THE TESTING EFFECT AND THE ITEM SPECIFIC VS. RELATIONAL ACCOUNT

Daniel John Peterson

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology.

Chapel Hill
2011

Approved by:
Kelly Giovanello
Joe Hopfinger
Peter Gordon
Neil Mulligan
Peter Ornstein
ABSTRACT

Daniel Peterson: The Testing Effect and the Item-Specific vs. Relational Account
(Under the direction of Neil Mulligan)

The testing effect refers to the phenomenon whereby being tested for some previously studied material leads to levels of recall above and beyond a simple re-presentation of the material. Research over the past 20 years has focused intently on the factors surrounding the memory improvements associated with tests with very little consideration for the underlying mechanism. Although several researchers have posited theories to explain why tests enhance memory, few accounts have been empirically tested. This paper reviews prior literature on the testing effect and considers whether it may be appropriate to conceptualize it within the item-specific vs. relational framework, an account used to explain a variety of memory phenomena including the generation effect. In the experiments presented here, the generation effect and the testing effect appear to be qualitatively similar memory phenomena as a design which yielded a negative effect of generation yielded a comparable negative effect of testing. These results suggest that generation and testing may be understood within a common framework offering one of the first empirically supported accounts to explain why tests (typically) improve memory.
# TABLE OF CONTENTS

LIST OF FIGURES ................................................................. v

Chapter

I. INTRODUCTION .............................................................. 1
The testing effect .............................................................. 2
Educational applications: In the laboratory .............................. 8
Educational applications: In the classroom .............................. 12
Explanations of the testing effect .......................................... 18
Explanations of the testing effect: The item-specific vs. relational account .. 25
The current study ............................................................... 34

II. EXPERIMENT 1 ............................................................... 40
Method ................................................................. 41
Results & Discussion .......................................................... 44

III. EXPERIMENT 2 .............................................................. 46
Method ................................................................. 46
Results ............................................................... 47
Discussion ............................................................. 48

IV. EXPERIMENT 3 ............................................................... 50
Method ................................................................. 52
Results & Discussion .......................................................... 53
LIST OF FIGURES

Figure

1. Data from Roediger and Karpicke (2006b) ........................................77
2. Data from Kang, McDermott, and Roediger (2007) .......................78
5. Data from McDaniel, Anderson, Derbish and Morrisette (2007) ....81
6. Data from Szpunar, McDermott, and Roediger (2008) ..................82
7. Data from Karpicke and Zaromb (2010) .......................................83
8. Data from Karpicke and Zaromb (2010) .......................................84
9. Data from Karpicke and Zaromb (2010) .......................................85
10. Data from Karpicke and Zaromb (2010) ....................................86
11. Data from Steffens (1999) .........................................................87
12. Data from Steffens (2007) .........................................................88
13. Mean proportion of targets recalled: Experiments 1-4 ..................89
14. Proportion of participants aware of the categorical relationship amongst the targets: Experiments 1-4 .................................90
15. Mean Adjusted Ratio of Clustering (ARC) scores: Experiments 1-4 ...91
16. Distribution of ARC scores: Experiments 1-4 ............................92
CHAPTER I

INTRODUCTION

Although tests are most often thought of as a means of measuring recall or learning, research suggests that testing can actually alter the memory trace it is intended to assess. The testing effect refers to the phenomenon whereby being tested on previously studied material enhances recall above and beyond the simple re-presentation of the material.

Though adoption of the term ‘testing effect’ is relatively new, the notion that testing can improve recall is not. Jones (1923-1924) was interested in the effect of testing on later retrieval in a classroom setting. In his study students listened to a one hour lecture of material consistent with what they might be presented in an actual class. Immediately afterwards, students in the testing condition were given a test of fill-in-the-blank and short answer questions on the content of the lecture. Control participants were present for the same lecture, though were not given any test. Following a variable delay (between 3 and 50 days later) students in both groups were given a final recall test to measure their recollection of facts from the original lecture. The data, collected from 600 students across 27 lectures revealed that recalling the lectures through intervening tests improved retention on a final recall measure. Interestingly, the benefit increased as the delay between original learning and final testing increased (more on this later).

Spitzer (1939) was interested not only in how testing affects recall in general, but also in the effect it had on forgetting initially recalled information. In his study he tested the entire population of sixth-graders in the Iowa City schools- 3,605 students. The students
studied one of two articles constructed for the purposes of the experiment: one on peanuts, another on bamboo. The articles were created with the intent to provide the children with new and factual material that would be similar to the sorts of texts they would come across in their school work. Spitzer divided the large sample size into several groups which were given either zero, one, two, or three intervening tests (given at variable schedules) before one final test which was given to all participants 63 days following the original reading of the article. He found that testing led to an average of 20% greater recall of facts relative to no test controls. In a result that would later be replicated (e.g. Roediger and Karpicke, 2006b; Karpicke & Roediger, 2010) the more intervening tests students took, the better recall performance was during the final test. Additionally the results indicated that with each additional intervening test, forgetting (defined as recalling a detail during one of the intervening tests but subsequently failing to recall it again during the final test) decreased such that with sufficient testing, forgetting was all but eliminated.

*The testing effect*

Though several other early researchers were interested in the effects of testing on memory, these two studies are highlighted because they provide prototypical examples of the types of research being done prior to the late 1960’s. Though these studies seem to clearly indicate the advantages of testing, these researchers failed to account for one important confound. In the experimental conditions (i.e. the participants who received intervening tests), participants had the added benefit of a re-presentation of the studied material compared to control participants. In other words, it could be that previous testing facilitated later recall (as these authors contended), or it could simply be that additional exposure to the
material (in the case of Spitzer, 1939, sometimes as many as three times over) led to the effect.

Carrier & Pashler’s (1992) study provides a good example of how modern research on the testing effect avoids this confound. In two experiments, participants were given a paired-associates task in which they had to learn a list of words paired either with a random digit (e.g. 9-wing) or its Eskimo cognate (e.g. yaquq-wing). All pairs were presented on a monitor one at a time for 10 seconds. Following the initial study period participants were presented with the paired associates a second time in one of two ways. In the restudy condition, participants were presented with the intact pairs in the same manner as before during the initial study phase. Conversely, in the testing condition, participants were presented with the first word or digit of the pair for five seconds in isolation (e.g. yaquq-_____), followed by the complete pair for an additional five seconds to provide feedback. In both conditions (manipulated within subjects) participants were instructed to say the response term aloud as quickly as possible. In the restudy condition, this simply meant reading the information presented on the screen, whereas in the testing condition participants had to try to retrieve the information from memory before the feedback was provided. Following a five-minute distracter, participants received a cued recall test on half of the items they initially studied. 24 hours later, the participants returned and were given a second cued recall test on the other half of the items which were not previously tested.

Critically, if learning only occurs during the study of material, then restudy items should show better cued recall performance relative to test items given the intact pairs were presented for 33% longer (20 seconds total in restudy condition vs. 15 seconds total in the test condition). The results from both experiments instead revealed better recall for the test
items, suggesting that testing facilitates memory performance. This boost in performance was found both five minutes following study and 24 hours later (the interaction between testing condition and delay was not significant). These results are important because the benefit of testing here cannot alternatively be explained by additional exposure to the studied material. Participants saw the restudy items intact for a longer duration, yet recalled fewer of these items relative to test items.

Perhaps an even more compelling demonstration of the testing effect comes from Roediger & Karpicke (2006b). In their study, participants were presented with prose passages to study for later recall. In Experiment 1 participants took an immediate free recall test without feedback following the initial reading (the study-test condition, or ST) or restudied the material via an additional reading of the passage (the study-study condition, or SS). A final recall test was administered to all the participants five minutes, two days or 1 week later. Figure 1 shows the results from this first experiment. After five minutes, there was a slight, though significant advantage for passages in the SS condition. However, at both the two day and one week retention intervals, there was a sizeable reversal, evidenced by a large ST advantage.

In second experiment the authors were interested in not only replicating the results from the Experiment 1 but also in determining whether additional tests lead to a larger testing effect (e.g. Spitzer, 1939). The design was similar to the first experiment though this time there were only two retention intervals (five minutes and one week) and three testing conditions, borrowed from Tulving’s (1967) pioneering work on test-enhanced learning: participants who studied the passage four times (SSSS), participants who studied the passage three times with one immediate free recall test without feedback (SSST), and participants
who studied the passage once and then immediately performed three free recall tests without feedback (STTT). The results from Experiment 1 were replicated, while additionally revealing that after one week, participants in STTT condition recalled slightly more details than those in the SSST condition replicating Spitzer’s (1939) findings that more tests leads to better retention (see also Karpicke & Roediger, 2010).

These results are significant for two reasons. Most importantly, this study demonstrates that the testing effect is driven by retrieval processes alone, and is not simply an artifact of feedback (see also Kuo & Hirshman, 1996; Toppino & Cohen, 2009). To grasp this point, it is helpful to compare these results with those of Carrier & Pashler (1992). In examining these two studies, an important question to consider is why after five minutes Roediger & Karpicke (2006b) found a significant advantage for the study only controls, while Carrier & Pashler found a significant advantage for the test items (i.e. a testing effect)? The answer is likely due to the presence (or absence) of feedback during the intervening tests. In Carrier & Pashler’s study when presented with test items, participants had to think back to the original study session to retrieve the paired associate. After five seconds the intact pair was again presented which served to reinforce accurate recall, or, in the case of a retrieval failure, served as an additional encoding opportunity. In this sense it is not so surprising that a significant testing effect was demonstrated only five minutes following the intervening test- after all participants here benefit both from the act of retrieval as well as an additional encoding period should they need it. These data alone leave open the question as to whether the testing effect is driven solely by the act of retrieval, or whether it is simply an artifact of corrective feedback. By demonstrating a significant and sizeable testing effect without giving participants any feedback, Roediger and Karpicke help solidify the notion that
retrieval processes alone drive the effect. This result gets at the core of the testing effect and is quite remarkable: participants in the STTT condition were only presented with a story one time, yet after one week recall 20% more information than participants who read the same story four times.

An additional reason the findings of Roediger & Karpicke (2006b) are so important is because by demonstrating an interaction between testing condition (recalling the story vs. rereading the story) and retention interval, there is concrete evidence that the underlying processes of testing are fundamentally different than those of (re)studying. Several early studies examining the testing effect demonstrated a main effect of testing such that testing was shown to improve retention at both short and long intervals (e.g. Carrier & Pashler, 1992). Without a functional dissociation of studying and testing, it is impossible rule out the idea that testing is simply more effective at inducing the same processes that are activated during study. If this were the case, though testing would still be a topic worth investigating, it would be a considerably less interesting phenomenon. However, by demonstrating an advantage for study items after five minutes, and an advantage for tested items after 48 hrs (a significant crossover interaction), it is clear that the processes underlying studying and testing are qualitatively different (for explanations of this interaction, see Explanations of the testing effect).

Prior to the early 1990’s research on the testing effect proper was relatively scarce, though the past twenty years has seen an explosion of studies aimed at examining the extent to which testing can improve recall. The testing effect has since been demonstrated with

---

1 Toppino & Cohen (2009) conducted a study which was effectively a replication of Carrier & Pashler (1992) without giving participants feedback in the testing condition and found results consistent with this explanation offered here. After a 2-min interval there was a small (nonsignificant) benefit for the restudy material, though after a 48 hr delay there was a significant advantage for the tested material, mirroring the results of Roediger & Karpicke (2006b).
word lists (e.g. McDaniel & Masson, 1985), paired associates lists, (e.g. Carrier & Pashler, 1992; Carpenter & DeLosh, 2005), multiple-choice tests (e.g. Butler, Karpicke, & Roediger, 2007; Butler & Roediger, 2008) and prose material (e.g. Roediger & Karpicke, 2006b).

Importantly, however, the effect is not limited to verbal material. While Carpenter & DeLosh (2005) demonstrated a testing effect in a face-name paired associates tasks, Carpenter & Pashler (2007) were interested in whether testing improved memory for exclusively non-verbal materials. In their study, participants were presented with two maps (across two different sessions) on a computer screen, one of which was presented with instructions for conventional studying. For the other map, participants were first presented with the intact map consisting of several features (e.g. roads, rivers, mountains, etc.). After 20 seconds, participants were shown a different image: the map they had studied, though missing one feature. The participants were instructed to covertly guess which item was missing. A series of these one-feature-absent-maps were presented such that the total time studying the map was equal to the other (control) map. Following a 30-minute distracter participants were instructed to draw both maps from memory. The results showed that participants’ map drawings were significantly better (i.e. accurately included more original features in the appropriate location) for maps learned in the testing condition relative to control.

Similarly, Johnson & Mayer (2009) demonstrated a testing effect with multimedia material. In their experiment, participants first viewed a short animated film about lightning. In the study condition, participants watched the film again, while in the testing condition, participants were instructed to provide a written explanation of how lightning works. Following a week delay, participants in the testing condition recalled more facts from the
film than the restudy condition. Taken together, these two studies are important for demonstrating that the testing effect is not a phenomenon limited to verbal material, but rather is generalizeable.

*Educational applications: In the laboratory*

Because of the obvious educational applications, many studies investigating the testing effect have used educationally relevant materials. Two of the studies reviewed previously specifically constructed their stimuli with ecological validity in mind. Carrier and Pashler (1992) employed an English-Eskimo paired associates task mimicking vocabulary learning in a second language, while Roediger and Karpicke (2006b) used prose passages intended to simulate the reading and retention of texts in the classroom.

