Effect of Multiple-Choice Testing on Memory Retention –
Cue-Target Symmetry

by

Cho Kin Cheng

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Department of Psychology
University of Toronto

© Copyright by Cho Kin Cheng (2014)
Effect of Multiple-Choice Testing on Memory Retention – Cue-Target Symmetry

Cho Kin Cheng
Doctor of Philosophy
Department of Psychology
University of Toronto
2014

Abstract

This thesis investigated the testing effect, which is the tendency of testing to enhance learning and memory retention. Specifically, the thesis examines the extent to which test-induced retention benefits extend to the questions as well as the answers; are these benefits symmetrical? The results in the laboratory experiment (Experiment 1) demonstrated that the symmetry in retention benefits is highly dependent on the accuracy in the initial test. For the items students answered correctly in the initial test, the retention benefit was found to be symmetrical across the questions and answers supporting the retrieval hypothesis of the testing effect. However, for initially incorrect items, the retention performance was substantially better for the questions than the answers suggesting an asymmetrical effect from testing, a result that is also consistent with the retrieval hypothesis. These results were replicated using educationally-relevant materials with a delay up to five and a half months in two hybrid studies combining actual classroom experiences with controlled assessments in the laboratory (Experiment 2A & 2B).

A series of follow-up studies were conducted to examine the difference in retention performance of the questions and the answers for the initial incorrect items. Particularly, the experiments focused on investigating various factors that might contribute to or eliminate the difference in
retention performance, including prior study (Experiment 3), verification feedback (Experiment 4A & 4B) and answer feedback (Experiment 5).

The thesis also touched upon various issues related to the temporal dimension of the testing effect. The previous notion that testing slows down forgetting rate is not supported by the data. Instead, testing seems to provide a short-term insulation against immediate forgetting, but then memory for the tested materials decays in the same way as the non-tested materials.
Acknowledgments

This work would not be possible without a number of people who have helped me and supported me throughout my graduate studies. I want to give my sincere gratitude to each and every one of them.

I would like to thank my supervisor, Steve Joordens. He took me on halfway through the program. He has opened up numerous opportunities for me. I am grateful for the trust he has given me. I am also thankful for all the time he spent fostering my professional and personal development. He is an amazing supervisor.

I want to thank my committee member who is also my mentor, Douglas Bors. His passion with statistics has inspired me enormously. His teaching has reshaped my understanding and appreciation of statistics and research method.

I want to thank George Cree. His ideas and insights have provided me a better perspective of my work.

I also want to thank my previous supervisor, Mark Schmuckler. Although I did not finish my dissertation in his lab, his mentorship has been invaluable to my research career. I would not have become the researcher I am without his years of guidance.

I have also had the good fortune to find myself surrounded by a truly special group of colleagues and friends, whose impacts cannot be understated. They have walked with me through ups and downs over the years. These individuals include, Sijing Wu, Dwayne Paré, Diane Mangalindan, Gelareh Jowkar, & Lisa-Marie Collimore.
# Table of Contents

Acknowledgments .......................................................................................................................... iv  
Table of Contents ............................................................................................................................ v  
List of Figures ................................................................................................................................... viii  
Chapter 1 Introduction ...................................................................................................................... 1  
  1 Classic studies ............................................................................................................................ 3  
  2 More recent studies .................................................................................................................... 7  
  3 Testing versus Re-studying ...................................................................................................... 10  
  4 Performance on a retrieval test ................................................................................................. 13  
  5 Relevant issues ......................................................................................................................... 16  
    5.1 Transfer of knowledge ...................................................................................................... 16  
    5.2 Negative effects of testing ............................................................................................... 21  
    5.3 Feedback ........................................................................................................................ 25  
  6 Summary and Roadmap ........................................................................................................... 30  
Chapter 2 Testing Effect and Cue-Target Symmetry ................................................................... 31  
  1 Experiment 1 ............................................................................................................................ 32  
    1.1 Method .............................................................................................................................. 34  
      1.1.1 Participants ............................................................................................................. 34  
      1.1.2 Materials ................................................................................................................. 34  
      1.1.3 Procedure .............................................................................................................. 35  
    1.2 Results ............................................................................................................................... 36  
    1.3 Discussion ......................................................................................................................... 41  
  2 Experiment 2A ............................................................................................................................ 44  
    2.1 Method ............................................................................................................................... 46  
      2.1.1 Participants ............................................................................................................. 46
2.1.2 Materials ............................................................................................................... 46
2.1.3 Procedure .............................................................................................................. 47
2.2 Results ............................................................................................................................... 47
2.3 Discussion ......................................................................................................................... 55
3 Experiment 2B ......................................................................................................................... 59
  3.1 Method .............................................................................................................................. 61
    3.1.1 Participants ............................................................................................................ 61
    3.1.2 Materials ............................................................................................................... 61
    3.1.3 Procedure .............................................................................................................. 61
  3.2 Results ............................................................................................................................... 61
  3.3 Discussion ......................................................................................................................... 66
Chapter 3 Cue-Target Asymmetry and Prior Study ................................................................. 69
  1 Experiment 3 ............................................................................................................................ 70
    1.1 Method .............................................................................................................................. 71
      1.1.1 Participants ............................................................................................................ 71
      1.1.2 Materials ............................................................................................................... 71
      1.1.3 Procedure .............................................................................................................. 71
    1.2 Results ............................................................................................................................... 72
    1.3 Discussion ......................................................................................................................... 76
Chapter 4 Cue-Target Asymmetry and Feedback ....................................................................... 79
  1 Experiment 4A ......................................................................................................................... 79
    1.1 Method .............................................................................................................................. 80
      1.1.1 Participants ............................................................................................................ 80
      1.1.2 Materials ............................................................................................................... 81
      1.1.3 Procedure .............................................................................................................. 81
    1.2 Results ............................................................................................................................... 81
List of Figures

Figure 1. Spitzer (1939): Mean proportion correct for the multiple-choice tests taken at the various delays after studying. Adopted from Roediger and Karpicke (2006a).

Figure 2. Experiment 1: Mean proportion recalled for the four experimental conditions in the cued-recall task.

Figure 3. Experiment 1: Mean proportion recalled for the two cueing directions for the initially correct and incorrect items in the cued-recall task.

Figure 4. Experiment 2A: Mean proportion correct for the three experimental conditions in the experiment task.

Figure 5. Experiment 2A: Mean proportion correct of the verbatim and modified conditions for the initially correct and incorrect items in the experimental task.

Figure 6. Experiment 2A: Regression lines with proportion correct as the function of days elapsed since the midterm for the three experimental conditions in the experimental task. (Beyond the zigzag line: mean proportion correct for the three experimental conditions in the experimental task in Experiment 2B).

Figure 7. Experiment 2B: Mean proportion correct for the three experimental conditions in the experiment task.

Figure 8. Experiment 2B: Mean proportion correct of the verbatim and modified conditions for the initially correct and incorrect items in the experimental task.
Figure 9. Experiment 3: Mean proportion recalled for the two cueing directions for the no study and study conditions separately for the initially correct and incorrect items in the cued-recall task.

Figure 10. Experiment 4A: Mean proportion recalled for the two cueing directions for the no feedback and feedback conditions separately for the initially correct and incorrect items in the cued-recall task.

Figure 11. Experiment 4B: Mean proportion recalled for the two cueing directions for the no feedback and feedback conditions separately for the initially correct and incorrect items in the cued-recall task.

