, re-studying led to better later memory performance than taking a multiple choice test (although a testing advantage was revealed in the conditionalized data, when the analysis was limited to the multiple choice questions that were answered correctly in the intervening phase).

Based on the retrieval hypothesis, or the idea that it is the act of retrieving or generating information from memory that conveys a testing advantage, Experiments 2a and 2b required participants to read a question stem, try to

retrieve the answer, and then answer the multiple choice question by recording the letter of the response they selected. The thinking here was that if the retrieval hypothesis holds, encouraging participants to try to come up with the answer to the multiple choice questions prior to seeing the responses may yield an overall testing advantage for multiple choice testing relative to standard multiple choice testing and re-studying. In Experiment 2a, covert retrieval of the response did not lead to better memory than taking a standard multiple choice test, and neither testing condition improved memory over re-studying the information (although a testing advantage emerged in the conditionalized data). In Experiment 2b, overt retrieval of the response resulted in the same pattern of results, indicating that instructing participants to retrieve answers to multiple choice questions prior to considering the presented alternatives is not sufficient to produce a testing advantage over re-studying or the standard method of multiple choice testing.

Experiment 3 implemented a version of the remember/know paradigm and resulted in better later memory for items rated Type A (in which specific details were recollected at the time of test) than items rated Type B (in which participants relied on familiarity to answer the question). Type B responses did not produce significantly better memory than re-studied items in the unconditionalized data (although a testing advantage emerged in the conditionalized data). Thus, in the cases where participants retrieved specific information from the passages, final memory performance was enhanced relative to conditions that were not accompanied by recollection. This finding suggests that recollective processes are sometimes used during multiple choice testing,

and that recollected items are better remembered than items processed using familiarity.

Implications for Multiple Choice Testing

Overall, the present study suggests that it may not be possible to take advantage of the testing effect with multiple choice tests. Across four different experiments, multiple choice testing failed to produce better memory performance on a final test than re-studying. This was true even when participants were instructed to engage in retrieval prior to reviewing the response alternatives. Based on these findings, the tentative conclusion is that multiple choice tests are not effective in enhancing subsequent memory performance.

However, there are several limitations to the current experiments. One methodological concern is that in Experiment 2b participants may not have been given adequate time to think back to the passage, retrieve the answer, and write it down before the four possible responses appeared. This is a difficult task, as evidenced by the finding that only 33% of the items were correctly retrieved in that condition. Perhaps if more time was allotted, as is often the case in a classroom setting where an entire class period is scheduled for one exam, participants would be more successful in retrieving the correct answers.

In addition, perhaps the present findings are limited to situations in which there is a relatively short delay until the final test. The current experiments were each conducted in single sessions of about an hour, which naturally limits the time available for the retention interval prior to the final test. In each experiment, the final test was administered 5 min after the intervening test was completed.

Although our lab has consistently demonstrated robust testing effects in standard list learning paradigms after a 5 min retention interval (e.g., Carpenter & DeLosh, 2006), perhaps with this protocol and with these materials, a longer retention interval is needed for a testing advantage to emerge. Indeed, prior research has indicated that often a memory advantage for re-studied materials is evident at short delays, with a testing effect emerging later. Roediger and Karpicke (2006b, Experiment 1), for example, administered an intervening free recall test for one prose passage and allowed participants to re-study a second prose passage. Participants then took a final free recall test either 5 min, 2 days, or 1 week later. The results indicated that re-studied items were remembered better than tested items after 5 min, but a testing advantage was evident after delays of 2 days and 1 week. Although Roediger and Karpicke did not employ a multiple choice intervening test, they did use materials very similar to those used in the current experiments. Thus, it is possible that a testing advantage would emerge for the protocol used here with a longer retention interval.

Finally, it is also generally true that in the testing effect literature, administering more than one memory test leads to a testing advantage of a greater magnitude than giving one test (Carpenter, Pashler, Wixted, & Vul, 2008; McDaniel et al., 2007; Roediger & Karpicke, 2006b, Experiment 2). Perhaps simply testing participants once prior to the final test was not sufficient to produce a testing advantage in the current experiments, but if one were to give several multiple choice tests, a testing advantage would emerge. Along these lines, performance on the intervening tests in the current experiment was relatively low,

lower than one would typically see on a course exam. Perhaps if multiple choice test performance were at a higher level, a significant testing effect would emerge. There is evidence that high intervening test performance increases the likelihood of obtaining a testing advantage (Kuo & Hirshman, 1996).