Over the years, researchers have gone to greater lengths to create more ecologically valid sets of materials. Butler & Roediger (2007) were interested in how testing could improve memory for lectures. In their study, participants watched a series of art history lectures on video. Over the course of three days, students saw three different 30-minute videos, each focusing on the work of one particular artist. Immediately following the presentation of each lecture, students were given a short-answer test, a multiple-choice test, or a focused restudy of the facts. For each video 30 facts were extracted such that 20 were tested (in one of the three conditions previously listed) and 10 were not (for a baseline comparison). In an attempt to simulate an average delay a student might see between the presentation of material and a test, the participants came back to the lab 30 days later for a final recall test. The results indicated that being given an intervening test led to improved retention compared to a focused restudy of the facts, an extension of previous findings to this more ecologically valid set of materials.
While many studies have strived to identify appropriate study materials for educational applications, the type of test used to examine the testing effect is also of interest. Universally, research has indicated that intervening tests of short answer or free recall result in the largest testing effect, regardless of the final retention test format (Glover, 1989; Kang, McDermott, & Roediger, 2007; McDaniel, Anderson, Derbish, & Morissette, 2007; Butler & Roediger, 2007).

Kang et al. (2007) provide a good example of a study that explicitly compared test formats. In their study (Experiment 2) participants read a series of articles from *Current Directions in Psychological Science* meant to provide participants with a set of materials similar to what may be read for an actual college course. Following each article, participants took an initial short answer test, took a multiple-choice test, or reread key statements from the text. Critically, the information sought in each of the question conditions was the same as that given in the reread condition. For example, a short answer question may have been, “*What is source confusion?*” The corresponding multiple-choice condition would have asked the same question while additionally providing four responses. In the read condition, participants would read, “*Source confusion occurs when one misattributes the content of a memory to the wrong source.*” In this way the same critical information is being tapped, though in a different manner for each condition. In both the short answer and multiple-choice conditions participants were given feedback on their responses. During a final retention measure (administered three days later) participants were given a final test on all of the articles read. Some of the questions were short answer while others were multiple-choice. With this design, some of the questions on the final test were identical to those answered previously (e.g. “*What is source confusion?*”). In other cases there was a
mismatch (in this case, being presented with the multiple-choice question on the same concept). The results (see Figure 2) reveal a benefit for short answer as an intervening test regardless of the final retention measure. These data support the notion that short answer tests (along with feedback) provide the most efficient means of maximizing recall for a given set of materials.

Further evidence that short answer tests are preferred over multiple-choice tests can be found in one of the negative effects associated with testing. While the effects of testing reviewed thus far are universally positive, the use of multiple-choice tests can unfortunately lead to negative suggestion effects. To give a simple example, consider a student taking a multiple-choice test who comes to a question of which he is unsure of the answer. He mulls over the potential responses until reasoning that choice (a) is the best response. Having convinced himself of the rationality of his choice, he is again likely to select this response if presented the same question at a later point (e.g. a final exam), the basic testing effect.

Problems arise, however, when choice (a) was actually the incorrect response to begin with\(^2\). This problem is a significant one because instructors strive to create plausible lures for multiple-choice tests to prevent students from simply reasoning through the correct response.

While research has empirically demonstrated that this negative effect of multiple-choice testing is a very real phenomenon (e.g. Roediger & Marsh 2005; Odegard & Koen, 2007; Butler & Roediger, 2008; Marsh, Agarwal, & Roediger, 2009), it also suggests that the use of corrective feedback can serve to diminish if not entirely eliminate it altogether. This

\(\text{\footnotesize\textsuperscript{2}}\) As Odegard & Koen (2007) show, this issue is especially detrimental when a “none of the above” option is included in the multiple choice question. That is because when this particular response is the correct one, answering the question correctly forces a student to only consider incorrect responses, and ultimately commit to a response alternative that does not factually complete the question. Their study revealed that including “none of the above” as the correct response served only to increase the already problematic negative effects of testing seen with multiple choice tests.
point is important because although the exclusive use of short answer tests may be ideal, the amount of grading required (especially as class size increases) may not always be practical. In their study Butler & Roediger (2008) had participants read various encyclopedia articles. Following a distracter participants were given a multiple-choice test with either no feedback, immediate feedback, or delayed feedback. Relative to no feedback both immediate and delayed feedback yielded greater retention during a follow up multiple-choice test.

An intuitive explanation for these results suggests that after selecting a wrong answer, receiving feedback helps correct the mistake for a future test. However, Butler, Roediger, & Karpicke (2008) demonstrate that feedback can also be important for correct responses. In a similarly designed study to Butler & Roediger (2008), participants gave confidence ratings following their initial multiple-choice responses. The data from a final cued recall test are provided in Figure 3. As one would expect, after originally selecting an incorrect response during the initial multiple-choice test, feedback served to greatly improve accuracy during the final recall. Interestingly, however feedback also had a significant impact on originally correct responses. Specifically, feedback served to reinforce originally correct responses which were rated as “guesses” or “low confidence” to improve final recall accuracy by as much as 40%. This research suggests that while it may be tempting to go back through a test only to focus on the questions a student got wrong, it is equally important to go back through the questions they got right to reinforce items for which they were only guessing, or unsure of their response.

A different problem associated with testing is that the endorsement of a frequent schedule of testing in the classroom takes up valuable class time and forces the instructor to spend considerably more time grading. As such, McDaniel, Howard, & Einstein (2009) set
out to determine if the benefits of testing could be packaged in a way such that students could easily benefit from them during their own study time. While basic vocabulary or fact learning is straightforward enough that students can easily test themselves on their own time (e.g. through the use of flashcards) the issue becomes more complicated as the information to be learned becomes more complex (e.g. understanding and retaining a Shakespearean play).

To address this, McDaniel et al. had students engage in a read-recite-review strategy requiring the students to read a passage, subsequently attempt to recall all of the details of that passage, and reread the passage once more to review. Students who engaged in the read-recite-review strategy recalled significantly more information from educational texts than both students who just read the passage twice, and students who read the passage twice while taking notes. This was true of both immediate and delayed tests. The authors concluded that the read-recite-review strategy can help students capitalize on the benefits of testing while not sacrificing class time to actually administer the tests.

*Educational applications: In the classroom*

Though most often demonstrated in the laboratory, the testing effect has garnered so much attention, in part, because of the practical applications in the classroom. One then may wonder how well the positive results of testing extend to an actual academic course. Though it may seem obvious that the benefits of testing should extend to the classroom, Roediger & Karpicke (2006a) point out several reasons why this is not a foregone conclusion. First and foremost the quantity of material for which a student is responsible is usually much greater than that tested in the lab. Though researchers have demonstrated positive effects of testing for texts, the amount of reading required in these studies pales in comparison with the amount of information a student in an introductory history or science course must master.
Another difference can be found in the delay between initial presentation and final test. In the lab participants are often tested a day or at most several days later, while a final exam in a class requires the student to be responsible for material presented several months prior. Also in an actual course (unlike an experiment), the material is presented in a variety of ways. For any given topic, a student will likely hear a lecture, read a textbook, and may additionally talk in discussion sections, or view relevant videos. Finally, while experiments explicitly control the amount of testing or restudying of the originally presented material, in the classroom, students will vary in terms of how much they study before a test. While an experimental design attempts to control all of these variables, in the naturalistic classroom setting, all of these factors are free to vary, raising the question of whether the positive effects of testing extend beyond the tightly controlled laboratory.

In a meta-analysis of studies dating back to 1923, Bangert-Drowns, Kulik, & Kulik (1991) examined the effect sizes of retention and academic performance across a series of classroom studies. In all, 35 studies were compiled (22 published, 13 unpublished) which systematically compared students taking relatively fewer exams compared with those taking more exams. The authors concluded that testing did increase performance on retention, though at a diminishing rate of return (see Figure 4). Further, the authors go on to state that most of the benefits of testing are marginal after factoring out the enormous benefits of taking just one test during the course of a class (relative to no test at all). While substantiating the general claim that testing can improve retention, the lukewarm endorsement of a frequent testing schedule led Roediger and Karpicke (2006a) to point out some of the shortcomings of the review. These authors point out that Bangert–Drowns et al. fail to analyze possible differences between the tests, nor do they include any information
about whether or not feedback was provided after testing. Perhaps most importantly, in 29 of the 35 studies there was no random assignment to testing conditions. Nevertheless, these results are important for demonstrating that the testing effect does extend to the classroom.

Recently, McDaniel et al., (2007) performed a classroom study which controlled for these aforementioned variables. In a web-based “Brain and Behavior” course, students at the University of New Mexico were given weekly reading assignments from the course textbook. For a given paragraph within the readings two facts were extracted which would later be tested. For one of those facts, students would be re-exposed to the information via testing (multiple-choice or short answer) or rereading. In this way final retention could vary by way of rereading, multiple-choice, short answer, or no re-presentation (dubbed no activity). The type of re-exposure was manipulated within participants such that across three weeks of material (each week corresponding to one over-arching topic) each student received a 10-item multiple-choice quiz in one of the weeks, a 10-item short answer quiz in another week, and a re-presentation of 10 facts to read in the third week (the type of exposure for any particular week’s facts were counterbalanced across students). Importantly, as in Kang et al. (2007), though the manner of re-exposure varied, the same information was tapped in each condition.

For example, for the week dedicated to neurons, a target fact was, “All preganglionic axons, whether sympathetic or parasympathetic release acetylcholine as a neurotransmitter.” In the short answer (SA) condition, participants were given the question “All preganglionic axons, whether sympathetic or parasympathetic release __________ as a neurotransmitter,” while in the multiple-choice condition (MC) they were given the same question along with four possible alternatives. In the read condition students simply read the
original fact again. For both testing conditions, feedback was given following the quiz. All of these conditions were compared against the control condition which consisted of related sentences from the same paragraph, though not re-presented in any form. For the example above the fact, “Parasympathetic postganglionic axons release acetylcholine as a neurotransmitter,” served as the no activity control. The quizzes were administered online and students were free to take them at the time of their choosing.

After three weeks of quizzes a unit examination was given on the 30 previously quizzed and read items as well as the 30 items which were not re-presented. Though the intervening tests varied the manner of re-presentation for each item, the unit exam consisted exclusively of multiple-choice questions. The data from the study are presented in Figure 5. As can be seen, the pattern established for laboratory studies holds: intervening tests improve final retention. The testing effect was evident with both short answer and multiple-choice items. Perhaps as interesting was the comparison of read items vs. no activity. The proportion correct for these items was identical suggesting that rereading the material (thought to be a tried and true method of studying) lead to no significant improvements above and beyond initial reading3. One additional point merits mention: despite the fact that the unit exam was multiple-choice, being tested via short answer lead to the best performance, replicating previous findings (Glover, 1989; Kang et al., 2007; Butler & Roediger, 2007).

The fact that the course was implemented online allowed for greater control with which to systematically examine the effects of testing. One may wonder, however, if the patterns hold in more traditional classroom settings. Leeming (2002) implemented an exam-

---

3 It is important to note, most lab studies have found this not to be the case, (e.g. Butler & Roediger, 2007) though follow up studies from McDaniel’s own research group (Callender & McDaniel, 2009) support the notion that rereading does not significantly improve recall.
a-day procedure with his “Learning and Memory” and “Introductory Psychology” courses, and while the strict experimental control seen with McDaniel at al., (2007) was noticeably absent, he nevertheless replicated the basic results in a traditional lecture-based course. Students in the summer-school courses met five days per week with each class lasting 100 minutes. Because his intent was to compare his students’ grades using this novel procedure against the grades of students from sections he had taught previously, Leeming made every effort to equate the classes, requiring him to maintain the same style and class format, use the same study questions, the same syllabus, and the same grading procedure and scale as he had implemented before.

The only difference from prior sections was that students in the exam-a-day section were given a short test during the first 10 to 15 minutes of each and every class rather than only four times per course as he had done before. Each test consisted of two short essay questions taken from the pool of study questions on the syllabus, and five short answer questions taken from the text. To provide feedback, after the exam he spent a few minutes discussing the correct answers, then went about teaching the material for that day.

As a result of the frequent testing, the final grades of the “Learning and Memory” students in the exam-a-day procedure were 8% higher\textsuperscript{4} than sections of the same class he had previously taught with only four exams. As one would expect the proportion of A’s and B’s increased, but perhaps more interesting, the proportion of D’s and F’s dropped from 21% to 2%.

Additionally, Leeming compared students from his own “Introductory Psychology” course with those from two other sections being concurrently taught with different instructors

\textsuperscript{4} Exam-a-day students’ final grades: M=89%, SD=8.13%, Traditional structure students’ final grades: M=81%, SD=17.61%
via a final retention test. Importantly, these students used the same textbook and covered the same chapters in the same order. Leeming constructed the final retention test as fairly as possible by randomly selecting short answer, short essay, multiple-choice and fill-in-the-blank questions from class exams previously administered in each of the three different sections. The results indicated that students from the exam-a-day section scored 6% higher than students from the other two sections\(^5\).

Taken together these results indicate testing not only improves student performance through the course of a semester, but additionally improves long term retention. In his discussion Leeming advocates for frequent testing despite the fact that his exam-a-day procedure cut into class time and necessitated a significant amount of additional grading time compared to his previous sections.

These results, however, are not without their limitations. Most egregiously, in the final grades comparison condition, Leeming is not blind to condition as he grades his students. There is, in all likelihood, a degree of experimenter bias in grading his exam-a-day students given he expects the manipulation to have a positive effect. Though these results alone would not be as compelling, when paired with the much more tightly controlled McDaniel et al., (2007) study, these complementary results indicate the positive effects of testing can be demonstrated both in the traditional and online classroom.

*Explanations of the testing effect*

The research reviewed here shows that testing can improve long term retention, but how? Though earlier studies were not able to rule out the role of additional exposure of the to-be-tested material, recent research has controlled for this confound and has isolated

\(^5\) Exam-a-day students’ final retention score: M=54%, SD=13.26%, Other sections final retention score: M=48%, SD=13.34%
retrieval components which drive the memorial benefits seen during final recall. However, it is not immediately clear exactly how retrieval improves retention. Traditionally, research has pointed to three factors: elaborative rehearsal, transfer appropriate processing, and release from proactive interference.