Figure 12. Experiment 5: Mean proportion recalled for the two cueing directions for the no feedback and feedback conditions separately for the initially correct and incorrect items in the cued-recall task.
Chapter 1
Introduction

Public education systems are experiencing great challenges as student-to-faculty ratios grow while budgets shrink. It is not a stretch to say that educators are being challenged to do more with less, which is prompting many educational researchers to reconsider existing practices with the goal of finding ways of enhancing learning. For instance, some researchers are examining how lectures are developed and considering alternatives (e.g., Aguaded-Gomez, 2013; Bassili & Joordens, 2008; Kikkas, Laanpere, & Poldoja, 2011; Wilson, 2013), others are creating new technologies intended to maximize active learning without increasing costs (e.g., Muntean, 2011; Paré & Joordens, 2008), and still others are considering the efficacy of new methods of connecting faculty and students (e.g., Cheng, Pare, Collimore, & Joordens, 2011; Guzdial & Turns, 2000; Rau, Gao, & Wu, 2008). This thesis will focus on another critical aspect of the educational experience, testing.

Testing is an integral part of modern education. Tests of various forms are frequently used across educational levels and disciplines to assess students’ progress and mastery of skills and knowledge. Students are constantly tested to allow comparison of performance with their peers or with some pre-determined standards. Aside from classroom testing, students also take standardized tests, such as SAT, GRE, and MCAT, and various professional licensing and certification tests. Performance on these tests are the basis for important social economical decisions including school admission, scholarship consideration, employment decision, etc. The idea of using tests as assessment devices is profoundly engraved in most people’s mind. In educational circles, this perspective is sometimes described as summative learning or as representing an assessment “of learning” perspective (Bennett, 2011; Roos & Hamilton, 2005).
Most clearly stated, this perspective highlights tests as devices used to measure previous learning. Aligned with this perspective, research on testing has been heavily focused on issues such as fairness and validity of tests (e.g., Sackett, Borneman, & Borneman, 2008; Tittle, 1975; Xi, 2010).

Although the values of using tests as assessment devices cannot be understated, they reflect only one facet of testing. It is crucial to recognize that tests do not only take static snapshots of students’ memory, tests also enhance learning and change the nature of students’ memory (Marsh, Roediger, Bjork, & Bjork, 2007). That is, the learning process does not shut down during a testing experience; it remains engaged. In fact, students are seldom more engaged than they are when they are being assessed, and research has shown that testing can foster future retention of the tested materials. In fact, the learning benefits of testing are often greater than if an equivalent amount of time is spent on re-studying the materials, and this holds true even when the performance on the test is far from perfect and no feedback is given. This phenomenon of improved retention performance through prior testing is referred as the testing effect (Roediger & Karpicke, 2006a), and it falls under a more general perspective of testing as formative learning or assessment “for learning” (Bennett, 2011; Roos & Hamilton, 2005). This powerful utility of testing has often been overlooked and under-utilized. While the main goal of this thesis is to examine the theoretical principles and underlying cognitive mechanisms of the testing effect, this thesis also sheds light on how to effectively use a test in a formative manner.

This thesis focused on the testing effect in the context of multiple-choice tests. The multiple-choice test format has been heavily used in modern education, especially in higher education. The prevalence of its use surpasses that of all other test formats (Hassmen & Hunt, 1994). It is
objective, versatile, and reliable (Aiken, 1987), which makes it an efficient way of testing, particularly for large classes.

This thesis integrates and replicates previous findings on the testing effect while it also examines a number of issues that either have not been examined before or have not been resolved. The project includes both pure laboratory-based studies as well as hybrid studies combining actual classroom experiences with controlled assessments in the laboratory. In order to properly frame the thesis research, previous relevant findings from the existing literature are first reviewed.

The remainder of this introduction is organized into two sections. First, I provide a review of the existing empirical data related to the testing effect, highlighting changes in design that occurred over time. Then, I highlight three issues related to the testing effect that have direct relevance to my research: the extent to which knowledge is transferred to novel contexts, the negative effects of testing, and the effect of feedback with respect to promoting memory retention and reducing the negative effects of testing.

1 Classic studies

The basic notion of a testing effect has a long history. Over 2000 years ago, Aristotle wrote in On Memory, “Exercise in repeatedly recalling a thing strengthens the memory”. William James (1890) also discussed about the phenomenon in his book, The Principles of Psychology, over a century ago:

A curious peculiarity of our memory is that things are impressed better by active than by passive repetition. I mean that in learning (by heart, for example), when we almost know the piece, it pays better to wait and recollect by an effort within, than to look at the book
again. If we recover the words the former way, we shall probably know them the next
time; if in the latter way, we shall likely need the book once more.

Abbott (1909) was the first investigator to confirm the testing effect through experimentation. A few other researchers also found empirical evidence of the phenomenon subsequently in the early twentieth century (Gates, 1917; Jones, 1923; Sones & Stroud, 1940; Spitzer, 1939; Thorndike, 1914). In the following section, two of the most influential classic studies, landmark studies by Gates (1917) and Spitzer (1939), are discussed.

Gates (1917) instructed young children across a range of grades to study two types of materials, nonsense syllables and brief biographies. The children first went through a study phase composed of two parts: in the first part, the children simply read the materials on their own, and then in the second part, they were instructed to look away from the materials and try to do a self-test, reciting the materials to themselves. During the self-test, or recitation, they were allowed to go back to the materials if they needed to refresh their memories. The proportion of time the children spent on studying and recitation were manipulated. They were given between 0 – 90% of the total time for recitation. That is, some children spent a lot of time studying and only a little time (say 10%) reciting, whereas others spent a little time studying and most of the time (say 90%) reciting, with other children falling at other points on this recitation continuum.

After the study phase, the children in Gates (1917) study were asked to recall as many items as they could and were tested again after a 3- to 4-hour delay. The results showed robust positive effects of recitation in almost all conditions except for one condition that involved first-graders. In general, the more recitation students performed, the better they recalled the items, even though more recitation meant less study time. However, the results also showed that, for the brief biographies, the effect of testing leveled off and even appeared to drop when the proportion of
time spend on recitation exceeded 60%, suggesting that some amount of study may be needed to sustain an optimal learning outcome. Gates concluded that recall attempts during learning are a good way to promote learning.

In another classic study, Spitzer (1939) tested 3605 six-grade students from nine cities within the state of Iowa. The students read 600-word articles that were relatively novel to them but were similar to materials they would read in school. Then they were tested with multiple-choice tests according to various schedules. The schedules differed in terms of when the tests were originally and subsequently administrated across the two months after students had studied the materials. These schedules were constructed such that different groups of students would be tested after the same delay, but for some of these groups this delay might be the first time they were tested whereas for others it was the second or third time they were tested. These comparisons allowed the researcher to assess the effect of prior testing on subsequent memory for the materials.

Several interesting findings were discovered in the Spitzer (1939) study, as presented in Figure 1. First, Spitzer found that the longer the first test was delayed, the worse was the performance on that test. The results showed that the drop in the performance was most rapid at the beginning after short delays but then it slowed down gradually until hitting asymptote at around three weeks, suggesting that the forgetting rate was an inverse function of time. Second, Spitzer found that taking a test dramatically slowed down the rate of forgetting; when students were given an initial test and then retested at a later time, their performance on the second test did not drop much from the first test at all. This is the basic testing effect, and it suggests that learning was somehow “set” at a specific, and generally high, level by the initial test. Third, because of the shape of the forgetting curve and the effect of testing on subsequent memory, the sooner the initial test was given after finished studying, the better students performed on the later tests.
When students took their initial test after three weeks, the test did not enhance later performance at all. These results together suggest that testing enhanced memory retention of the materials by slowing down the rate of forgetting and it is effective as long as the initial retrieval test was administrated when students can still recall or recognize the materials with a relatively high level of accuracy.