Some investigations of the testing effect have incorporated feedback, which is one way ensure that the correct answer is processed (e.g., Butler et al., 2007; Butler & Roediger, 2008). Although the current study did not formally include feedback, in Experiments 2a and 2b, some form of feedback was available. After retrieving the answer from the question stem prompt, participants saw the four possible responses appear. If the response they had retrieved was among the alternatives, this may have provided validation that the answer they retrieved was correct. Likewise, if they retrieved a response and it was not among the four alternatives that followed, this provided an indication that their initial response was incorrect. Despite this potential feedback in the multiple choice plus retrieval condition, overall performance did not exceed that of items tested in the standard multiple choice condition.

Given the limitations outlined previously, further research should be conducted to assess whether multiple choice testing produces a memory advantage in comparison to re-studying information. One way to address potential timing issues could be to allow participants to self-regulate the time taken during the multiple choice test. Participants could be allowed to read the multiple choice question stem on a computer screen, think about the answer, and type it in when they were ready, after which the four possible responses would

appear. While forfeiting some experimental control, this method would be similar to what students actually do when they take a test in class. Alternatively, a follow-up study could be conducted in which the same basic study design as the experiments reported here was employed, but rather than a 5 min retention interval, a 24 hr break could be introduced prior to the final test. Previous research indicates that if a testing advantage is going to emerge, a 24 hr retention interval should be sufficient (e.g., Carpenter et al., 2008). Finally, future research should investigate ways to boost intervening test performance, possibly by giving more than one intervening test, in order to examine the effects on memory for tested items.

Should a testing effect emerge in future studies with multiple choice tests, it would be interesting to assess whether memory for specific items is driving this testing effect. In Experiment 3, a testing advantage emerged for items rated Type A, but not Type B or Guess. With a longer delay, one might speculate that items initially answered based on recollection of specific details (i.e., Type A) would be more resistant to forgetting than other tested items or re-studied items (see Carpenter et al., 2008 for a discussion of testing insulating against forgetting). At longer delays this advantage might manifest itself as an overall testing effect. Thus, in future studies it would be interesting to track performance over time for Type A, Type B, Guess, and re-studied items over a longer retention interval.

Implications for the Retrieval Hypothesis

Overall, the results of the present set of experiments do not provide strong support for the retrieval hypothesis of the testing effect as applied to multiple choice testing, at least for the materials and procedure used here. According to the retrieval hypothesis, one may not expect to observe a testing effect with standard multiple choice tests, since standard multiple choice tests do not require the active generation or retrieval of information from memory. This is, in fact, what was found in Experiment 1. However, Experiments 2a and 2b required participants to covertly or overtly retrieve answers from memory prior to seeing the response alternatives. To the extent that participants do, in fact, engage in active retrieval of the information, the retrieval hypothesis would predict that this retrieval would enhance subsequent memory. However, Experiments 2a and 2b did not yield an advantage for tested items over re-studied items, even in the retrieval conditions.

The present findings also question the assertion that retrieval is uncommon in standard multiple choice tests. In Experiment 2b, participants were asked to overtly retrieve and record the answer to the multiple choice questions in one of the conditions. Although later overall memory performance for information appearing in this retrieval condition did not exceed performance for items in the standard multiple choice condition, participants did correctly retrieve 33% of the items on the intervening test. Because there was no difference in later memory performance despite this fact, it may be the case that participants were

spontaneously engaging in retrieval in the standard multiple choice condition as well.

Indeed, in Experiment 3, when participants were asked to go back and rate their responses to the multiple choice questions they just completed, 40% of the responses were rated as Type A, or responses selected based on retrieval of specific details from the passage. Note that this was the case even though participants were not asked to retrieve the answers and did not know about the rating task until after they had completed the multiple choice test. Interestingly, Type A responses were better remembered on the final test than any other condition, indicating better later memory for those items for which retrieval processes were spontaneously engaged. Although this boost was not enough to lead to an overall testing advantage in the current study, it is an interesting finding and provides some support for the importance of retrieval processes in multiple choice testing. Thus, the results from Experiment 3 suggest that multiple choice exams could be an effective tool for learning in the classroom setting provided that retrieval of the information is encouraged and familiarity-based responding is discouraged.