The notion that elaborative rehearsal improves memory is a relatively well-established phenomenon (e.g. Craik & Tulving, 1975). The elaboration of a memory trace multiplies the available retrieval routes facilitating recall at a later period because a greater number of retrieval cues can cue the trace (Bjork, 1988). Testing is thought to induce elaborative rehearsal through difficulty. In other words, by forcing the participant to retrieve an item from memory, the participant must engage in more elaborate processing to retrieve that item relative to the processing it would take to simply reread the item. To illustrate this point, Auble & Franks (1978) gave participants a series of sentences which were initially incomprehensible, (e.g. The notes were sour because the seam split.) It was not until being given the disambiguating cue (bagpipe) that the sentences made sense. Critically, this cue was given five seconds before, immediately after, or five seconds after the target sentence. The results indicated that the longer the lag of the disambiguating cue following the target sentence, the better the sentence was recalled at test. That is, the longer participants spent trying to figure out the sentence (making more of an effort towards comprehension) the better memory was for that sentence.

More recently Carpenter & DeLosh (2006) have reported an analogous finding with regard to the testing effect. In their study (Experiment 3), participants first studied a series of word lists presented on the computer screen. Participants saw four lists containing eight words each. Following the presentation of each list and a brief distracter, participants were
either presented with the list again, or were given a modified cued recall test. In this task, the participants were given the first one to four letters of one of the previously presented eight words (each list was constructed such that within a list each word began with a different letter. In this way the presentation of a single letter during cued-recall still constrained correct responses to only one item) and were instructed to recall the correct word. For example, for the word cabin, a participant could see a cue as difficult as c_ _ _ _, or as easy as c a b i _. After all four lists had been administered, and following a distracter, a final memory test was given in which participants were instructed to write down as many of the words as they could recall from any of the four lists. The results revealed a main effect of testing (i.e. cued recall performance greater than rereading) but more importantly, indicated that the fewer letters provided during the intervening cued recall test the better final recall was for that item. The authors argued that fewer letters during cued recall resulted in a more difficult test, increasing elaborative processing, which facilitated recall during the final test.

Pyc & Rawson (2009) offer further support for this theory while operationally defining difficulty in a different manner. In this study, participants were instructed to learn Swahili-English word pairs. After an initial presentation of each of the intact pairs, participants were given a cued recall test in which they were presented with a Swahili word and were instructed to provide the English cognate. Each cue was presented multiple times requiring the participant to retrieve each English target multiple times. Critically, the authors manipulated difficulty in two ways: through increasing the interstimulus interval (ISI, the number of trials between successive presentation of the same Swahili cue), and by manipulating criterion level (number of times the English target had to be successfully retrieved before it was dropped from further testing). As ISI increases, difficulty increases,
as did performance on a final cued recall test. Criterion, conversely, is slightly more complicated. As criterion increases, difficulty decreases because recalling an item more often makes it easier to recall at a later point. This predicts that as the number of correct target retrievals increases, the incremental benefit on a final test performance will decrease (the same function which Bangert-Drowns et al., 1991 identified with their meta-analysis), exactly what the authors found.

Interestingly, even failure to retrieve can be understood as desirable difficulty with respect to the testing effect. Kornell, Hays, & Bjork (2009) provided participants with a series of questions to learn. Half of the items were real questions with real answers (e.g. *What is the only word the raven says in Edgar Allen Poe’s poem The Raven?*) whereas the other half were fictitious questions (e.g. *What was the last name of the person who panicked America with his book Plague of Fear?). For half of the items, participants were presented with the question along with the answer, for the other half participants were presented with the question in isolation for eight seconds, during which the participant was to provide a response, followed by the answer for an additional 5 seconds (feedback). The authors were interested in what effect trying to retrieve an answer would have on questions for which there was no correct answer. Surprisingly, there was a testing effect even for these fictitious items. Trying to recall the answer to a fictitious trivia question during the study phase enhanced the encoding that took place when its answer was presented during feedback. Taken together, these studies offer considerable support for the notion that elaborative rehearsal via increased difficulty is a significant explanatory factor with respect to the testing effect.

Though some have suggested that effortful retrieval is the primary factor in explaining the testing effect (e.g. Roediger and Karpicke, 2006a) transfer appropriate
processing (Morris, Bransford and Franks, 1977) is also thought to play a significant role. The basic tenant of TAP posits that the degree to which processes at encoding are again invoked at retrieval, the greater memory will be enhanced. The notion is similar to Tulving & Thomson’s (1973) encoding specificity theory, and on the face is a very reasonable explanation for the testing effect. By giving intervening tests you are giving relevant testing practice for the information which will again be tested later.

Interestingly, this notion is advocated even though in their most stringent interpretation, TAP predictions have been repeatedly refuted (Roediger & Karpicke, 2006a; Carpenter & DeLosh, 2006). As reviewed earlier, intervening tests of short answer have repeatedly yielded the largest testing effect regardless of the final test used (Glover, 1989; Kang et al., 2007; Butler & Roediger, 2007; McDaniel et al., 2007). According to TAP this should not be the case. If the final test is to be multiple-choice, TAP would suggest the most appropriate intervening test (that is the intervening test which should lead to the largest testing effect) would also be multiple-choice. However, research has repeatedly failed to support this idea.

Carpenter and DeLosh (2006, Experiment 1) demonstrate this phenomenon explicitly. In this experiment participants were presented with 16 eight-word lists in a similar fashion to the previously reviewed Experiment 3. Following the presentation of each list, participants were given one of three intervening tests: recognition, cued recall, or free recall. After the presentation of all 16 lists and following a brief distracter, participants were given a recognition test for all 96 items, a cued recall test (given the first letter of each word) for all 96 items, or a free recall test. The results were consistent with previous findings: free recall as an intervening test led to the highest levels of performance regardless of the final test type.
Contrary to the predictions of TAP, memory performance was not highest when there was a match between intervening test and final retention test.

Though no support for the TAP account at this fine grained level has been demonstrated, it is still universally cited as a contributor to the testing effect for its ability to parsimoniously explain the basic phenomenon: being tested for material previously learned leads to better testing performance later on. The processes involved at encoding (testing-in this case intervening tests can be thought of as an additional encoding period) are invoked again at retrieval which leads to improved memory. That tests of free recall and short answer consistently yield the largest effect likely speaks more to the strengths of the effortful retrieval hypothesis (free recall is certainly more difficult than a test of multiple-choice where the correct answer is provided amidst a few lures) than the shortcomings of the TAP account.

Though these two accounts are the most universally cited explanations for the testing effect, Szpunar, McDermott, & Roediger (2008) demonstrate that testing has the added benefit of insulating against the buildup of proactive interference (previously learned material interfering with the retrieval of newly learned material). In their study, participants were presented with five 18-word lists. Following the presentation of each of the first four lists, half the participants were given a distracter task, and half were given a free recall test for the previously presented list. Importantly, following the presentation of the fifth and final list, all participants were given a free recall test (for that, the final list; list five). Of interest was the number of inter-list intrusions (erroneously recalling items from lists 1-4) participants made during the list five recall in the testing vs. no testing conditions, as well as recall for the list five items during a final recall test (participants were asked to recall words
from all five lists) administered 30 minutes after the completion of the list five recall. The results (Figure 6) show a clear advantage for the tested material: not only did participants in the test condition produce less than 10% of the intrusions compared to the no test participants, they also correctly recalled more than twice the number of words both on the initial list five test, and on the final recall test. The authors explained the results in terms of a source monitoring framework. Testing serves to clearly segregate and distinguish the lists in memory, permitting more efficient source monitoring and a reduction of cue overload.

Additional support for this theory comes from the demonstration of a negative effect of testing similar to the one seen with multiple-choice tests. In a study investigating the effects of testing on eyewitness memory, Chan, Thomas, & Bulevich (2009) demonstrated that an immediate free recall of details following an eye-witness event intensified the recall of post-event misinformation. According to the authors, the increase in misinformation was due, in part, to the fact that testing after the witnessed event reduced the level of proactive interference that the original witnessed event exerted on new learning (in this case, the misinformation).

This theory offers a unique component to our understanding of why testing improves memory, especially in the more naturalistic classroom setting. Over the course of a semester the average student will be presented with an array of related concepts and ideas which he or she will not only have to learn but discriminate (e.g. compare and contrast these concepts). A student in a typical mid-term/final-only course will be more vulnerable to the buildup of proactive interference (for example, confusing anaerobic respiration with the previously learned aerobic respiration). By testing the student following the presentation of each
concept, that retrieval period can serve to insulate against the buildup of proactive interference which reduces confusion and aids in final test performance.

Though several explanations have been offered as to why testing enhances memory no research has conclusively explained the aforementioned interaction between testing condition and retention interval (e.g. Roediger & Karpicke, 2006b). Authors have, however, offered speculation as to why retention of restudied material degrades so much more quickly than tested material (e.g. Wheeler, Ewers, Buonnano, 2003; Roediger & Karpicke, 2006b; Toppino & Cohen, 2009). Some of the explanations offered may provide additional insight as to how testing improves memory.

Roediger & Karpicke (2006b) suggest the interaction is much like the spaced vs. massed practice effect. At short intervals, there is a benefit for massed studying, whereas for longer intervals, there is a benefit of spaced studying. While this may be able to explain why recall performance of the SS performance quickly falls over time (reading and rereading a passage can justifiably be considered a massed studying manipulation) it is probably inaccurate to conceptualize the ST condition as a spaced studying manipulation. That is because the critical component to spaced studying is the intervening time between encoding sessions, a factor which the ST condition does not manipulate.

Conversely, other authors (Wheeler et al., 2003; Toppino & Cohen, 2009) point to a theory proposed by Bjork & Bjork (1992) to explain the interaction. With respect to item memory, Bjork & Bjork (1992) distinguish between an item's retrieval strength (current accessibility) with its storage strength (degree of learning). The probability of recalling an item is determined entirely by the item’s retrieval strength. This explains why it is easier to remember the room number of the hotel you have been staying at for the past few nights (low
storage strength, high retrieval strength), and much more difficult to remember the telephone number from your childhood home (high storage strength, low retrieval strength).

As this theory relates to the testing effect, both testing and restudying are thought to increase storage strength, but successful retrieval (through testing) has the added bonus of enhancing retrieval strength. This fact alone, however, cannot explain the interaction, prompting Wheeler et al., (2003) to further theorize that enhancements of retrieval strength are longer lasting than enhancements of storage strength. As such, longer intervals will yield greater benefits of testing, offering a plausible explanation for the interaction between testing condition and retention interval. However, thus far, no empirical evidence has been offered in support of this theory.

*Explanations of the testing effect: The item-specific vs. relational account*

Perhaps the most compelling theory to explain how testing improves memory is borne out of a recent study from Karpicke & Zaromb (2010). In the study, the authors were interested in examining the relationship between the testing effect and the generation effect.

On the surface, the two phenomena bear a strong resemblance. The testing effect is evidenced by improved recall as a result of generating a response during an intervening test. Similarly, the generation effect refers to the phenomenon whereby material which is actively generated results in improved recall performance relative to material that is passively read (see Mulligan & Lozito, 2004 for review). In a typical generation study, participants will be instructed to learn a list of items such as antonym word pairs (e.g. *north-south*) under two conditions. In the read (control) condition, participants would be presented with both words to read aloud, whereas in the generate condition, they would have to generate the target antonym given the cue in isolation (e.g. *north-___*). Typically information which has been
generated in this way is recalled better than information which has been passively read (e.g. Slamecka & Graf, 1978; Jacoby, 1983; Masson & MacLeod, 1992; Mulligan, 2001; 2002).

Traditionally, the generation effect has been conceptualized within the item-specific vs. relational framework (Hunt & McDaniel, 1993). The account asserts that when learning a list of words, one does not merely encode each individual item in isolation. Rather, associations between the items are also formed. Item-specific processing can be considered the processing of features unique to a stimulus, whereas relational processing can be thought of as the processing of common features shared by a set of stimuli such as the connections between cues and targets, and between the individual items and the list as a whole.

Consider a study list composed of a series of word pairs (e.g. hot-cold, north-south). Item-specific processing can be understood as the degree to which you attend to the particular features of each item, in this case, each word individually. Relational processing, conversely, can be further broken down into several subcomponents. The connections or associations made between the cue and target (in this case, antonyms) can be thought of as cue-target relational processing. The connections made between each word pair in the study list can be considered inter-item relational processing. In the example given, perhaps you make the connection that it is hot in the south and cold in the north. Inter-item relational processing also refers to the connections one makes between each item and the list as a whole, in this case how you relate one particular antonym word pair to all of the other antonyms presented in the list.

Importantly, different tests of memory rely differentially on these types of processing. Free recall relies both on relational information to help delimit potential responses and item-specific information to accurately select among these potentially correct items. In contrast,
recognition tests (deciding whether an item is old or new) do not require the generation of potential responses, so relational information is less important. As a result item-specific processing dominates this type of test. Still other tests rely heavily on relational processing. Tests of order memory, in which participants are required to reconstruct the order in which a series of items were presented, rely exclusively on inter-item relational processing.

According to the item-specific vs. relational account at its most basic level, generation contrasts a common (read) and an unusual (generate) condition. Unusual stimuli, such as items to be generated, attract more attention for interpretation resulting in greater item-specific processing. This enhancement in item-specific processing, however, comes at the cost of relational processing, specifically inter-item relational processing. This is thought to be because we have a limited amount of resources with which to encode an item. Therefore, with generated items, there is necessarily a trade-off such that the enhancement of item specific processing limits the resources available for the encoding of inter-item relations.

Common stimuli, such as read items, attract less processing of item characteristics, allowing for greater processing of the inter-item associations. During a mixed list presentation of both generate and read items (as one would find when generation is manipulated within subjects), the disruption of inter-item processing caused by the occasional generate item uniformly hinders recall of both read and generate items. However, the generate items uniquely benefit from the enhancement in item specific processing. This accounts for the basic generation effect: improved recall for generated material compared to read material.
This account has been applied to explain a variety of memory effects (see McDaniel & Bugg; 2008, for the item-order account, a variant of the item-specific vs. relational framework) and is preferred because it can explain not only the basic scenario where generation improves memory, but also situations in which there are null or even negative effects of generation. The basic positive effect of generation outlined above is typically found under the following conditions: (1) when generation is manipulated within subjects and (2) when free recall or recognition is used to measure retrieval. Changing either of these conditions can nullify and even reverse the effect.