Figure 1. Spitzer (1939): Mean proportion correct for the multiple-choice tests taken at the various delays after studying. Adopted from Roediger and Karpicke (2006a).

Although the results of these early studies on the retention benefits of testing were compelling, interest in the testing effect faded because of the rise of the interference theory and interest in the study of forgetting (Roediger & Karpicke, 2006a). In the context of research focusing on
forgetting, repeated testing was often seen as an empirical confound, and therefore research on
the testing effect all but disappeared.

2 More recent studies
A few decades later, interest in the effect of testing was revived when contemporary
psychologists questioned and debated the nature of the learning process in the late 1950s and the
early 1960s (Roediger & Karpicke, 2006a). These psychologists often studied learning by
administering a series of study and test trials in an alternating manner. Learning materials were
presented in the study trials, and tests were given to determine what students had learned. The
study and test trials were usually repeated numerous times, and the performance of all the test
trials was plotted to depict the rate of learning (e.g., Tulving, 1962). The result was a learning
curve with a shape of a log function. The curve illustrated that most learning occurs early on and
the amount of learning decreases over time. One critical assumption underlying this kind of
research is that learning occurs only during study trials when students are exposed to the
materials, and test trials are considered to be merely neutral events that only measure what
students have learned on the previous study trials. Thus, despite the previously described works,
this new paradigm essentially ignored them by assuming that all testing does is summative.

In 1967, Tulving called this summative assumption into question once again. He had students
learn a list of 36 words. The students went through a series of 24 trials, a combination of study
and test trials. The students were presented with the same 36 words in the study trials and were
instructed to free-recall the words in the test trials. The study trials and test trials were
administrated in three different schedule conditions. In the standard condition, study and test
trials alternated for all 24 trials (STST). In the repeated-study condition, students were showed
the materials for three consecutive trials before they were given a test (SSST), and the sequence
repeated for 24 trials. Finally, in the repeated-test condition, students studied the materials once and then were tested three times consecutively in each sequence (STTT).

If learning only occurred in the study trials, then the students in the repeated-study condition should perform better than the students in the other two conditions, while the students in the repeated-test condition should evidence the worst performance. These predictions reflect the total number of study trials students experienced in each condition: the students in the repeated-study condition went through 18 study trials, whereas the students in the standard and repeated-test conditions were only given 12 and 6 study trials respectively. Surprisingly, despite the large difference in the number of study trials, all three conditions produced strikingly similar learning curves. The learning curves for the standard condition and the repeated-study condition were essentially the same. The students in the repeated-test condition seemed to be disadvantaged early on but quickly caught up. After the 24 trials, the students from the repeated-test condition performed only slightly worse than the students from the other two conditions. This slight difference was argued to be a result of the fact that students in the repeated-test condition experienced a longer delay between the final test trial and the study trial prior to that, which deprived them from using their short-term working memory (Glanzer & Cunitz, 1966).

The results from Tulving (1967) again suggested that tests serve a function beyond simply measuring the amount of learning. They also facilitate subsequent recall to a similar extent as spending the same amount of time re-studying the materials. These surprising findings sparked renewed interest in the testing effect. A number of studies successfully replicated the findings with similar methodologies (Birnbaum & Eichner, 1971; Donaldson, 1971; Lachman & Laughery, 1968; Roediger & Karpicke, 2006b; Rosner, 1970), and some studies also extended the findings to different materials, such as word pairs (Izawa, 1966, 1967).
As exemplified by Tulving’s (1967) study, the traditional procedure for studying the testing effect involved multi-trial learning in which students go through numerous repetitions of intertwining study and test trials. Some researchers argued that, in this procedure, the effect of testing could be confounded with the effects of the later study of previously tested items, thereby preventing straightforward interpretations of the results (Carrier & Pashler, 1992; Kuo & Hirshman, 1996; Thompson, Wenger, & Bartling, 1978). These researchers argued that tests can equip students with the knowledge of recall-ability of the tested materials. This knowledge can then be used to guide future encoding of the materials in the subsequent study trials. Students can strategically allocate more or less time and resources on the basis of recall-ability of the materials. To counter this possibility, recent studies have almost completely abandoned the multi-trial learning procedure and have established a relatively standard procedure to assess the effects of testing on memory retention.

More recent studies of the testing effect use paradigms generally in line with the following description. Students are first given some materials to study (in some cases, students are not given the chance to study some or all of the materials, either because it is a manipulation for the assessment of the effect of initial study or because the materials involved general knowledge which students were assumed to have acquired to a certain extent), then they complete one or more retrieval practice test(s), and, after a delay, they are given a final test. The performance difference between groups that have been given a retrieval test(s) and those that have not been given any test, or the performance between materials that have been tested previously and materials that have not been tested, is used to assess the effects of testing on memory retention. Studies usually find that groups tested previously perform better in the subsequent retention task or materials that were tested previously would be more likely to be retained. Using this
procedure, studies have found that prior testing enhanced the future retention of a wide range of materials, including word lists (Wheeler, Ewers, & Buonanno, 2003), word pairs (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005; Runquist, 1983, 1986a, 1986b), picture-word pairs (Glover, 1989; Wheeler et al., 2003), face-name pairs (Carpenter & DeLosh, 2005), maps (Rohrer, Taylor, & Sholar, 2010), general knowledge (e.g., Fazio, Huels, Johnson, & Marsh, 2010; Marsh, Lozito, Umanath, Bjork, & Bjork, 2012; McDaniel & Fisher, 1991), and complex prose materials (e.g., Chan, McDermott, & Roediger, 2006; Duchastel, 1981; Glover, 1989; Karpicke & Blunt, 2011; Marsh, Agarwal, & Roediger, 2009; Roediger & Marsh, 2005).

The results of a growing number of studies have converged and provided ample evidence supporting the beneficial effects of testing. They provide strong support that tests are not neutral events that only assess the level of learning, but that they also have profound impacts on learning and future retention. In the follow section, I will discuss how the effect reacts to various manipulations and describe some of the factors shown to modulate the testing effect.

3 Testing versus Re-studying

In the early studies on the testing effect, researchers often compared a condition in which students studied some materials and then took a delayed retention test with a condition in which students studied, took an initial test, and then took the delayed retention test. The studies usually found that students performed better in the latter condition (e.g., Gates, 1917; Spitzer, 1939). Yet, this research design fails to account for a crucial factor. When an item is retrieved from memory, this action would inevitably result in a re-processing and re-presentation of the item. Thus, it is possible that testing simply provides additional processing and exposure time to the retrieved item, and this might be the reason for the beneficial effects of testing (Thompson et al., 1978). This argument is referred as the amount-of-processing hypothesis. It suggests that the
testing effect is simply a result of additional exposure to materials causing over-learning of the set of materials that can be retrieved. This idea is consistent with the conventional view that retrieval itself does not have any influence on memory. By this logic, retrieving an item in a retrieval test is similar to re-studying it for an additional time and, thus, providing an equivalent amount of time for re-studying should boost retention to the same extent as taking a test. Yet, data from studies that equated exposure time portrayed a more complex pattern.