However, based on the overall pattern of findings in the experiments reported here, more research is needed to further evaluate the retrieval hypothesis of the testing effect with multiple choice tests. Overall, the data from the current study show that a testing advantage does not emerge when multiple choice testing is compared to re-studying, and this is true even with the addition of an induced retrieval activity (Experiments 2a and 2b). Further experimentation

on multiple choice testing is warranted, however, particularly studies that include a manipulation of variables that have produced robust testing effects in past work (i.e., studies with longer retention intervals, the addition of multiple intervening tests, etc.). The preliminary conclusion based on the current data, however, is that instructors should not rely on multiple choice tests if they are interested in reaping the benefits of testing as a tool to enhance memory.

REFERENCES

Butler, A.C., Karpicke, J.D., & Roediger, H.L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13,* 273-281.

Butler, A.C. & Roediger, H.L., III. (2008) Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory and Cognition, 36,* 604-616.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268-276.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition, 36,* 438-448.

Chan, J. C. K. & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 33,* 431-437.

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

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior, 21,* 451-467.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork and R. A. Bjork (Eds.), *Handbook of perception & cognition: Memory* (pp. 317-344). San Diego, CA: Academic Press.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16,* 309-313.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392-399.

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

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109,* 451-464.

McCabe, D. P. & Geraci, L. D. (2009). The influence of instructions and terminology on the accuracy of remember-know judgments. *Consciousness & Cognition, 18,* 401-413.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494-513.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371-385.

Odegard, T.N. & Koen, J.D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory, 15,* 873-885.

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

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

Roediger, H. L., III & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 1155-1159.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26,* 1-12.

Yonelinas, A.P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

APPENDIX A

Mean proportion correct on the final short answer tests for all experiments: Unconditionalized and conditionalized on correct intervening test performance. (Standard deviations in parentheses, MC+R = Multiple Choice plus Retrieval)

	Unconditionalized	**Conditionalized**
Experiment 1		
Multiple Choice	.49 (.17)	.72 (.21)
Re-studied	.56 (.19)	-
No Test	.30 (.14)	-
Experiment 2a		
Multiple Choice	.52 (.19)	.68 (.17)
MC+R	.51 (.16)	.68 (.17)
Re-studied	.55 (.19)	-
Experiment 2b		
Multiple Choice	.49 (.15)	.69 (.16)
MC+R (letter)	.49 (.14)	.70 (.14)
Re-studied	.51 (18)	-
MC + R (retrieved)	-	.93 (.11)
Experiment 3		
Type A	.76 (.18)	.82 (.16)
Type B	.43 (.24)	.58 (.29)
Guess	.11 (.17)	.24 (.31)
Re-studied	.45 (.19)	-

APPENDIX B

Experiment 3: Multiple choice response rating instructions

Now I'd like you to look back over the multiple choice questions you just answered and think about why you selected the response you chose for each particular question. In each case, you should write down next to the question whether it was a Type A memory, a Type B memory, or a guess.

Type A

You should select a Type A response if you remembered specific details from one of the passages while you were answering the question. In this case you might have had specific images or feelings in mind relating to the information you recalled as you were answering the question. For example, if you give a Type A response, perhaps you virtually "saw" again or had a specific memory of the passage you read while answering the question. Maybe you remembered what you were thinking about at the time that you read that particular fact in the passage, or you thought of a particular association you made between that fact and something else. A Type A response might also be given because when you were answering the multiple choice question you remembered a personal association you made when you first read that fact in the passage. Or, you might have thought about where that fact appeared in the passage as you were answering the question. In these cases where you had a vivid or conscious recollection of an answer being in the passage at the time you made your response on the multiple choice question, you should write down "Type A" next to that question.

Type B

You should select a Type B response if, while you were answering the question, you did not specifically recollect details from when you were reading the passage. For these questions, you knew the answer was in the passage, and it seemed more familiar than any of the other responses, but you didn't have a vivid or conscious recollection of actually reading it in the passage. You should also select Type B if you just knew the answer to the question but didn't recollect it from the passage specifically. In these cases you should write down "Type B" next to the multiple choice question.

Guess

It is also possible that you may not have remembered specific details of reading the answer in the passage, you might not have known the answer, or even felt the answer you selected was familiar. In this case you may have made a guess.

For example, some of the answer choices may have looked unlikely for some reason so you selected the one that seemed most likely to be right. If you truly guessed as to the answer, write "Guess" next to the multiple choice question.