First consider the issue of how retrieval is measured. Studies employing either free recall or recognition tests typically find positive effects of generation because both tests rely upon item-specific processing, which generation is thought to enhance. By contrast, consider now a test of order reconstruction. In these tests, participants are provided with all of the previously presented items, though in a scrambled order. The participant’s job is to reconstruct the original presentation order. Here item encoding is of little importance because all the items are provided for you. Rather, the associations among items are critical - specifically inter-item relational processing. According to the item-specific vs. relational account, because generation disrupts inter-item processing, participants who have generated a list of words typically are worse at reconstructing the order of the items relative to those who simply read the words, a negative effect of generation (Nairne, Riegler, & Serra, 1991; Mulligan, 2002; Serra & Nairne, 1993).

Similarly, the implementation of a within-subject design is critical for demonstrating a generation effect. Recall that with respect to within-subject designs, generate items disrupt order processing for both generate and temporally contiguous read items. However, the
generate items uniquely benefit from the enhancement in item-specific processing, resulting in a positive generation effect on a later recall test. By contrast, in between-subject designs participants are presented with a list of all generate or all read items. In this case, the pure list of generate items benefits from item-specific enhancements, but this benefit is offset by the disruption in inter-item processing. Pure lists of read items produce less item-specific processing but inter-item processing is left intact. As a result, recall on the two lists is relatively comparable. Indeed, those studies which have manipulated generation between subjects with free recall tests typically find no generation effect (e.g. Hirshman & Bjork, 1988; Grosofsky, Payne, & Campbell, 1994). This result also holds true when generation is manipulated within subjects in pure blocks or lists (Mulligan & Peterson, 2008). In this study participants studied two lists of words, one encoded under generate instructions and one encoded under read instructions. As the item-specific vs. relational account would predict, there was no effect of generation on free recall. Importantly, however, this effect is specific to free recall—studies employing a recognition test of memory typically see a benefit of generation, regardless of whether it is manipulated between or within subjects. This is thought to be because disruptions in relational processing are of little consequence to recognition, a test relying exclusively on item-specific processing (McDaniel & Bugg, 2008).

This between-subject/within-subject distinction is particularly important because it demonstrates a potentially critical difference between the generation effect and the testing effect: unlike generation, the benefits of testing are found in both within-subject and between-subject designs. Though most studies opt for a within-subject design because it provides more power, those that have used a between-subject design still find a positive
effect of testing (e.g. Carpenter & DeLosh, 2006; Karpicke & Roediger 2007; Kornell et al.,
2009; Roediger & Karpicke, 2006b; Wheeler et al., 2003)\(^6\).

Recently, Karpicke & Zaromb (2010) developed a clever experimental design to compare these two memory phenomena directly. In the first experiment participants were presented a series of words to read (e.g. love, diet) without any warning of an impending memory test. In the second phase of the experiment, participants were presented with a series of semantically related cue-target pairs in which the cue was novel and the target was an item from the previously presented word list in Phase 1 (e.g. heart-love, eat-diet). In the read condition, participants silently read the intact word pairs. In the testing condition participants were presented with the cue in isolation (e.g. heart-____) and were instructed to complete the word pair with a semantically related word from the previously studied items in Phase 1. In the generate condition participants were presented with an intact cue and a fragment of the target (e.g. heart-l-v-, eat-di-). Participants were instructed to complete the fragment with the first word that came to mind. Critically these fragments had more than one plausible semantic completion (e.g. heart-love / heart-live, eat-diet / eat-dine). The pairs were constructed based upon word frequency norms such that one completion was more likely to be produced than others (in this case love and diet, the words presented in Phase 1).

With this design the authors were able to create a scenario in which the only thing that differed between the generate and testing conditions was whether participants were instructed to retrieve the information from Phase 1, or generate the information themselves.

\(^6\) This is an important theoretical point relevant for the endorsement of testing as an educationally beneficial strategy. If the benefits of testing were only found when contrasted with other items which were not tested (as would be the case for generation), you create a scenario in which only half of the material to be learned can benefit from testing. Because the improvements in recall seen with testing are not limited to within-subject designs, this is not a concern.
Importantly, the base rate of completion in Phase 2 (recalling or generating love and diet) did not differ between conditions.

In Phase 3 all participants were given a free recall test in which they were instructed to recall targets presented previously during Phases 1 & 2 (see Figure 7). As previous research had demonstrated, because of the between-subject design, there was no effect of generation (Hirshman & Bjork, 1988; Grosofsky, Payne, & Campbell, 1994; Mulligan & Peterson, 2008), though there was a robust testing effect (Wheeler et al., 2003; Carpenter & DeLosh, 2006; Roediger & Karpicke, 2006b; Karpicke & Roediger 2007; Kornell et al., 2009).

The second experiment looked much like Experiment 1, though now with generation and testing manipulated within subjects. This was accomplished through a change to the Phase 2 procedure. In Experiment 1 read vs. generate vs. testing was manipulated completely between subjects (i.e. three different groups). In the second experiment, Phase 2 consisted of only two groups: generate and testing. In the generate group, half of the items were presented to be read, whereas the other half were presented to be generated. Likewise, in the testing group, half of the items were presented to be read, whereas half the items were presented to be recalled.

The data (Figure 8) reveal both a significant generation effect and significant testing effect. These first two experiments are important for two reasons. First and foremost it demonstrates that this unusual design (especially with respect to the generation condition) can yield both a generation and testing effect. Additionally, these experiments replicate the finding that the generation effect occurs only during within-subject designs whereas the testing effect occurs both with a within and between-subject design.
In comparing the generation and testing conditions (Experiment 2), there was an interaction such that the difference between testing and reading was larger than the difference between generating and reading (i.e. the testing effect was larger than the generation effect). These data alone are somewhat ambiguous. This could be taken as evidence that the generation effect and the testing effect are two different phenomena. However, it does not rule out the idea that both generation and testing operate within the same item-specific vs. relational framework. It simply could be that the item-specific benefits seen with testing are larger than those seen with generation. Examining the testing effect and generation effects from Experiments 1 & 2 together help illustrate this point explicitly (Figure 9). Moving from a within-subject to a between subject design eliminates the generation effect as one would expect. However, it is not as if testing is immune to the design shift; the testing effect is similarly reduced in the between subject manipulation. The benefit of testing is still significant, of course, though perhaps implies these two phenomena are more similar than the interaction in Experiment 2 would suggest.

To more directly answer this question it is necessary to understand exactly how testing and generation each affect relational processing, an issue which Karpicke & Zaromb address with Experiment 3. This experiment was identical to Experiment 1 (the between-subject manipulation), though instead of a free recall test at Phase 3, participants were given an order reconstruction test (see Figure 10 for results). As demonstrated in prior studies, order recall of generate items was worse than read items. However, the same order disruption occurred with testing, as there was no difference between the testing and generation conditions.
This result is particularly important because it offers the first empirical evidence that testing may be understood within the item-specific vs. relational account. Though the authors take the results from Experiments 1 and 2 to suggest that generation and testing are qualitatively different phenomena, the data, considered as a whole, suggest otherwise. Though the testing effect was larger than the generation effect, the most parsimonious explanation is that both generation and testing disrupt order processing and enhance item processing, though the item enhancement is greater in the testing condition. In other words, the difference may be more quantitative than qualitative. This may also explain why between-subject designs of generation typically find null results, while those with testing still find a testing effect. With generation, the disruptions of order processing roughly counterbalance the enhancements in item processing, leading to no net improvement. Conversely, with testing, item-specific enhancements presumably are sufficiently large to more than compensate for the disruption in order processing.

*The Current Study*

Consider the testing effect and the generation effect together: in one case a participant is generating (i.e. recalling) information from episodic memory (the testing effect), whereas in the other, a participant is generating information from semantic memory (the generation effect). Despite this apparent similarity, over the past twenty years researchers have argued that generation and testing are two qualitatively distinct memory phenomena (e.g. Carrier & Pashler, 1992; Roediger & Karpicke, 2006b; Kornell et al., 2009; Karpicke & Zaromb, 2010). Though it is certainly possible that there is something fundamentally different (and especially advantageous) to episodic generation relative to semantic generation, there
currently is no evidence which can conclusively dissociate these two effects (Karpicke & Zaromb, 2010).

Further complicating this issue is the relationship between generation/testing and the enactment effect. The basic tenant of the enactment effect suggests that action phrases (e.g. *break the pencil*) which have been acted out will be remembered better than those which have been passively read or observed (see Engelkamp, 1998 for a review). Like generation and testing, the item-specific vs. relational account has been offered to explain why enactment improves memory (e.g. McDaniel & Bugg, 2008).

As reviewed earlier, certain experimental paradigms can yield negative effects of generation (and perhaps testing). The same is true with enactment. Steffens (1999, Experiment 2) gave participants a list of verb-object pairs to symbolically enact or read (manipulated between subjects). The object in each pair came from one of several taxonomic categories (e.g. animals, vehicles, etc.). Verbs were manipulated between subjects such that some participants were presented with verbs which emphasized the semantic relationship between the objects (e.g. *stroke-cow, stroke-horse, stroke-dog, stroke-cat*). In the example here, ‘*stroke*’ emphasizes each of the objects as a group of domesticated animals. In contrast to these congruent verb-object pairs, other participants saw incongruent pairs where the verbs did not emphasize the semantic relationship between the objects (e.g. *paint-cow, lift-horse, chain-dog, dress-cat*).

Following encoding participants were given a free recall test (Figure 11). There was a significant enactment effect for the congruent verb-object pairs, though this effect reverses for incongruent pairs. Steffens (1999) described the results within the item-specific vs. relational framework. We know that enactment serves to enhance item-specific and cue-
target relational processing. With respect to the congruent verb-object pairs, cue-target relational enhancements in particular help offset the typical disruption of inter-item relational processing enactment causes because each cue relates to multiple targets. As a result, the semantic grouping of the targets is facilitated. This organization, along with the item-specific enhancements from enactment more generally, help those who enacted the congruent items to recall more objects than those who simply read the congruent items (i.e. a positive enactment effect). Now consider the incongruent items. Enactment still enhances item-specific and cue-target relational processing. However, symbolically lifting a horse or dressing a cat does not aid participants in associating the objects with one another. This, in combination with enactment’s baseline disruption of inter-item processing, results in worse recall performance than those who read the incongruent pairs.

The results here highlight the similarities between the different memory effects (generation, testing and enactment) suggesting a common theoretical framework for interpretation. Looking at all three effects together, one can conceptualize each as an active processing manipulation. In each case, the more active processing condition (generating, testing, enacting) results in superior memory relative to the more passive control (reading, restudying, observing). At a broader level, research clearly states that active or self-initiated encoding produces better memory relative to passive or perceptual encoding. For example, research on persuasion typically shows that self-generated arguments are better remembered than arguments supplied by a speaker (e.g. Petty, Ostrom, & Brock, 1981). Similarly, educational research suggests students retain more information from active as opposed to passive learning strategies (e.g. Kalem & Fer, 2003; Michael & Modell, 2003). Given such a clear and parsimonious framework for interpreting all of these effects together, it is important
to conclusively determine: is, as researchers have suggested, the testing effect really different than other memory effects? If so, then the testing effect truly is unique and qualitatively different from other memory effects.

Alternatively, as suggested by the analysis above, testing might fit into a general theoretical framework suggesting a new theoretical analysis of this effect. If we are to assume that testing operates under the principles outlined in the item-specific vs. relational account, we should be able to engineer a situation in which testing results in worse memory relative to restudy, a negative testing effect. Such a demonstration would be significant for several reasons (1) it would solidify the notion the testing effect can best be understood in terms of the item-specific vs. relational account (2) it would have important theoretical implications to suggest that testing is not always a beneficial study strategy as the current understanding suggests and (3) it would be the first empirical demonstration of a negative testing effect in free recall (i.e. not just worse order memory). Though researchers have labeled the detriments associated with testing on multiple-choice tests as a ‘negative effect’, this is not the most appropriate label. In this case you recall a lure better as a result of selecting it on an earlier test. This is still the basic testing effect (improved memory for material recalled previously) it just happens to for information that was initially erroneously recalled. No study has ever before demonstrated that retrieving information during an intervening test can result in worse memory performance for that same material on a final test.

The critical question, of course, is how to engineer such a scenario. If the generation and testing effects share a common underlying explanation, then situations which result in a negative generation effect should also lead to a negative testing effect. Burns (1990)
provides just such a model (see Burns 1992; Steffens & Erdfelder, 1998 for other similar negative generation effects). In his study (Experiment 1C) participants were presented with a series of cue-target word pairs consisting of two rhyming words (e.g. moon-spoon)\(^7\). Importantly, each target in the cue-target pair was an exemplar from a taxonomic category. For example, the target *spoon* in the word pair *moon-spoon* would fall under the taxonomic category of silverware. Across the list there were six target exemplars from each category (e.g. *wife-knife, cork-fork,* etc.). These exemplars were presented in a blocked fashion such that all six cue target pairs from one category were presented in sequence, followed by six from another category, etc. Participants encoded the word pairs in one of two conditions (manipulated between-subject). In the read condition, participants were presented with the intact pairs to read aloud. In the generate condition, participants were presented with fragments (*moon-sp___*) and instructed to complete the target fragment with a word rhyming with the cue. Afterwards, participants in both conditions were given a free recall test.

According to the item-specific vs. relational account (Hunt & McDaniel 1993; Steffens, 1998; McDaniel & Bugg, 2008) generation increases item-specific and cue-target relational processing while disrupting inter-item relational processing. In this case increases in item-specific and cue-target processing should enhance processing of each cue-target pair in isolation and emphasize the rhyming nature of each pair. However, this information would not be useful in the retrieval of the targets during a free recall test. Conversely, generation should disrupt inter-item relational processing, which would suggest participants would be less likely to identify and encode the semantic relationship among the items on the list (e.g. recognizing that the targets *spoon, knife* and *fork* are all examples of silverware). Unlike the rhyme information, the ability to identify the taxonomic categories would be extremely

\(^7\) The author does not provide a list of stimuli used, so examples here are arbitrarily chosen
useful as a retrieval cue during the free recall test. To illustrate, if you remember spoon was on the list, and additionally that it was part of a greater set of silverware exemplars, you will be more likely to also recall knife and fork by capitalizing on the relational nature of the items. As such, the trade-off account predicts that participants in the generation condition would actually recall fewer targets than the read group, a negative generation effect. This is precisely what Burns (1990) found: participants in the read condition recalled more targets than those in the generate condition.