A number of studies directly compared the effects of testing and re-studying but provided inconsistent results. Some studies found that re-studying lead to better retention than taking a retrieval test(s) (Chan et al., 2006; Hogan & Kintsch, 1971; Thompson et al., 1978), some studies found that the effects of the two are similar and comparable (Hogan & Kintsch, 1971; Wheeler et al., 2003), while others still found an advantage for taking the retrieval test(s) (Carrier & Pashler, 1992; Duchastel & Nungester, 1981; Kuo & Hirshman, 1996; Nungester & Duchastel, 1982; Roediger & Karpicke, 2006b; Thompson et al., 1978; Wheeler et al., 2003). These inconsistent findings suggest that there may be mediating factors that mitigates the effectiveness of a retrieval test on future retention. One of the important factors may be related to the retention interval.

In Hogan and Kintsch (1971) and Thompson et al. (1978), students went through repeated-study (SSSS) or a repeated-test (STTT) learning schedule (one cycle), then they were tested after a short delay (within 20 minutes) and were tested again after two days. The students in the repeated-study condition showed a considerably higher level of retention initially after a short delay. However, after two days, the pattern changed completely; although the students in both conditions retained less materials in general, the students in the repeated-test condition retained more materials than the students in the repeated-study condition in one of the studies (Thompson et al., 1978), and the retention outcomes for the two conditions were equivalent in the other study
(Hogan & Kintsch, 1971). The students in the repeated-study condition seemed to forget the materials at a faster rate than the students in the repeated-test condition.

Wheeler et al. (2003) recognized that the within-subject design employed in Hogan and Kintsch (1971) and Thompson et al. (1978), in which the students were tested in both retention intervals, might not be the optimal research design; the first delayed test might have influenced performance on the second delayed test. Thus, Wheeler et al. (2003) re-examined the issue with a between-subject design. They had students go through a repeated-study procedure or repeated-test procedure, then take the final retention test either at a five-minute or at a two-day retention interval. Their results replicated Hogan and Kintsch (1971). They found that the students in the repeated-study condition performed significantly better than the students in the repeated-test condition at the five-minute retention interval, but the retention difference disappeared after two days. In a follow-up experiment, they extended the retention interval of the delayed test from two days to a week, and they found that the retention performance in the repeated-study condition continued to drop at a faster rate and was inferior to the repeated-test condition after a week. The results were consistent with Thompson et al. (1978).

Roediger and Karpicke (2006b) replicated the Wheeler et al. (2003) findings and further suggested that the number of retrieval tests might also be a contributing factor. They compared the retention outcomes of three learning schedules with different number of study trials and test trials (SSSS, SSST, and STTT) after five minutes and one week. Again, they found that the retention performance of the SSSS condition was better than the STTT condition after a short delay, and the pattern reversed after a long delay. The retention performance for the SSST conditions fell between the SSSS and STTT conditions in both retention intervals.
All these studies attempted to equate exposure to the materials in the repeated-study and repeated-test conditions and were still able to demonstrate the robust advantages of taking tests, thereby emphasizing the fact that taking tests is different from simply re-studying. This is exciting because the usual repeated-study condition, in fact, provided an even greater exposure to the materials than the repeated-test condition given that, in the repeated-test condition, students were only re-exposed to the materials that they could retrieve in the initial test(s). All these findings suggest that the effect of testing is fundamentally different from the effect of studying in terms of their respective effects on memory retention. These studies, along with Spitzer (1939), suggest that taking a retrieval test can change the course of retention and dramatically slow down the rate of forgetting.

4 Performance on a retrieval test

Some researchers have argued that the overall performance level on a retrieval test is another important factor that modulates the effectiveness of testing on memory retention (Carrier & Pashler, 1992; Wenger, Thompson, & Bartling, 1980). Carrier and Pashler (1992) claimed that testing will not contribute to memory retention unless the students succeed in retrieving the materials in the initial test. Studies conducted by Roediger’s research group have reported direct evidence to support this claim (Butler, Karpicke, & Roediger, 2007, 2008; Butler & Roediger, 2008). In their studies, students studied prose materials, then were first tested with a multiple-choice retrieval test, and then again with a cued-recall retention task after a variable delay (five minutes, one day, or a week). The results consistently showed that the performance on the retention task varied greatly dependent on the accuracy on the retrieval multiple-choice test. Students performed substantially better on the items that they got correct in the initial multiple-choice test than the ones that they got wrong. In all of the experiments they reported, the
retention performance for the initially incorrect items was about 10% or less, compared to over 50% for the initially correct items. The large discrepancy of the retention performance for the two types of items, along with the near-baseline performance for the initially incorrect items, provided support for the claim that the retention benefits of testing only apply to the materials that students were able to retrieve initially.

Similar results were found when the researchers experimentally manipulated the performance/accuracy on the retrieval test through various indirect means (Butler, Marsh, Goode, & Roediger, 2006; Butler & Roediger, 2008; Marsh et al., 2009; Roediger & Marsh, 2005). The results of these studies have provided additional supports for the positive relation between the retrieval test performance and the test-induced retention benefits. For instance, Roediger and Marsh (2005) used a multiple-choice format retrieval test, and they varied the difficulty of the retrieval test by manipulating the number of alternatives. They found that as the number of alternatives on the multiple-choice test increased, the performance on the retrieval test and the retention performance in the final cued-recall task (after a 5-minute delay) both decreased. The same pattern of results held even if the final retention task occurred a week later (Butler & Roediger, 2008) or if the final retention task was a free-recall task (Experiment 2 & 3: Butler et al., 2006).

Other studies have manipulated students’ level of knowledge before they take the initial test. For example, in another experimental manipulation of Roediger and Marsh (2005), students were only presented with half of the materials before taking the initial multiple-choice test, essentially creating a study and no study condition. The results showed that lack of prior study led to poor performance in the multiple-choice test and, in turn, led to a lower level of retention in the cued-
recall task. Butler and Roediger (2008) extended these findings showing that providing extra study opportunities would, instead, promote retention compared to a standard study condition.

Likewise, Marsh et al. (2009) found different levels of enhanced retention among students with different education backgrounds while using the same study materials. They found that when administrating the same SATII test to a group of high school students and a group of high-achieving university students, the high school students displayed worse performance in the initial test and showed a lower level of performance in the subsequent retention task compared to their university counterparts. These studies varied a number of factors that always led to the same direction of performance fluctuations in both the initial retrieval test and the subsequent retention task, suggesting the effectiveness of testing on memory retention is mediated by the performance on initial retrieval test.

Testing can alter and slow down the rate of forgetting; however, what can be retained is determined by what can be retrieved during the initial test(s). Together, these two factors provide a simple explanation of the crossover interaction of the retention performance between testing and re-studying over different retention durations. On one hand, re-studying gives students a chance to encode all the materials again, while, on the other hand, testing only allows students to process the materials that they can retrieve. This leads to a superior performance for the re-study condition over the test condition over a short delay. However, the re-studied materials are soon forgotten when the re-study phase is over. In contrary, the memory of tested materials was strengthened and subsequently deteriorates at a much slower rate. After some delay, the amount of the materials that are retained in the re-study condition would drop below that in the test condition creating the interaction that the studies have documented. At what delay the cross happens is dependent, at least partially, on the performance on the initial test.
5 Relevant issues

The current thesis will examine the testing effect in more detail in hopes of providing further clarity with respect to the relevant underlying processes. The specific experiments reported provide data relevant to three other issues associated with the testing effect; the extent to which knowledge can be transferred to novel contexts, the negative effects of testing, and the effects of feedback on promoting memory retention and reducing the negative effects of testing. In this section I describe each of these issues and highlight relevant previous research.

5.1 Transfer of knowledge

Are the retention benefits of testing restricted to retrieved information from the initial test? Also, can the knowledge retained through testing be used flexibly to construct new responses or answer different questions, or be applied to new contexts or circumstances? The testing effect would have limited utility if the effect is restricted to learning a specific response to a specific question in a specific context. As discussed in the previous session, previous research has demonstrated that the retention benefits of testing are transferable, at least, across different test formats. In this section, the issue of transferability would be discussed more broadly and in greater detail.

Some studies have examined whether testing only facilitates the retention of the responses from the initial test or whether the facilitation extends to information in the questions as well. Carpenter, Pashler, and Vul (2006) had students study paired associates (i.e., Cue - Target), and then students were tested initially in the forward manner (Cue → ?). After a delay, the students were given a retention task during which they were either tested with a cued-recall test or a free-recall test. In either test format, they were asked to recall either the targets or the cues from the
initial test. Carpenter et al. (2006) found that prior testing led to better retention for both the targets and the cues compared to re-studying both words of the pairs.

In another study, McDaniel, Anderson, Derbish, and Morrisette (2007) used educationally relevant materials in a classroom setting and found a benefit of initial testing over re-studying on final retention tests in which the required responses were embedded in the questions in the initial test. Specifically, they gave students weekly quizzes based on the course reading materials. In these quizzes, students were either tested, in a multiple-choice format or in a short answer format, or they were presented with statements that corresponded to the items in the quizzes. Subsequently, in the midterm tests and final exam, students were asked to retrieve other words in the statements that were not retrieved in the initial test. The researchers found that prior testing, regardless of the test formats, led to better retention of the non-retrieved embedded words than did re-studying.

Together these studies suggest that students are not simply learning the answers to specific questions as they are tested. Instead, it seems that the entire test experience is being committed to memory with the benefits of testing extending to the question (or cue) that was associated with the answer (or target). Thus, transfer at least applies to the specific questions that prompted the retrieval of the answers.

Other studies have investigated whether information retained through testing can be transferred or applied to novel questions. Rohrer et al. (2010) showed that testing promoted transfer to a final test that consisted of novel questions but for which the required responses remained the same as those prompted in the initial test. Specifically, they had grade school students study two fictional maps. After the study phase, the students took a test on one map, in which they had to match the name of cities with locations, and they re-studied the other map. After a day delay, the
students were given a transfer test consisted of questions in which they were given the name of two cities and had to recall the name of the city that lies between the two. The results showed that taking an initial matching test promoted better transfer performance than re-studying the map despite the change in the specific question being answered.

Butler (2010) further demonstrated that testing can promote transfer to a final test that consisted of novel questions and that required novel responses as well. Butler had students study prose materials and then either repeatedly re-study or take multiple tests on the materials. An example of a test question is the following: “Bats are one of the most prevalent orders of mammals. Approximately how many bat species are there in the world?” (Answer: “More than 1,000 bat species have been identified.”). One week later, they took a final test with new inferential questions which required novel responses. An example of an inferential question is the following: “Approximately what percent of all mammal species are species of bat?” (Answer: “If there are about 5,500 species of mammals and more than 1,000 species of bat, then bats account for approximately 20% of all mammal species.”). Butler found that repeated testing produced superior transfer on the final test compared to repeated studying.

In a follow-up study, Butler (2010) further demonstrated that the effect extended to new inferential questions from different knowledge domains. For example, the initial question might ask “A bat has a very different wing structure from a bird. What is the wing structure of a bat like relative to that of a bird?” (Answer: “A bird’s wing has fairly rigid bone structure that is efficient at providing lift, whereas a bat has a much more flexible wing structure that allows for greater maneuverability.”), while an inferential question can be “The U.S. Military is looking at bat wings for inspiration in developing a new type of aircraft. How would this new type of aircraft differ from traditional aircrafts like fighter jets?” (Answer: “Traditional aircrafts are
modeled after bird wings, which are rigid and good for providing lift. Bat wings are more flexible, and thus an aircraft modeled on bat wings would have greater maneuverability.”). This is remarkable because transferring knowledge to a different knowledge domain is consider to be a far transfer along a single contextual dimension, according to the taxonomy proposed by Barnett and Ceci (2002). As such, these findings suggest that the learning underlying the testing effect occurs at a deeper conceptual level not just at a shallow surface level.

This conclusion is further bolstered by studies that have explored whether the testing effect extends to new inferential questions about previously untested but related materials. Previous studies demonstrated that testing facilitates mnemonic retention of non-tested information from the study materials (Foos & Fisher, 1988) and facilitates the transfer of this information (McKenzie, 1972). Recently, the issue has been re-examined (e.g., Chan, 2009, 2010; Chan et al., 2006). For example, in Chan et al. (2006), students studied prose materials, and then were given an initial retrieval test or an opportunity to re-study statements extracted from the study materials. These statements corresponded to the items in the retrieval test. After a day delay, the students took another test that included tested/re-studied materials and non-test/non-restudied materials. For example, an item in the initial test might be “Where do toucans sleep at night?” (Answer: “Tree holes”), the subsequent test might ask “What other bird species is the toucan related to?” (Answer: “Woodpeckers”). Although the non-tested/non-restudied materials were not presented in either way during the test/re-study phase, they were presented in the study phase and were related to the tested/re-studied materials. The results showed that testing can indeed benefit the retention of non-tested but related materials, and this benefit does not occur if students were merely given the chance to re-study. Chan et al. (2006) referred to this phenomenon as retrieval-induced facilitation.
Chan et al. (2006) argued that the facilitation stemmed from the fact that students often consciously think about related or relevant information when they try to search for an answer to a question. The conscious retrieval of related information might promote the retention of this information allowing students to perform better at a final retention test that tapped into this related information. Supporting their theory on the mechanism underlying the phenomenon, Chan et al. (2006) demonstrated that the benefit disappeared if students were instructed to employ a narrow retrieval strategy by which they consciously avoid thinking about related information when searching for the answers.

When attempting to think about and understand the transfer of the testing effect better, it is useful to more deeply consider how transfer itself might occur. Barnett and Ceci (2002) argued that the memory demands of a transfer task can be broken down into three components: recognition, recall, and execution. Students have to recognize when it is appropriate to use previously learned knowledge to a new context. They also have to be able to spontaneously recall the knowledge that was learned earlier. Lastly, they have to be able to apply that knowledge to successfully execute the transfer task.

Given the retention benefits of testing, taking a test probably improves and strengthens students’ performance on the recall component allowing a better to chance to complete a transfer task. Furthermore, Butler (2010) suggests that testing might also improve students’ performance on the execution component of the transfer process by enhancing students’ ability to apply the knowledge. Testing requires students to produce a response to a question from memory. This process provides an opportunity for students to practice using and applying the information they have previously learned in a different context (i.e., studying versus testing). Additionally, the testing process might also foster a better and deeper understanding of the information (McDaniel,
Howard, & Einstein, 2009). All these processes associated with testing together provide a stronger basis for students to apply and transfer their knowledge to different contexts and domains.

5.2 Negative effects of testing

Although testing provides considerable mnemonic benefits, testing can also lead to negative consequences. This is particularly problematic for tests that contain misinformation, such as multiple-choice tests. Given the typical structure of a multiple-choice item, students are not only shown the correct answers, they also exposed to multiple incorrect alternatives (i.e., misinformation) at the same time. For example, in a typical four-alternative multiple-choice item, students would be exposed to one correct answer along with three incorrect alternatives.