This design provides a nice model for looking at a potential negative testing effect for a couple reasons. First and foremost it is a clear, theoretically motivated design which resulted in a significant negative generation effect. Additionally, it helps mitigate the issue related to testing’s significant enhancements to item-specific and cue-target relational processing. In Karpicke & Zaromb (2010) the improvements to these two types of processing associated with testing so far outweighed disruptions in inter-item relational processing, there still was a significant testing effect in recall. With this new design, the item-specific enhancements should still improve free recall. However, by implementing a series of semantically unrelated rhyming cue-target pairs, the enhancements in cue-target relational processing should emphasize the rhyming nature of each pair. Because participants will be instructed to recall only the targets in a free recall test, this type of processing should be less helpful.

Again, if generation and testing share a common underlying explanation, adoption of this general design should allow for the demonstration of a negative testing effect. This would prove a theoretically significant demonstration for the reasons outlined above. If however, we demonstrate a positive effect of testing, the results will be equally meaningful.
First and foremost, it would suggest that the testing effect should not be understood in terms of the item-specific vs. relational account, contradicting results from Karpicke & Zaromb (2010). Perhaps more importantly, it would provide the most convincing evidence that testing is a qualitatively different memory phenomenon- a notion that many researchers endorse, but an idea for which there is still little compelling evidence (Karpicke & Zaromb, 2010).
CHAPTER 2

EXPERIMENT 1

Given the clear motivations outlined above the first experiment followed the design of Burns (1990, Experiment 1C) with modifications to make it apply to the testing manipulation. In the first phase of the experiment participants were presented with a list of 36 word pairs (e.g. wife-knife) to read silently. Each word pair consisted of a rhyming cue and target. The targets included six different exemplars (e.g. spoon, knife, fork, etc.) from six different taxonomic categories (e.g. silverware, animals, fruits, etc.). Each cue rhymed with its accompanying target, but was not itself a member of any of the six target categories. The word pairs in this first phase were presented in a randomized order. Following the presentation of the word pairs in Phase 1 participants completed a brief math distractor.

In the second phase of the experiment, participants encoded the cue-target pairs a second time. The pairs were presented in one of two ways depending on condition (manipulated between subjects). In the restudy condition participants were presented with the intact word pairs with instructions to read each word pair aloud. In the retrieval condition, participants were given the cue in isolation (e.g. wife- ) with instructions to recall the rhyming target word presented during Phase 1. Regardless of whether participants correctly recalled each target, feedback was provided in the form of the intact word pair (e.g. wife-knife). After an additional math distractor, participants in both conditions were given a free recall test and instructed to recall any of the targets presented during Phases 1 & 2.
A Note on the Original Research Plan and a Preview of the Results of Experiments 1 and 2

The first two experiments are presented here in the historical order conducted. As originally proposed, Experiment 1 was to compare retrieval with restudy, and given a positive testing effect (which was demonstrated) Experiment 2 was to employ a similar design comparing generate with read. In this sense Experiment 2 was to be a control experiment to verify that the present experimental methods would replicate the negative generation effect reported in prior research. This is necessary to cleanly interpret the positive testing effect in Experiment 1- a significant testing effect would not really imply qualitative differences between the testing and generation effects unless a negative generation effect could be demonstrated under comparable circumstances. Because Experiment 2 failed to demonstrate a negative generation effect, the results of Experiment 1 do not provide clear evidence for differences between testing and generation. Given the limited interpretability of these data, the discussion of Experiments 1 and 2 will be limited in scope.

Method

Participants

34 undergraduates from the University of North Carolina participated in exchange for course credit.

Materials

The stimulus set consisted of 36 cue-target word pairs. Targets (6 exemplars from 6 different taxonomic categories) were taken from the category norms of Van Overschelde, Rawson, & Dunlosky (2004), an updated version of Battig & Montague’s (1969) category norms. Targets had a mean rank frequency of 4.45, and a mean frequency of 62 (Kucera &
Francis, 1967). A rhyming cue was selected to accompany each target, which itself was not a member of any of the 6 target categories. (See Appendix 1 for a complete list of materials used).

**Design**

Experiment 1 used a pure-list between-subjects design. Half of the participants were randomly assigned to the testing condition, and half of the participants were randomly assigned to the restudy condition.

**Procedure**

Like Karpicke & Zaromb (2010) the experiment consisted of three phases. In the first phase participants were presented with the 36 cue-target pairs in a pseudo-randomized order with the constraint that no target items in adjacent serial positions were from the same category. Each word pair was presented one at a time on a computer monitor for 4 sec each with a 500 ms interstimulus interval. Participants were instructed to read the word pairs silently and remember them for a later (unspecified) memory test. After the initial presentation, participants were given a 5 min distractor task of math problems.

In the second phase of the experiment, participants were presented with the word pairs again in one of two manners. In the restudy condition participants were told they were going to be given a second opportunity to learn the word pairs for the impending memory test. Intact pairs were presented for 6 sec\(^8\) each with instructions to read the word pairs aloud. In the testing condition, participants were told they were going to practice recalling the information for the impending memory test. Here participants were given each cue in

---

\(^8\) This timing is based upon Karpicke & Zaromb, where pilot testing revealed 4 sec was the average time participants needed to recall a word in the testing condition of Phase 2. This plus the 2 sec of feedback which will be given in the testing condition results in the 6 sec restudy participants will see the word pair.
isolation (e.g. *wife-*) with instructions to recall the accompanying, rhyming target presented during Phase 1\(^9\). Participants read the cue aloud and recalled the target aloud as well. Participants had 4 sec to retrieve the target after which point feedback was provided in the form of the intact word pair (e.g. *wife-knife*) presented for 2 sec. If the participant had failed to recall the target, or had inaccurately recalled the target during the allotted 4 sec, they were instructed to read the newly presented (i.e. correct) target aloud. Importantly, unlike Phase 1, the word pairs in Phase 2 were blocked by taxonomic category such that all six cue-target pairs from the silverware category were presented, followed by all six cue-target pairs from the fruit category, etc., although the participants were not explicitly informed of this list structure. Following Phase 2, participants were given a second 5 min math distractor task.

In the final phase of the experiment, participants were given a blank sheet of paper and pen to recall the targets presented in the first two phases of the experiments. Participants were reminded that the targets were the second word in each of the word pairs. This free recall test lasted 5 min. Following the free recall test participants were given an awareness questionnaire to determine whether or not they were aware of the categorical nature of the targets. Participants were first asked if they noticed anything unusual about the previously presented targets. They were then asked if they noticed any similarities amongst the targets. If, from these two questions, it was clear that the participant was aware of the categories, they were then asked whether they used the categories to help guide their retrieval of the

\(^9\) One may wonder if participants in the recall condition are simply generating a rhyming word rather than recalling one from Phase 1. In a pilot study participants were given the cues in isolation with instructions to generate the first rhyming word which came to mind. On average, participants happened to generate the correct target 15.6\% of the time (5.6 matches out of 36 items). Because retrieval performance during Phase 2 was considerably higher than that, it is reasonable to assume participants were not merely generating a rhyming target.
targets during the final recall test. Finally, participants were debriefed and thanked for their participation.

Results & Discussion

Phase 2 cued recall performance:

During Phase 2, participants in the retrieval condition had to retrieve the accompanying rhyming target from Phase 1 given each cue. Participants correctly recalled 46% of the targets given each cue. This was considerably higher than the 15.6% of targets correctly generated in the pilot study. This helps support the argument that participants in the retrieval condition were not merely generating a rhyming target, but were rather episodically retrieving the targets from Phase 1. Of course, on those trials when the target was not recalled it was still presented as feedback.

Final recall performance:

The proportion of targets recalled in each of the two conditions can be found in Figure 13. The significance level for this and all other analyses was set to .05. An independent sample t-test revealed that participants in the retrieval condition recalled significantly more targets than those in the restudy control (i.e. a positive testing effect), \( t(32) = 2.48, p=.02 \).

Two additional measures can help shed light on how participants used the categorical nature of the targets to help guide retrieval. The first is awareness of the categories. During the awareness questionnaire participants were asked if they were aware of the categories, and if so whether or not they used that information to help guide their retrieval of the targets. In both this and all other experiments to be reported, in each and every instance in which the participant was aware of the categories, they indicated they used this information to help
guide their retrieval of the targets later at test. As such, awareness was coded as a binary variable of either being aware of the categories, or being unaware of the categories. The proportion of participants who were aware of the categories in each condition can be found in Figure 14. Though there was a numerical advantage for participants in the retrieval condition to be more aware of the categories, the difference was not significant, $\chi^2 (1, N=27) =2.05$, $p=.1510$.

Category clustering is another measure which can be helpful reveal the extent to which each group relied on the categorical structure of the material to support target search. Clustering was measured via the adjusted ratio-of-clustering (ARC) scores (Roenker, Thompson & Brown, 1971), which have a value of 0 for chance-level clustering, positive values for above-chance clustering, and a value of 1 for perfect clustering. The mean ARC scores for each experiment can be found in Figure 15. An overall analysis of the ARC scores indicated the distribution of the scores was non-normal (see Figure 16). As such, prior to hypothesis testing for this and all other experiments the ARC scores underwent a square root transformation: $\sqrt{1-ARC}$. The constant 1 was used to render negative ARC scores positive. This transformation yielded a closer approximation of a normal distribution (Figure 16), important to fulfill the assumptions of parametric tests.

The mean ARC score for each condition in Experiment 1 can be found in Figure 15. There was little evidence for category clustering. Neither of the groups mean ARC scores differed significantly from 0, nor did they differ from each other ($t < 1$).

---

10 This measure was introduced after the seventh participant, so data is only available on 27 of the of the 34 participants run
CHAPTER 3

EXPERIMENT 2

The intent of the design of Experiment 1 was to render inter-item information important for later memory. The more a participant realized the categorical nature of the targets, the more this could be used to help guide retrieval later during the final free recall test. The item-specific vs. relational framework suggests that participants in the retrieval condition should be less likely to pick up on this because more resources are devoted to cue-target relational processing. Because this rhyming information is relatively less helpful, the item-specific vs. relational framework predicts a negative effect of testing. Contrary to this prediction, testing actually improved memory performance. These data, then, suggest the testing effect is truly unique, and cannot be explained by the aforementioned account.

However, these data alone must be interpreted with caution. The rationale for employing this particular design was that it was one which had previously elicited a negative generation effect (Burns, 1990). Given a positive effect of testing was found, it is necessary to demonstrate the design, as constructed, can actually yield said negative generation effect. If so, then the positive effect from Experiment 1 can be justifiably interpreted as before. If not, it would suggest that the design is not sensitive enough to demonstrate any negative effect, and the positive effect of Experiment 1 would be less interpretable. As such, the design of Experiment 2 was a basic replication of Burns’ (1990) original study to demonstrate a negative generation effect.
Method

Participants

34 undergraduates from the University of North Carolina participated in exchange for course credit.

Materials

The materials were the same as those used in Experiment 1.

Design

Experiment 2 used a pure-list between-subjects design. Half of the participants were randomly assigned to the generate condition, and half to the read condition.

Procedure

Participants studied the word pairs in one of two ways. In the generate condition, participants were presented with the intact cue along with the first two letters of the target (e.g. wife – kn ). Participants were told to read the cue word aloud and to generate aloud a rhyming word (the target) using the first two letters provided\textsuperscript{11}. In the read condition, participants were presented with the cue-target pairs intact, and simply read both words aloud. In both conditions participants were told to remember the pairs for a later (unspecified) memory test. Importantly, during this single presentation the pairs were all grouped together based on the categorical membership of the targets, as in Phase 2 of Experiment 1. Following the encoding period, participants completed a 5 min math distractor, and were then given a 5 min free recall test to recall the previously presented targets.

\textsuperscript{11} Because the cues and targets rhymed, the first two letters provided constrained completion to one and only one response, the correct target. As such performance during this phase (i.e. correctly generating the target) was 100%
**Results**

The proportion of targets correctly recalled in each of the two conditions can be found in Figure 13. Contrary to expectations, there was no difference between the conditions \( t(32) = .41, p=.69 \). Indeed, the non-significant mean difference was in the direction of a positive rather than a negative generation effect. According to a 2 (Experiment 1 vs. Experiment 2) \( \times \) 2 (retrieve/generate vs. restudy/read) ANOVA the pattern of results did not significantly differ between Experiments 1 and 2 (\( F(1,32) = 1.57, p=.22 \)). The proportion of participants aware, and the mean ARC scores for each condition can be found in Figures 14 and 15 respectively. Significance testing indicated no differences either for awareness (\( \chi^2 (1, N=33) = 0.03, p=.86 \), or clustering (\( t(29) = .39, p=.70 \)).

**Discussion**

Unlike Burns (1990), Experiment 2 did not demonstrate a negative generation effect. The rationale for using this design was that it was one that had already been shown to yield a negative generation effect. Because Experiment 2 did not demonstrate such an effect, the positive testing effect in Experiment 1 becomes much less interpretable. It could be the case, as suggested earlier, that testing improves memory even with a design that emphasizes inter-item relational processing. More likely, however, it could be that the design as constructed is simply not sensitive enough to demonstrate a negative effect of either testing or generation.