Remmers and Remmers (1926) coined the term “negative suggestion effect” to refer to the possibility that students may learn incorrect information as a result of exposure to false test items or alternatives.

Numerous studies have demonstrated that exposure to incorrect information can have a significant impact on latter memory. Brown (1988) and Jacoby and Hollingshead (1990) showed that exposure to misspelled words increased the likelihood of misspelling in a subsequent oral spelling test. In another study, students were asked to judge the fame of a series of famous and non-famous names. Students were found to be more likely to call non-famous names “famous” if they were presented the names a day earlier (Jacoby, Kelley, Brown, & Jasechko, 1989). Brown and Marsh (2008) also demonstrated that exposure to campus scenes increased the likelihood that students would believe that they had previously visited locations that they had never actually been to, after a delay of one or three weeks.
Some debate exists with respect to whether the negative suggestion effect is a relevant concern in the context of testing. Researchers have argued that students have knowledge that they will encounter false information on tests, and they believe that exposure to this misinformation will not have any lasting effect on memory (Ebel, 1972; Ross, 1947; Ruch & Stoddard, 1927). Yet, a number of studies have provided data that show otherwise. Hasher, Goldstein, and Toppino (1977) asked students to rate the perceived validity of plausible statements on various areas of general knowledge in an initial session. Students then came back for two sessions, after two and four week intervals, to complete the validity rating again. The results showed that there was a significant increase in perceived validity ratings for the statements that were presented in the previous sessions, and the increments were found for both objectively true and objectively false statements.

Toppino and Brochin (1989) used a similar methodology to examine the negative suggestion effect in true-false tests. They asked students to study passages about U.S. presidents and then do a true-false test based on the materials from the passages. The students were invited to come back a week later to rate the perceived validity of a series of statements about the U.S. presidents, some of which were taken directly from the items in the true-false test. The results showed that prior exposure to an item on a true-false test increased the students’ validity ratings regardless if the statements were objectively true or false.

In a follow-up study, Toppino and Luipersbeck (1993) extended the findings to show that the effects remained robust even if the statements were paraphrased and were not verbatim repetition from the true-false test. Furthermore, they also demonstrated the same effect with a two-alternative multiple-choice test suggesting that negative suggestion effect may apply to any test that contains misinformation. However, it should be noted the perceived validity ratings of the
objectively false statements never reached the level of the objectively true statements even after testing.

More recent studies have confirmed that taking multiple-choice tests indeed causes negative consequences on future memory performance. In a study conducted by Roediger and Marsh (2005), students were given a multiple-choice test after they studied prose materials. Then, after a five-minute delay, the students were given a cued-recall task to assess their retention of the materials. The cued-recall task included items that were tested in the multiple-choice test and items that were not tested. The results of this experiment showed that prior multiple-choice testing prompted students to produce lures in the cued-recall task (i.e., intrusion errors), even for the students who were warned against guessing. The findings have been replicated in a number of studies (e.g., Butler et al., 2006; Marsh et al., 2009) and with a longer delay (Butler & Roediger, 2008; Fazio, Agarwal, Marsh, & Roediger, 2010). These results suggest that test-takers can indeed acquire false knowledge from a multiple-choice test.

Roediger and Marsh (2005) also found that intrusion in the cued-recall retention task varied as they manipulated the difficult of the initial multiple-choice test by varying the number of lures. Particularly, they found that the intrusion in the cued-recall retention task increased with the number of lures in the initial multiple-choice test. This effect was largely a result of students reproducing the incorrect lures they had chosen in the initial multiple-choice test (i.e., persistence of errors). The effect of the number of lures in the multiple-choice on intrusion in the cued-recall task remained significant even though the analysis was restricted to the selected lures; in contrast, the unselected lures were rarely produced in the cued-recall task, and the rate of production of unselected lures was not related to the number of lures. The same pattern of results
held even if the final retention task occurred a week later (Butler & Roediger, 2008) or if the final retention task was a free-recall task (Experiment 2 & 3: Butler et al., 2006).

Marsh et al. (2009) provided another piece of evidence supporting the idea that intrusion errors are primarily a result of students re-producing the previously selected lures. In their study, students were assigned into one of the two instruction conditions, a free-response condition or a forced-response condition. In the free-response condition, the students were told that responding to an item incorrectly in the initial test would be penalized while skipping an item would not affect the score, and they were instructed to answer strategically to earn the best score possible. In the forced-response condition, the students were told that they should respond to every item even if they had to guess. The results of the study showed that the students committed less intrusion errors in the free-response condition suggesting that simply selecting a lure might increase the chance of intrusion.

Of course, if students are learning while they are taking tests, it makes perfect sense that they could learn incorrect associations as well as correct ones. Thus, the intrusion errors, or more specifically, the persistence of errors, can be viewed as the dark side of the testing benefits typically highlighted in discussions of the testing effect. Students are forming connections between questions and answers, and they do so regardless of whether the answers are correct or not.

Though all these findings point to a relation between committing an error on a multiple-choice and intrusion in the retention task, this relation is not completely due to a bias for maintaining the same response. Using a different paradigm, Schooler, Foster, and Loftus (1988) found that endorsing a lure reduced the probability ratings of the correct answer even when the endorsed lure was not included in the final test.
Overall then, the persistence of errors suggests that students do indeed learn while being tested, yet they learn incorrect associations at the same time as they learn correct ones. That is, when they chose or produce incorrect responses, those responses are subsequently remembered in association with the initial questions. Given the widespread use of multiple-choice testing in modern educational contexts this finding is of high importance. Imagine a final exam in which the multiple-choice component is answered with 70% accuracy. In this case, the last thing students in this course might do is to learn incorrect associations for 30% of the materials they are tested on! Any factors that might mediate this negative effect would thus be very important, one of which is feedback.

5.3 Feedback

The premise that providing feedback after an initial retrieval test would promote performance in subsequent tests seems obvious; however, researchers in motor learning contexts have shown that providing feedback may improve performance during training sessions, yet it often does so at the expense of longer term retention (e.g., Rosenbaum, Carlson, & Gilmore, 2001; Schmidt & Bjork, 1992). The implication is that providing feedback prevents individuals from engaging in a deeper processing during learning, and therefore later retention is compromised.

However, recent studies have demonstrated that the motor learning results do not generalize to learning of higher cognitive skills. For example, Pashler et al. (2005) had students study Luganda-English word pairs and take two consecutive cued-recall tests during the initial session, then the students were invited back to take another cued-recall test a week later. These researchers found that providing answer feedback, in which the feedback messages were the correct answers, further promoted retention in subsequent tests relative to merely taking a test without feedback. The feedback effect was found in the second test in the initial session and in
the subsequent test after a week delay. Similar results have been replicated in various studies using different materials and different types of retrieval tests (e.g., Butler et al., 2007, 2008; Butler & Roediger, 2008).

Although studies have found that the mnemonic benefits of testing are often greater than the negative consequences of acquiring misinformation (e.g., Butler et al., 2006; Butler & Roediger, 2008; Fazio, Agarwal, et al., 2010; Marsh et al., 2009; Roediger & Marsh, 2005), feedback can further enhance the net benefits of testing by maximizing the positive benefit while minimizing the negative effect simultaneously. Butler and Roediger (2008) conducted an experiment in which students studied prose materials and were then given a multiple-choice retrieval test. The students were given feedback for some of the items. A week later, the students showed superior performance on recall tasks for the items presented with feedback in comparison to items presented without feedback. Furthermore, the negative effect of intrusion was dramatically reduced when feedback was provided. The results of this study suggest that providing feedback can dramatically increase the efficiency of using a test as a learning tool. It is important to provide feedback after any type of test, but it is particularly critical for recognition tests, such as multiple-choice tests and true-or-false tests, in which students are exposed to false information during testing.