Though the results of Experiment 2 suggest the design was not sensitive enough to test the original research question, two important points can be drawn from the results. First, the results are generally consistent with prior research. A between subjects design yielded a positive effect of testing, though a null effect of generation (Hirshman & Bjork, 1988; Grososky, Payne, & Campbell, 1994; Wheeler et al., 2003; Carpenter & DeLosh, 2006;
Roediger & Karpicke, 2006b; Karpicke & Roediger 2007; Kornell et al., 2009; Karpicke & Zaromb, 2010). Of course, as detailed earlier, there are two competing explanations to account for this difference. More importantly for the present study, there was a significant testing effect in Experiment 1. The explanation behind the effect itself is difficult to explain at this point, but the fact that one can be demonstrated at all, especially at such a short retention interval, will be especially important in light of the results in Experiment 4.
CHAPTER 4

EXPERIMENT 3

The critical question, of course, is why? Why did Burns (1990) demonstrate a negative generation effect, whereas Experiment 2 did not? The answer seems to lie in the fragility of the negative generation effect itself. Though Experiment 2 was in most respects a direct replication there were five components which differentiated the two studies. (1) The word pairs used in each of the experiments were different. (2) In Burns’ study, participants saw each of the pairs on a projector rather a computer screen. (3) In Burns’ study participants were told to focus on the target word, whereas in Experiment 2, participants were told to remember both words. (4) In Burns’ study there was no distractor phase; following the presentation of the final word pair, participants were immediately given the free recall test. (5) In Burns’ study participants were presented with each word pair for a longer duration (15 sec vs. 6 sec in Experiment 2 reported here).

Though each component by itself (or any combination of these components) could explain the discrepant results, there was no reason a priori to expect that the decision not to replicate these particular factors would yield a qualitatively different pattern of results. Steffens & Erdfelder (1998) employed a similar design (rhyming cue-target pairs, categorically related targets), and found a comparable negative generation effect. In that study all but one of the aforementioned factors were also violated: the stimuli (i.e. word pairs) were different, the pairs were presented on a computer screen, the participants were not instructed to focus on the target, and the duration was short (7 sec). There was one
commonality, a lack of a distractor phase, but there is no theoretical reason to suggest the addition of a 5 min math distractor would render an otherwise negative generation effect null. These two designs (Burns, 1990; Steffens & Erdfelder, 1998) vary along several other dimensions yet yield the same result. Given this, it was reasonable to assume the negative generation effect would be a rather robust effect (i.e. insensitive to minor differences between designs). The results of Experiment 2, however, suggest otherwise.

To answer my original research question it is critical to demonstrate the same pattern of results as Burns (1990). This was the goal for Experiment 3. Because the results of Experiment 2 suggest this negative generation effect may be more delicate than previously assumed it was necessary to reexamine which factors might be changed to more closely replicate Burns’ original study. One obvious step would be to use the same word pairs, though this proved impossible. From a theoretical perspective this should not be problematic, as the pairs in the current study were constructed in a similar fashion from the same database of words.

The next factor considered was the timing. In Experiment 2, participants saw the word pairs for 6 sec each whereas in Burns (1990) participants were presented with the pairs for 15 sec each. More time for each trial should allow for more overt rehearsal. According to the item-specific vs. relational account the type of rehearsal should differ between the two conditions. In the read condition, more time for each item would allow more time for inter-item relational processing. In other words, the more time the participant is given the more time they would have to pick up on the relational nature of the targets, facilitating later memory for those targets. Conversely, in the generation condition, additional time should afford the participant more opportunity to think about the cue-target relational information.

12 Through a personal correspondence I learned Daniel Burns was no longer in possession of the original stimuli
This rhyming information should presumably be less helpful later during a free recall test of the targets. In short, more time should be preferentially more beneficial to the read condition, and thus more likely to elicit a negative generation effect. Given this reasoning, in Experiment 3 presentation rate for the word pairs was increased to 15 sec.

Because the goal of Experiment 3 was to demonstrate a negative generation effect, and not a fine-grained test of why the results of Experiment 2 differed from Burns (1990), two other factors were changed in addition to the timing. Specifically, participants were told to focus their attention on the target words, and the distractor phase was removed. Following the presentation of the final word pair, participants were immediately given the free recall test.

**Method**

**Participants**

56 undergraduates from the University of North Carolina participated in exchange for course credit, or $10.

**Design, Materials, and Procedures**

The design, materials and procedures were identical to Experiment 2 with the following exceptions. (1) Prior to the presentation of the word pairs, participants were told it was important to focus attention upon the targets. This was the extent of the guidance provided, and participants were still not given any explicit information about the memory test to follow. (2) During encoding, each word pair was presented for 15 sec. Again, the pairs were presented either intact (for participants in the read condition) or with the cue intact, and the first two letters of the target (for participants in the generation condition). (3) In this third experiment there was no math distractor. Following the presentation of the final word pair
participants were provided instructions about the memory test which would immediately follow. The time between the end of the Phase 2 and the beginning of Phase 3 was about 30 sec.

Results & Discussion

The proportion of targets recalled in both the generate and read conditions can be found in Figure 13. There is a numerical pattern for participants in the read condition to recall more than those in the generate condition, and is of the same magnitude of both Burns (1990) and Steffens & Erdfelder (1998); in all three studies participants in the read condition recalled roughly 10% more targets than those in the generate condition. However, in the current study the variability within each condition was quite high. As a result the difference failed to reach significance $t(54) = 1.55, p=.13$.

In comparing overall rates of recall with those from Experiment 2, the changes implemented had the predicted effect. A 2 (Experiment 2 vs. Experiment 3) × 2 (generate vs. read) ANOVA indicates participants in Experiment 3 recalled more targets ($F(1,43) = 50.58, p < .01$) than those in Experiment 2. This was presumably driven by the increase in awareness ($\chi^2 (1, N=89) =12.61, p < .01$) and more category clustering ($F(1,34) = 27.21, p < .01$), two components which increased substantially relative to Experiment 2 (Figure 14).

Comparing the read and generate conditions within Experiment 3, there was a marginally significant trend for participants in the generate group to have more awareness of the categories as the item-specific vs. relational account would predict $\chi^2 (1, N=56) =3.31, p=.06$. With respect to category clustering, there was a numerical advantage for more clustering in the generate group compared to the read group (Figure 15) though this was not significant $t(48) = .56, p=.58$. 
The changes implemented in Experiment 3 were successful in that the desired pattern emerged: better recall in the read condition compared to the generate condition. In Experiment 2 awareness levels in both conditions were comparable, around 50%. Of the changes introduced in Experiment 3, two likely had a direct impact on awareness levels. Cueing participants into the targets likely played a role in the higher awareness levels relative to Experiment 2, though this should have been a comparable enhancement for both conditions (i.e. a main effect). Conversely, the additional time, as outlined earlier, should preferentially bias the participants in the read condition to be more aware of the categories (i.e. an interaction), a pattern which was generally found, though was not significant $F(1,42) = 1.65, p = .21$. This of course is only one plausible explanation, it could in fact be some interaction between the two factors. Because the changes were instituted simultaneously, it is impossible to discern.

It is important to reiterate, however, that the most critical component, the negative generation effect, was not significant. This result lends more evidence to the notion that the negative generation effect is simply more volatile than originally thought. Despite the marginal effect, because Experiment 3 was a direct replication of Burns (1990) in every meaningful way, and because the pattern of results were qualitatively similar, the decision was made not to collect more data for Experiment 3. I felt running a substantial number of additional participants simply to reach the .05 threshold would have been an inefficient use of both time and resources (i.e. participants), both of which were at a premium.
CHAPTER 5
EXPERIMENT 4

Despite the fact that the results of Experiment 3 were not as conclusive as one might hope, the pattern was clear enough, and suggested that the design as currently constructed had a reasonable chance of demonstrating a negative testing effect should there be one. Accordingly, the design of Experiment 4 was clearly motivated: a conceptual replication of Experiment 3 comparing participants retrieving or restudying previously presented word pairs.

Along with these two conditions, a third between-subject condition was added. Following a preliminary examination of the initial data, an interesting pattern began to emerge: participants in Experiment 4 (in both the retrieval and restudy groups) were recalling fewer targets than participants in Experiment 3. This was surprising given participants in Experiment 4 were exposed to the words twice, whereas participants in Experiment 3 were exposed to them only once. This is discussed in further detail in the General Discussion, but briefly, there are two possible explanations for this pattern. (1) The presence of Phase 1 itself (i.e. seeing the list of word pairs in a scrambled, non-structured order) so strongly biased participants to think of the items as unrelated, that even during Phase 2 when the targets were presented in categorical blocks, they had trouble noticing or effectively using the categorical structure. (2) This biasing occurred not simply due to the presence of Phase 1, but rather because participants thought back to Phase 1 during the encoding of the Phase 2 lists, either
because they were explicitly instructed to do so (as in the retrieval condition) or because they
did so incidentally (as may happen in the restudy condition).

Both of these hypotheses offer plausible explanations for why performance for the
restudy controls in Experiment 4 was lower than the read controls in Experiment 3. To tease
these two possibilities apart, a third group was added. For participants in this condition, as
with the other two conditions, each word pair was presented intact in a scrambled order
during Phase 1. During Phase 2 the cue was presented intact, and participants were
instructed to generate the target using the first two letters provided (e.g. wife – kn ).
Critically, no mention was made that the word pairs were the same as those from Phase 1. If
the process of thinking back is what is critical, then performance should be comparable (or
superior) to generation performance in Experiment 3 (and greater than the retrieval condition
of the present experiment). If, conversely, the mere presence of Phase 1 biases encoding
during Phase 2, performance should be significantly worse than the generate condition of
Experiment 3 (and comparable to the retrieval condition). Either way this condition was
critical for illuminating the origin of the unanticipated discrepancy in memory performance
between the two experiments.

Method

Participants

84 undergraduates from the University of North Carolina participated in exchange for
course credit, or $10.
Design

Experiment 4 used a pure-list between-subjects design. One-third of the participants were randomly assigned to the restudy condition, one-third to the retrieval condition, and one-third to the generation condition\textsuperscript{13}.

Materials and Procedures

The materials and procedures were identical to Experiment 1 with the following exceptions. Like Experiment 1, Phase 1 was the same for each of the three conditions. During Phase 2, participants were presented with each word pair for 15 sec. In the restudy condition, participants were presented with the intact pair for 15 sec. In the generation condition participants were told that they were going to be presented with another series of words. The words were the same word pairs from Phase 1 though participants were not informed of the relationship. Following the instructions for generation, participants were presented with the cue intact and the first two letters of the target (e.g. wife – kn ) for 15 sec each. In the retrieval condition participants were presented with the cue in isolation for 10 sec. During this time the participant was instructed to recall the accompanying rhyming target presented earlier. After 10 sec, the correct target was presented for feedback for 5 seconds. The proportion of time for retrieval vs. feedback was the same as in Experiment 1. Also during Phase 2 (as in Experiment 3), participants were told to focus their attention on the target words. Lastly, there was no math distractor. Following the presentation of the final pair, participants were immediately given the 5 min free recall test.

\textsuperscript{13} The generation condition was included following the 28\textsuperscript{th} participant run (14 participants in both the restudy and retrieval conditions). Following this a new subject counterbalance was created in which participants were randomly assigned to one of the three different conditions (restudy, retrieval, generate). Because there was to be an equal number of participants in each condition (28) participants necessarily were more likely to fall into the generate condition.
Results & Discussion

Phase 2 cued-recall performance:

During Phase 2, participants in the retrieval condition had to retrieve the accompanying rhyming target from Phase 1 given each cue. Participants correctly recalled 63% of the targets given each cue.

Final free recall performance:

The proportion of targets recalled in each of the three groups can be found in Figure 13. A one-way ANOVA revealed a difference between the mean recall rates of the groups $F(2,81) = 4.33$, $p = .02$. Planned contrasts revealed that participants in the restudy condition performed better than both those in the retrieval condition $t(54) = 2.331$, $p = .02$ and the generate condition $t(54) = 2.513$, $p = .02$. There was, however, no differences between generate and retrieval $t(54) = .464$, $p = .65$.

The awareness data (Figure 14) reveal a pattern consistent with the theory that participants in the restudy control recalled more at least in part because they were more likely to be aware of the categories, though this was not significant $\chi^2 (2, N=84) = 2.67$, $p = .26$. The same is true for the clustering data (Figure 15). A one-way ANOVA indicated there were group differences in category clustering in the three groups $F(2,75) = 3.14$, $p = .05$. Planned contrasts revealed greater category clustering in the restudy condition compared to retrieval ($t(49) = 2.53$, $p = .02$), a marginal, though nonsignificant trend of more clustering in the restudy condition compared to generate ($t(50) = 1.55$, $p = .13$), and no difference between retrieval and generate ($t < 1$).

A 2×2 ANOVA revealed participants in Experiment 4 recalled more targets ($F(1,43) = 24.38$, $p < .01$), were more aware of the categories ($\chi^2 (1, N=88) = 6.78$, $p < .01$) and
clustered more based upon category membership \((F(1,37) = 11.23, p < .01)\) than participants in Experiment 1, again indicating the changes instituted had the desired effect.

The results from this experiment are important for several reasons. First, this is the first demonstration of a true negative testing effect. Retrieving the targets during Phase 2 caused participants to recall fewer targets during a final recall test compared to those who simply studied the word pairs an additional time. That a negative testing effect was demonstrated is certainly exciting; as importantly, though, this is one of the first studies to empirically test and demonstrate support for a theory which explains why tests improve memory.

These results offer compelling evidence to suggest that the testing effect operates upon the same principles as other, more well understood effects such as generation. Tests do not simply enhance memory in some global sense. Consistent with the item-specific vs. relational account testing enhances item-specific and cue-target relational information at the cost of inter-item relational processing. This idea is supported not only by the negative effect in free recall demonstrated here, but also by the negative effect testing had on order memory in Karpicke & Zaromb (2010, Experiment 3).
CHAPTER 6

GENERAL DISCUSSION

Though there has been an abundance of research on the beneficial effects of testing on memory, relatively little research has been conducted to come to a better understanding of exactly why this is the case. A consideration of seemingly similar memory phenomena (like the generation effect) and an analysis of the results of Karpicke & Zaromb (2010) suggest that the testing effect might fit well within an established framework, the item-specific vs. relational account, which has been offered to explain a variety of memory phenomena including the generation effect. This theory predicts that if the underlying explanation of these effects is the same, experimental designs which yield a negative effect of generation (worse memory performance following self generation) should also yield a negative effect of testing. The results of Experiments 3 and 4 unequivocally support this idea: a design which yielded a negative effect of generation yielded a qualitatively similar negative effect of testing. That is, participants who retrieved previously presented targets during Phase 2 recalled fewer of those targets on a final free recall test compared to a control group who simply restudied the cue-target pairs. This is the first demonstration of a true negative testing effect- that recalling some information at Time 1 inhibited participants’ ability to recall that same information later at Time 2 (more on this later).