Feedback can potentially improve the effect of testing in two ways: by correcting errors and by maintaining correct responses. There is solid evidence that feedback can reverse errors (Butler, Fazio, & Marsh, 2011; Butler et al., 2007, 2008; Butler & Roediger, 2008; Butterfield & Metcalfe, 2001, 2006; Fazio, Huelser, et al., 2010; Guthrie, 1971; Pashler et al., 2005). In contrast, the evidence supporting the notion that feedback can help maintain correct responses is less clear. Some studies have demonstrated that when a correct response is produced in the initial
retrieval test, feedback does not enhance subsequent retention (Guthrie, 1971; Pashler et al., 2005). For example, Guthrie (1971) had students recall missing words in sentences they had studied previously. The students were given feedback for some of the items in the initial test. Guthrie (1971) found that when the students made an error in the initial test, providing feedback strongly facilitated subsequent performance; however, when the students did not make an error in the initial items, feedback was not helpful. Similarly, in Pashler et al. (2005), when the final cued-recall performance was examined as a function of whether the responses in the initial test were correct or incorrect, the results suggested that feedback facilitated retention performance of the items with incorrect responses and missing responses, but made little difference for the retention of the items with correct responses one week later.

Butler et al. (2008) pointed out that these studies might not have allowed proper assessment of the benefits of providing feedback after correct responses because, in all cases, students were allowed to volunteer or withhold responses in the initial retrieval test. Butler et al. explained that there is a strong positive relationship between the confidence of a response and willingness to volunteer that response, such that students have a tendency to volunteer high-confidence responses and withhold low-confidence responses even when they are objectively correct. With the way the experiments were designed and set up in Guthrie (1971) and Pashler et al. (2005), very few, if any, low-confidence correct responses in the initial retrieval test might have been recruited. These authors argued that those are the items that are most likely to benefit from feedback.

Butler et al. (2008) were able to demonstrate positive effects of feedback on accurate responses using a forced-response multiple-choice test as the initial retrieval test. They found that providing feedback substantially improved the benefit of prior testing on the subsequent cued-
recall test, and the benefits applied to both initially correct and incorrect responses. They found that students not only corrected a lot of the initially incorrect responses if they were provided with feedback than if they were not, they also maintained a greater proportion of the initially correct responses if feedback was provided.

The results of studies discussed so far were primarily based on answer feedback; providing the student with the correct answer after a response is made. Yet, there are numerous types of feedback which can range from a full representation of the original study materials that allow students to determine the accuracy of their responses (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008) to verification feedbacks that simply indicate whether the responses are correct or incorrect (Marsh et al., 2012). Feedback may also include other information such as representation of the questions and students’ responses, both of which provide contextual information to assist the processing of the feedback, which might be particularly important when the feedback is given after a delay.

Studies have generally found that when correct answers are included in the feedback, which allows students to both evaluate the accuracy of their responses and encode the correct responses if necessary, the future retention of the materials are often enhanced (Butler et al., 2007, 2008; Butler & Roediger, 2008; Pashler et al., 2005). This holds true even if students have to look up the correct responses from the study materials (Agarwal et al., 2008) or discover the correct responses within the items such as in answer-until-correct multiple-choice items (Butler et al., 2007).

However, if feedback only indicates whether a response is correct or incorrect, the retention benefit depends on multiple factors: the format of the retrieval test, and the accuracy and confidence of the responses on the retrieval test. In Pashler et al. (2005), students studied word
pairs in a different language and were tested initially with a cued-recall task, and then were tested again a week later. In this study, the students were allowed to withhold a response in the retrieval at their own discretion. The researchers found that providing verification feedback did not provide any retention benefit in the subsequent cued-recall task compared to providing no feedback. Fazio, Huelser, et al. (2010) replicated the findings with prose materials; however, they also found that when students were forced to respond in the retrieval test, verification feedback led to a greater persistence of low-confidence correct responses from the retrieval test to the final retention task than if the students were not given feedback. The effect was as robust as in the answer feedback condition (Fazio, Huelser, et al., 2010).

In a recent study, Marsh et al. (2012) argued that if a response on a recall test is labeled as incorrect, there remains an infinite number of possible alternatives, whereas knowing a multiple-choice selection is incorrect allows students to limit the possible correct alternatives. They thus argued that verification feedback provides students with more information if the retrieval test is in a multiple-choice format than if it is in a recall format. Marsh et al. (2012) were able to demonstrate that providing verification feedback could indeed lead to additional retention benefits over providing no feedback using a multiple-choice retrieval test. They showed that providing verification feedback helped to maintain correct responses as effective as answer feedback. Providing verification feedback also helped to correct errors, although they are not nearly as efficient as answer feedback. The ability of verification feedback in correcting errors was found to be inversely related to the number of alternatives in the multiple-choice retrieval test (Marsh et al., 2012).
6 Summary and Roadmap

The goal of this introduction was to present the evidence in favor of the testing effect and then to highlight some of the factors related to it, factors that will be relevant to a variety of manipulations performed in my thesis experiments. The chapters that follow represent my attempt to replicate and extend these findings as follows.

In Chapter 2, I investigated the extent to which the effect of testing extends to the non-retrieved information such as the cues or information embedded in the cues. The chapter includes a pure laboratory-based study (Experiment 1) as well as two hybrid studies combining actual classroom experiences with controlled assessments in the laboratory (Experiment 2A & 2B). The two hybrid studies also touched upon various issues related to the temporal dimension of the testing effect.

The studies in Chapter 2 revealed that the retention effect from testing was equally robust for the targets and the cues when students responded correctly in the initial test; however, when students responded incorrectly in the initial test, they retained the cues better than the targets when prompted with the counterparts in the delayed test. In Chapter 3 and 4, a series of studies were conducted to examine the difference in retention performance of the cues and the targets for the initial incorrect items. Particularly, the experiments focused on investigating various factors that might contribute to or eliminate the difference in retention performance, including prior study (Experiment 3), verification feedback (Experiment 4A & 4B) and answer feedback (Experiment 5).
Chapter 2
Testing Effect and Cue-Target Symmetry

In most of the studies on the testing effect, students study some materials, then they take an initial test(s), and after a delay, they take a final test in which they were given the same questions (i.e., cues) from the initial test and were required to retrieve the same answers (i.e., targets) (e.g., Carrier & Pashler, 1992; Roediger & Marsh, 2005). Studies utilizing this methodology have repeatedly demonstrated that testing can promote the retention of the targets over long intervals suggesting the memory of the retrieved information is strengthened. It is of interest whether, and if yes, to what extent, this mnemonic benefit extends to the non-retrieved information that is presented in the test, such as the cues or information embedded in the cues.

As highlighted in the introduction to this thesis, recent studies have demonstrated that the retention benefit from testing is not restricted to information that was retrieved in the initial test. Several studies have found that the testing effect extends to non-tested but related information (Chan, 2009, 2010; Chan et al., 2006; Foos & Fisher, 1988). Other studies have found that the information that is strengthened through the process of testing is transferable. Not only can this information be used to answer distinctly different questions (Rohrer et al., 2010), and be transferred and applied flexibly to questions that required different responses (Butler, 2010), but it even transfers to questions that are in a different contextual domain (Butler, 2010).