Though the results of Experiments 3 and 4 are clear in their support of this idea, getting to a point where this theory could be appropriately tested proved more challenging
than anticipated. Experiment 1 demonstrated a positive testing effect, but the null effect of Experiment 2 severely limited the extent to which those data could be interpreted. In light of Experiments 3 & 4, it is prudent conclude that the original design simply was not sensitive enough to demonstrate any negative effect. This was likely due to the fact that the design of the first two experiments made it quite difficult to pick up on the relational nature of the stimuli. The modifications made to Experiments 3 & 4, including longer stimulus duration and explicit instructions to focus on the target, significantly increased both awareness of the categories and clustering based on the categories. This in turn significantly enhanced overall recall. With this new design which rendered identification of the relational structure easier, the expected negative generation effect emerged. This same design yielded a qualitatively similar negative testing effect.

The results from Experiment 1, however, are useful in one sense. When considering the negative effect of testing, one may contend that at short delays, testing often does not improve memory. Indeed, when participants are not provided with feedback, tests typically do not immediately enhance memory (e.g. Roediger & Karpicke 2006b). This is because at short delays, participants in the retrieval practice condition are at a severe disadvantage: if they fail to recall an item during Phase 2, they do not get the additional exposure to the material as the restudy controls do. Given this, it is not surprising that following short delays, studies without feedback fail to demonstrate any benefits of prior testing. Of course at longer delays, the forgetting rates for participants who engaged in retrieval practice are much smaller, the crux of what makes tests special from a memory standpoint.

Three points help mitigate the concern of an insufficient delay. (1) With feedback, positive testing effects typically emerge immediately, even at the shortest delays (e.g. Carrier
& Pashler, 1992; Butler et al., 2008; Marsh et al., 2009). In these designs participants get the benefits of retrieval, yet also are provided with the information in the event of retrieval failure. (2) Participants in Experiment 4 retrieved significantly more targets during Phase 2 (63%) relative to participants in Experiment 1 (46%), $t(43) = 3.99$, $p < .01$. Despite this Experiment 4 demonstrated a negative effect of testing while Experiment 1 demonstrated a positive effect of testing. If one were to be concerned that the failure to demonstrate an effect in Experiment 4 was in some way due to the combination of a short delay and inadequate exposure to the material (the typical recipe for a null effect of testing), this data pattern helps argue otherwise. (3) Experiment 1 yielded a significant, positive effect of testing after a delay of only 5 min.\footnote{While there was no distractor phase (and thus no delay) in Experiment 4, the retention intervals for Experiments 1 & 4 are comparable because the presentation rate in Experiment 4 was slower. In Experiment 1 the average retention interval between a given item in Phase 2 and the free recall test was 6 min, 40 sec while for Experiment 4 it was 4 min, 23 sec.} This is the strongest point to this effect: the same materials as those used in Experiment 4 produced a significant testing effect in Experiment 1. The effect itself, of course, is difficult to interpret. As detailed earlier, the changes made to Experiments 3 & 4 were instituted to emphasize inter-item relational processing (i.e. make the categories more salient). Given this, it is reasonable to assume that in Experiments 1 & 2, when inter-item processing was not accentuated, the benefits of item-specific processing won out. This is evidenced by a significant testing effect in Experiment 1 and the numerical trend of a positive generation effect in Experiment 2. This analysis, of course, is merely speculation and very post hoc. Regardless, the presence of the effect is meaningful: the materials and basic design of the experiments were sensitive enough to show a positive effect had there been one. In sum, the lack of a positive testing effect in Experiment 4 should not be attributed to an inadequate delay (more on the negative testing effect and delays in Future}
Though the first two experiments are limited in the extent that they can inform our understanding of the testing effect, Experiments 3 & 4, conversely, are much more interpretable. Memory performance in the two studies was explicitly compared via a 2 (Experiment 3 vs. Experiment 4) × 2 (read/restudy vs. generate/retrieval) ANOVA. The analysis revealed a highly significant main effect of encoding $F(1,54) = 7.448$, $p=.01$ with no hint of an interaction $F(1,54) = .12$, $p=.74$. This analysis establishes that in each experiment participants in the control conditions recalled significantly more targets than those in the generate/retrieval conditions. More importantly, it substantiates the notion that the pattern of results was the same in both experiments. Given the marginal negative generation effect in Experiment 3, this is especially informative. The analysis suggests the pattern was qualitatively the same, though the effect just happened to fall short of the significance threshold in Experiment 3. Across all four experiments, the generate and retrieval conditions appear remarkably similar lending further credence to the notion that both should be interpreted within the same framework.

Though these data clearly support one hypothesis, namely that testing may be better understood within the item-specific vs. relational framework, it is important to consider the implications with other theories as well. Because taking a test is a complex task, it is reasonable to posit that there may be multiple avenues through which testing improves memory. Accordingly, finding support for one theory does not necessarily preclude other explanations. For example, the data here do not argue that tests cannot insulate against the buildup of interference from previously studied material (i.e. the release from proactive interference hypothesis, Szpunar et al., 2008). These data can, however, speak to the
Transfer Appropriate Processing account. As previously reviewed, TAP is generally cited as an important mechanism to account for the benefits of retrieval practice: retrieving information at Time 1 helps one to better retrieve the same information again at Time 2. This endorsement comes in spite of the fact that more fine-grained predictions borne out of the account are regularly falsified. Specifically, TAP predicts that the greater the match between tests at Time 1 and Time 2, the greater the memory gains should be. This would suggest, then, that if the final test were multiple-choice, for optimal gains, the initial test should also be multiple-choice, as this provides the best match with regards to test type. Contrary to this prediction, research has consistently demonstrated that regardless of the final test format, retrieval practice with open-ended tests leads to the greatest memory gains (e.g. Carpenter & DeLosh, 2006; Roediger & Karpicke, 2006a).

Interestingly, the results of Experiment 4 demonstrate just the opposite pattern. At a general level, the TAP account is clearly not supported: taking a test during Phase 2 led to worse performance on a test during Phase 3. The data do, however, fit better in a more fine-grained analysis. During Phase 2, the information practiced, the cue-target relations, did not match well with the information needed to be successful during Phase 3, the inter-item relations. Because of this mismatch, TAP can account for the negative effect. That the theory is so malleable is concerning because it is unclear in what way the account may be falsifiable. Regardless, it is helpful to consider other theories when interpreting results like these.

*The negative effects of retrieval on later memory performance*

At a more general level, it is not unique to demonstrate that retrieval can inhibit some later aspect of memory performance. Consider the basic retrieval induced forgetting
paradigm (Anderson, Bjork, and Bjork, 1994). In a typical study participants will be presented with several category – exemplar word pairs to study (e.g. fruit – orange, fruit – banana, tool – hammer, tool – wrench). During a retrieval practice phase, certain categories are practiced and others are not. Categories for which some of the exemplars receive retrieval practice are called Rp categories, whereas those categories which receive no practice are called Nrp categories.

During the retrieval practice phase a participant may see fruit – or____. Orange in this case would be an example of an Rp+ item; the category of fruits was one which was a part of the retrieval practice (Rp) phase, and it is denoted with a + because that item in particular, orange, was one which was practiced. Conversely, banana was not included during the retrieval practice phase, so banana would be denoted as an Rp- item; an item for which the category was part of the retrieval practice phase, but the individual item itself was not practiced. Finally lets assume that the category tool is not included at all in the retrieval practice phase. All of these pairs then are denoted as Nrp.

Later during a final cued recall test participants are presented with each of the categories (fruits, tools, etc.) with instructions to recall any of the exemplars presented/recalled at any point of the experiment. Not surprisingly, participants recall more of the Rp+ items than the Rp- items. Interestingly, however, participants recall fewer Rp-items compared with Nrp items. In other words, the participant above would be more likely to recall hammer or wrench compared with banana even though the category of fruits was practiced and tools was not. The common explanation offered for this phenomenon is one of inhibition (Anderson & Spellman, 1995). When recalling orange during the retrieval practice phase you must inhibit competitors, such as the semantically related banana. This
inhibition persists long enough such that during the final cued recall test, banana is still inhibited and therefore less likely to be recalled than exemplars from a category which was not practiced at all.

Now consider retrieval induced forgetting with the negative effect of retrieval practice in Experiment 4. In both cases retrieving information during an intermediary phase results in poorer performance on a later memory test. Does that mean that the negative effect in Experiment 4 should be labeled as retrieval induced forgetting? The answer is clearly no because the two memory phenomena differ along both methodological and theoretical dimensions in very important ways. First consider the methodological differences. With retrieval induced forgetting, while it is true that retrieval practice does lead to a negative effect on memory, importantly, it is isolated to Rp- items; those items in particular which are not practiced. The actual items which are practiced, the Rp+ items, do quite well, demonstrating the typical positive testing effect compared to non-practiced (Nrp) items. In Phase 2 of Experiment 4, all of the targets are subject to retrieval practice and yet they are all recalled poorly on the final recall test relative to controls. Just as important are the theoretical differences in how researchers account for the effects. Participants remember fewer Rp- items because successful practice with the Rp+ items requires inhibition of the Rp-items. In the current study there is no reason to expect inhibition for the target items because they are actually recalled during Phase 2. The inhibitory account would predict that these items are not inhibited at all, so there would be no basis to predict a negative testing effect for these later items.

Even with a specific focus on studies of retrieval practice, it may seem, contrary to the prior the assertion, that this is not the first negative effect of testing. Indeed authors have
labeled certain consequences of testing as ‘negative testing effects.’ Such effects have been demonstrated with multiple-choice tests (Roediger & Marsh 2005), in studies of eye-witness memory (Chan et al., 2009), and in tests of order reconstruction (Karpicke & Zaromb, 2010). A closer examination, however, reveals important differences in the negative effect demonstrated here with those previously documented. Roediger & Marsh (2005) and Chan et al., (2009) both demonstrated that inaccurately recalling information on an earlier test (an incorrect response on a multiple-choice test or retrieving misinformation about a previously witnessed event) increased the likelihood of again recalling that information during a later test. This is certainly a negative consequence of tests, but not a negative testing effect per se. In both cases retrieval of the incorrect information resulted in a greater likelihood to retrieve that same (mis)information again. This is the basic testing effect— it just so happens that the information was inaccurately recalled to begin with. Conversely, Karpicke & Zaromb (2010) demonstrated that retrieving targets in a cued recall test during Phase 2 impaired participants’ ability to accurately reconstruct the original presentation order of the targets from Phase 1. Though this more closely approaches a negative testing effect as defined here, again, it is not the same. In their study participants recalled the targets themselves during Phase 2, not the order of said targets. Because the order of the targets was never a focus of the retrieval practice, it is less surprising that order reconstruction during Phase 3 was impaired. Indeed, in all of Karpicke & Zaromb’s (2010) experiments entailing item memory, the items retrieved in the practice phase were better recalled on the final test. By contrast, in the current study, participants who recalled targets during Phase 2 recalled fewer of those very same targets during Phase 3 compared to controls. This is the purest definition of a negative testing effect, and is a unique and theoretically informative demonstration.
As detailed earlier a preliminary investigation of the Experiment 4 data halfway through collection revealed something quite surprising: participants in Experiment 4 were presented with the list of words to learn twice yet were recalling about 13% fewer targets than participants in Experiment 3 who were presented with the same list of words only once; in other words a negative effect of repetition. This is confirmed by the aforementioned 2×2 ANOVA explicitly comparing the two experiments; there was a highly significant main effect of experiment $F(1,54) = 8.56, p=.01$. To my knowledge this is the first such demonstration: A group of participants presented with information twice (through spaced repetition no less- the lists were separated by a 5 min math distractor) recalling significantly less than a group given a single presentation, a remarkable finding.

As detailed previously there are two possible explanations to account for this pattern. It could be that the scrambled presentation of the cue-target pairs during Phase 1 biased participants so strongly to encode them as unrelated items, that even during the blocked presentation of Phase 2, participants were less able to capitalize on the categorical relatedness of the targets. Alternatively, it could be that this biasing occurred only because participants thought back to Phase 1 (either explicitly as participants in the retrieve condition were instructed to do, or incidentally as the restudy controls would have done).

To test these competing ideas a generation condition was included halfway through data collection of Experiment 4. It was identical in every way to the generation condition of Experiment 3, with the exception that participants in the generation condition of Experiment 4 saw each of the word pairs presented intact (and in a scrambled order) in a separate presentation (Phase 1) prior to the generation task in Phase 2 when the pairs were blocked by
category. Critically, prior to Phase 2, participants were not informed that the word pairs were the same as those presented earlier, and completion of the task was sufficiently easy enough (as indicated by perfect generation performance in Experiment 3) that it did not necessitate thinking back to Phase 1.

If the presence of Phase 1 itself biased participants to encode the targets as unrelated during Phase 2, memory performance in the generate group of Experiment 4 should be worse than performance of the generate group in Experiment 3. Conversely, if it is really the act of thinking back to or recalling Phase 1 that was leading to the curious effect observed, then the generate group of Experiment 4 should be at least comparable to (if not better than) generate performance in Experiment 3.

Recall performance from the participants who generated the word pairs in Phase 2 supported the former hypothesis: performance was significantly lower than the group of participants who generated the word pairs in Experiment 3 \( t(54) = 2.47, p=.02 \). These data support the idea that the scrambled nature of Phase 1 so strongly biased participants to think of the stimuli as unrelated, that even during Phase 2 when the targets were presented in categorical blocks, they had trouble noticing or effectively using the categorical structure. This idea is supported by the item-specific vs. relational account. During Phase 1, all participants, regardless of condition, were presented with intact cue-target pairs. Because encoding required only basic reading, attentional resources were available both for cue-target processing and inter-item processing. Because the pairs were scrambled, unless a participant was particularly adept at picking up on the scrambled structure, analyzing the relations between targets (especially targets presented concurrently which never belonged to the same category) would lead the participant to conclude the targets were simply random words.
without association. When participants were presented with the pairs again in Phase 2, though the targets were now blocked by category, the ‘random words’ notion from earlier inhibited participants’ ability to notice the structure.

There are two major limitations to the analysis above. (1) The explanation is entirely post hoc. The difference in memory performance between the read controls and restudy controls, while theoretically explainable, was entirely unanticipated. (2) The comparisons being made are cross-experimental. The two groups being compared are two groups from two different experiments, and so consequently there was no random assignment to condition. That said, all the data were collected from the same population of students over a single 8-week period. In other words, there is no reason to suspect that the participants in the two conditions differed in any appreciable way.

Nevertheless, this point would be unquestionably more compelling if demonstrated in a single experiment with random assignment and clear a priori hypotheses. Data collection for this very study is currently under way. Participants are being presented with the same word pairs in one of two conditions. The read condition is structured identically to the read condition of Experiment 3: participants are being presented with the word pairs once blocked by category. The restudy condition is structured identically to the restudy condition of Experiment 4: participants see the pairs once in Phase 1 in a scrambled order and again in Phase 2 in a blocked order.

Assuming the pattern of results holds from the cross-experimental comparison (i.e. a negative effect of repetition), the logical follow up would be a replication with cued recall rather than free recall. Participants would be presented with each of the cues one at a time, in a random order with the instruction to recall the accompanying target presented earlier.
According to the hypothesis above, the explanation for the negative effect of repetition (like the negative testing effect in Experiment 4) in free recall centers on a failure to identify the categorical membership amongst the different targets. Specifically, the presence of the initial scrambled list hinders participants’ ability to notice the relationships amongst the various targets during the blocked list in Phase 2. Awareness of this categorical structure is crucial in free recall because it can serve as a powerful retrieval cue. Such a retrieval cue, however, would be substantially less useful in a cued recall test. Here performance would be solely influenced by the degree to which you have associated the cue to the target, and more exposure (i.e. through an additional presentation) should serve to facilitate performance. In this regard the additional presentation should lead to better memory in a cued-recall test, a reversal of the expected pattern from the experiment currently being run. Such a demonstration would be an interesting and theoretically cohesive package of experiments.

Educational implications

The results here provide an explicit demonstration that in some contexts retrieval can actually hurt memory performance. What then, does that mean for students in the classroom? Although prior research has shown retrieval may sometimes hurt memory related to that which is tested, current conceptualizations of retrieval practice indicate that memory for the material which is practiced is always enhanced (e.g. Roediger & Karpicke, 2006b). Taken together, the results of Experiment 4 presented here and Karpicke & Zaromb (2010, Experiment 3) suggest this may be too broad an endorsement.

Consider the results of Experiment 4. A negative effect of testing in free recall is novel, but certainly cannot be taken as evidence that tests are not generally beneficial in this context. This experiment was designed to test a theory; the point was not to mimic an
ecologically relevant learning situation, rather it was to understand how and why retrieval improves later memory. The implications to be drawn from this, however, do have bearing on everyday learning situations: tests may hinder one’s ability to make relations among items. This idea is explicitly borne out in Karpicke and Zaromb (2010, Experiment 3) where testing led to worse performance on a order reconstruction test relative to a restudy control. This finding touches on perhaps the most likely scenario where tests may be detrimental: concepts where the order information itself is important and meaningful. Consider a student in an introductory biology class who is learning glycolysis, a multi-step process which converts sugar into energy. Here order is especially important because later steps cannot logically take place unless and until earlier steps have already taken place. Failure to recall order in this case prohibits an understanding of the basic process. Importantly, however, because the ordering of steps is not arbitrary, the structure of the material constrains the order of the steps, a component not found in arbitrarily ordered word lists. As such, this remains an open, empirical question and an important venue for additional research (see Future Directions).

*Future Directions*

Research on testing over the past 20 years has been disproportionately focused on understanding what happens when we take a test. As a result the mechanism underlying testing has been poorly understood. The results here offer one of the first empirically tested and supported hypotheses to understand exactly why tests improve memory. However, additional research aimed at understanding the mechanism is important both specifically with respect to the theory offered here and at a more general level.

With respect to the experiments reported here, a useful follow up would be to
replicate the basic design of Experiment 4 with different types of tests. This reasoning is similar to that outlined in *Memory performance in Experiment 3 vs. Experiment 4*. The explanation for the negative effect in Experiment 4 was critically dependent upon the properties guiding successful retrieval in a free recall test. When asked to recall the previously presented targets in a free recall test, the categorical nature of the targets served as an extremely useful retrieval cue should the participant pick up on it. According to the item-specific vs. relational account because participants in the retrieval condition focused more on the cue-target relations rather than the inter-item relations, they were less likely to pick up on this structure, and consequently they recalled fewer targets. In a test of recognition or cued-recall, awareness of the categories is not nearly as important. Take, for instance, a cued recall test in which the participant has to recall each target given its rhyming cue. In this case, the relations between the targets and cues, something the tradeoff account predicts would be enhanced, becomes critically important. The account therefore predicts a reversal of the pattern from Experiment 4, namely a positive testing effect. Such a demonstration would offer further support to the theory that the testing effect operates upon the basic tenants of the item-specific vs. relational framework.

Additionally, more thought should be given to what effect, if any, a delay might have on this negative testing effect. Carrier & Pashler (1992) tested participants on half of a previously presented set of word pairs following a 5 min delay and the remainder of the pairs following a 24 hr delay and found no time × recall interaction. In other words, the effects of testing did not qualitatively change with a delay. Accordingly, it is reasonable to predict

15 It is worth mentioning that other studies have found time × recall interactions (e.g. Roediger & Karpicke, 2006b), though critically those were designs which did not employ feedback, and therefore are not an appropriate comparison.
that the current negative testing effect would similarly be insensitive to such a delay. Of course Carrier & Pashler’s (1992) results focused on the positive rather than negative effects of testing, which is why this would be a question worth testing empirically. Importantly, should a positive effect emerge following a longer delay, this would not necessarily indicate qualitative differences with generation. This is because researchers have not looked at the negative generation effect to see what happens following such a delay. For a complete understanding of the negative testing effect itself and how it relates to a negative generation effect, both Experiments 3 & 4 should be replicated following a 24 delay.

As discussed earlier, at a general level, it is important to approach the understanding of testing with the idea that tests may improve memory through multiple routes. Because one explanation need not preclude others, our understanding of testing may be enhanced not only through an examination of the theories already posited, but also through the exploration of new ideas. Recently Pyc & Rawson (2010) demonstrated that tests help improve memory by supporting the effective use of mediators during encoding. Specifically, tests help one to create and better remember novel, self-generated mediators linking cues and targets. This is a novel contribution to our current understanding of tests and highlights this point well: the results are not in conflict with other supported theories, yet provides an additional dimension to our understanding of how and why tests may improve memory.

The issues raised in Educational implications also merit further investigation. Many researchers have argued that tests enhance performance in all learning contexts, though the current study, along with other evidence, suggests this may be too bold a claim. Do tests hinder memory performance for materials, such as glycolysis, which have a meaningful order to them? Though research has demonstrated tests negatively impact order memory (Karpicke
and Zaromb, 2010) the stimuli in the laboratory do not accurately reflect material to be learned in the classroom. In a typical laboratory study order memory is measured through the reconstruction of arbitrarily ordered, unrelated words. A process such as glycolysis, conversely, has a logical ordering of the steps, as earlier steps must necessarily take place before later steps.

Accordingly it is not a foregone conclusion that testing will disrupt order recall in ecologically valid learning contexts. To explore this point further, we can presume that testing would increase the item-specific processing of each of step in isolation. The relations between the steps (i.e. inter-item relational processing) may be disrupted, but this perhaps is of no consequence. With a solid recollection of each of the steps in isolation, given the steps can only proceed in one (or very few) logical order(s), a participant may be able to construct the appropriate order after the fact. Here successful order (re)construction has less to do with inter-item relational processing (and therefore may be unaffected by the disruptions associated with testing) but instead is made possible through reasoning out the logical progression of the steps.

Additionally, it is worth examining what happens when you engage in retrieval practice for the order information itself. The negative effects of testing on order reconstruction in Karpicke & Zaromb’s (2010) study followed recall of the items in a cued recall test. This may be too far removed from how a student may actually study such a process. If the order is that which is recalled during Phase 2, how would that impact order reconstruction in Phase 3? How would this relate to a restudying control?

The reasoning and theories offered here are merely speculative. Accordingly, it is important to empirically test: does testing disrupt order recall for educationally relevant
materials? If it does, this will suggest there are scenarios in which testing is not a beneficial learning strategy. If we are to generally endorse testing as an effective classroom tool, it is important to identify specific conditions (few though they may be) where testing is not beneficial. Conversely, if testing does not disrupt order recall with educationally relevant materials, these data would add to a growing literature which suggests tests serve as a robust learning tool across a variety of situations.
**Figure 1.** Mean proportion of idea units recalled from a prose passage after a 5-min, 2-day, or 1-week retention interval as a function of whether participants studied the passages twice or studied them once before taking an initial test. Error bars represent standard errors of the means. From Roediger and Karpicke (2006b).
Figure 2. Mean final-test performance as a function of intervening task. Error bars show 95% confidence intervals. MC, multiple-choice; SA, short answer. From Kang, McDermott and Roediger, 2007
Figure 3. Proportion of correct responses on the final cued-recall test in Experiment 1 as a function of response confidence and learning condition for responses that were incorrect (left side) and correct (right side) on the initial multiple-choice (MC) test. From Butler, Roediger, and Karpicke, (2008)
Figure 4. Expected effect size for classroom testing as a function of the number of tests given during a semester-long course. From Bangert-Drowns, Kulik, and Kulik (1991).
**Figure 5.** Unit exam performance on quizzed versus nonquizzed items, collapsed across units, in a “Brain and Behavior” course. Error bars show 95% confidence intervals. MC, multiple-choice; SA, short answer. From McDaniel, Anderson, Derbish, and Morrisette (2007)
Figure 6. Mean initial and final List 5 recall performance in Experiment 1A as a function of condition. Error Bars display standard errors of the mean. From Szpunar, McDermott, and Roediger (2008).
Figure 7. Proportion of items recalled on a final free recall test, Phase 2 manipulated between subjects (Experiment 1). Adapted from Karpicke and Zaromb (2010).
Figure 8. Proportion of items recalled on a final free recall test, Phase 2 manipulated within subjects (Experiment 2). Adapted from Karpicke & Zaromb (2010).
**Figure 9.** A comparison of the generation effect (proportion of recalled items previously generated – proportion of recalled items previously read) and testing effect (proportion of recalled items previously tested – proportion of recalled items previously read) in Experiments 1 & 2. Adapted from Karpicke & Zaromb, (2010).
**Figure 10.** Order reconstruction performance for the Read, Generate, and Recall conditions in Experiment 3 plotted as a function of serial position. From Karpicke & Zaromb (2010).
Figure 11. Proportion of objects recalled on a final free recall test (Experiment 2). Adapted from Steffens (1999).
Figure 12. Percentage of action phrases recalled and percentage of action phrases recalled in the correct order (Experiment 2). Adapted from Steffens (2007).
**Figure 13:** Mean proportion of targets recalled in Experiments 1-4. Error bars are standard errors of the mean.
Figure 14: Proportion of participants aware of the categorical relationship amongst the targets. Error bars are standard errors of the mean.
Figure 15: Mean Adjusted Ratio of Clustering (ARC) scores for Experiments 1-4. Error bars are standard errors of the mean.
Figure 16: Distributions of the ARC scores from Experiments 1-4 before (top) and after (bottom) a square root transformation
## APPENDIX 1

CUE-TARGET PAIRS USED IN EXPERIMENTS 1-4

<table>
<thead>
<tr>
<th><strong>Target:</strong> A metal</th>
<th><strong>Target:</strong> A part of the Human body</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feel Steel</td>
<td>Beg Leg</td>
</tr>
<tr>
<td>Cold Gold</td>
<td>Linger Finger</td>
</tr>
<tr>
<td>Win Tin</td>
<td>Bread Head</td>
</tr>
<tr>
<td>Pickle Nickel</td>
<td>Doe Toe</td>
</tr>
<tr>
<td>Class Brass</td>
<td>Sand Hand</td>
</tr>
<tr>
<td>Think Zinc</td>
<td>Hose Nose</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Target:</strong> A four-footed animal</th>
<th><strong>Target:</strong> A fruit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Force Horse</td>
<td>Tape Grape</td>
</tr>
<tr>
<td>Swear Bear</td>
<td>Wear Pear</td>
</tr>
<tr>
<td>Vow Cow</td>
<td>Teach Peach</td>
</tr>
<tr>
<td>Cheer Deer</td>
<td>Drum Plum</td>
</tr>
<tr>
<td>Blouse Mouse</td>
<td>Time Lime</td>
</tr>
<tr>
<td>Rig Pig</td>
<td>Tune Prune</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Target:</strong> A kitchen-utensil</th>
<th><strong>Target:</strong> A Transportation Vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wife Knife</td>
<td>Jar Car</td>
</tr>
<tr>
<td>Cork Fork</td>
<td>Plus Bus</td>
</tr>
<tr>
<td>Moon Spoon</td>
<td>Puck Truck</td>
</tr>
<tr>
<td>Ban Pan</td>
<td>Cane Plane</td>
</tr>
<tr>
<td>Disk Whisk</td>
<td>Hike Bike</td>
</tr>
<tr>
<td>Skate Plate</td>
<td>Coat Boat</td>
</tr>
</tbody>
</table>
REFERENCES


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*(8), 873-885.