These findings suggest that the mnemonic facilitation of testing is not strictly specific to the retrieved information being tested under the same condition as in the initial test. If what is being strengthened is the associative link between a cue and a target, then one would expect prior testing to improve the retention of the cue as well if the target was used as a prompt. Indeed,
studies have confirmed that retention of the cues is enhanced after taking a test (Carpenter et al., 2006; McDaniel et al., 2007).

Given the studies described above, it does indeed appear that the association between the cue and target is being strengthened by testing. This then gives rise to a second question. Is the facilitation of target retrieval given the cue as strong as the facilitation of the cue retrieval given the target? That is, are facilitation effects symmetrical as one might expect if the sole cause of facilitation was the strengthening of the association between cue and target? If so, this would suggest that a relatively simple process of strengthened associations is underlying the testing effect. However, it is also possible that the strengthening of links is more complex. If asymmetries in facilitation are observed that will prompt questions about what other factors are relevant, and what those factors then tell us about the processes underlying the testing effect.

1 Experiment 1

A study conducted by Carpenter, Pashler and Vul (2006) provides us with some relevant data on the symmetry of facilitation issue. They conducted two experiments with minor experimental differences. In both experiments, students were first given 40 word pairs to study in the study phase. Then in the second phase, they were re-presented with half of the pairs (re-study trials) and were tested on the other half (testing trials). After 18 to 48 hours of delay, the students were administrated a cued-recall task in which they were either given the same cue words from the initial test and were asked to recall the matching target words or they were given the target words and were asked to recall the matching cue words. The results of both experiments showed a testing effect for both cued-recall conditions suggesting that testing improved the retention for both the cues and the targets. However, the results of the two experiments were not consistent in terms of the magnitude of the testing effect in the different conditions. In one of the experiments,
students showed a better retention performance when they were asked to recall the targets when they were given the cues than vice versa. However, the other experiment did not show the same difference; the effects in both conditions were equivalent; that is, symmetrical.

In addition to the inconsistent findings, there are some methodological concerns with the Carpenter et al. (2006) study. First, the cues and the targets were explicitly defined in the study phase. The cue word was always positioned on the left hand side inside a box labeled as the “cue”, while the target word was always presented on the right hand side inside another box labeled as the “target”. With the cue and the target defined explicitly during the study phase, it is possible that the students encoded them differently compared to a situation in which the relationships were not highlighted and explicitly defined.

The second concern with this study was with respect to the format of the initial test they employed. In the testing trials, a cue appeared on the screen first by itself, for four seconds, then the matching target appeared and remained on the screen with the cue for another two seconds. The students were instructed to do a self-test and try to retrieve the target before it appeared on the screen. No data was available for the self-test, which made it impossible to know whether the students actually attempted to retrieve the target for all the trials. It also prohibited any form of conditional analysis with respect to the performance accuracy in the initial test. This is problematic because the retention performance has been found to vary greatly dependent on the accuracy in the initial test (e.g., Butler et al., 2007, 2008; Butler & Roediger, 2008).

Lastly, because the target was re-presented subsequently, presumably to provide feedback, it is not certain how much of an impact this feedback had on the performance in the final recall task. Past studies have demonstrated that feedback can greatly improve the efficiency of a retrieval test (Butler et al., 2007, 2008; Butler & Roediger, 2008). Therefore, Carpenter et al. (2006)
might not have provided an accurate assessment of the relative strength of the tested-induced retention of the cues and the retention of the targets.

Experiment 1 of this thesis was designed to address the methodological issues with Carpenter et al. (2006) and thereby provide a strong re-examination of the symmetricity of the testing benefits. It also included an additional control condition in which the pairs were neither tested nor re-presented to provide an assessment of the baseline performance.

1.1 Method

1.1.1 Participants

The participants were 50 university students who were taking an introductory psychology course at University of Toronto Scarborough. They participated in the experiment in exchange for course credit. Data from an additional participant was excluded because of an experimental error.

1.1.2 Materials

The materials for the experiment consisted of 60 word pairs. They were selected from the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998). All the pairs were weakly associated in the forward and the backward directions; both the forward cue-to-target associative strengths and the backward target-to-cue associative strengths were below 0.1, for example, table and cloth. An additional 10 word pairs were selected to be used as fillers.

For each individual word among the experimental pairs, three additional associated pairing words were chosen from the University of South Florida Free Association Norms to be used as lures in a multiple-choice test. The forward associative strengths between the experimental word and the lures were below 0.1.
1.1.3 Procedure

The experiment consisted of three phases: the study phase, the multiple-choice test/re-study phase, and the recall phase. In the study phase, the participants were given an opportunity to study the word pairs. The word pairs were presented at the center of a computer screen one pair at a time for three seconds each. The experimental pairs were presented in a random order with five filler trials presented at the beginning of the sequence, before the experimental pairs and the five of the fillers presented at the end, after all the experimental pairs. The fillers were implemented in attempt to reduce the impact of the primacy and recency effects (Rundus, 1971). Which word in a pair was presented on the left or on the right was randomly assigned for each trial.

In the multiple-choice test/re-study phase, 30 of the experimental pairs were randomly selected to be included in the test (i.e., the test pairs). For each trial, one of the words in a word pair was randomly selected and presented as a cue, and the participants were instructed to select the matching target, from four alternatives, all of which were weakly associated with the cue word. Additionally, 15 experimental word pairs were re-presented, for three seconds each (re-study pairs). As was the case in the study phase, the left-right order of the words was randomized. The multiple-choice and the re-study items were randomly intermixed across experimental trials. The remaining 15 experimental word pairs were the control pairs; they were neither tested nor re-presented in this phase. Upon finishing, the participants played Tetris for 15 minutes before taking the delayed test.

Finally, in the last part of the experiment, the recall phase, the participants were given a cued-recall task. In each trial, one of the words from an experimental pair was presented, and the participants were instructed to recall and type out the matching word. All 60 experimental word
pairs were tested in this phase. For the 30 word pairs used in the multiple-choice test, half of them were tested in a forward manner, in which the participants were presented with a cue from the multiple-choice test and were instructed to recall the associated target word, while the other half were tested in a backward manner, in which the participants were presented with a target from the multiple-choice test and were instructed to recall the associated cue word. 1

1.2 Results

The accuracy on the multiple-choice test was 0.741 ($s = 0.137$). Overall, there was no significant difference in multiple-choice accuracy between the pairs that were subsequently tested in the forward manner ($\bar{x}_{MC\_test\_forward} = 0.735, s = 0.155$) and the pairs that were subsequently tested in the backward manner ($\bar{x}_{MC\_test\_backward} = 0.748, s = 0.153$), $t(49) = -0.674, d = -0.095, p = 0.504$.

The cued-recall performance across the four experimental conditions (control, re-study, test-forward, and test-backward) is presented in Figure 2. Analysis of this figure suggests, in general, that re-studying seemed to provide a benefit relative to not re-studying, but that the benefit of testing was greater still, though seemingly equivalent across the forward and backward directions.

---

1 Participants were prompted to rate how confident they were with their response on a 6-point Likert scale after they made a response in the initial test and in the final retention task for all the experiments. The confidence rating results were very similar to those of the performance data, therefore they were not reported.
Figure 2. Experiment 1: Mean proportion recalled for the four experimental conditions in the cued-recall task.

These data were analyzed with a one-way repeated-measure analysis of variance (ANOVA). The effect of Condition was significant, $F(3, 147) = 45.572, MSE = 0.012, \eta_p^2 = 0.482, p < 0.001$. Taking the initial multiple-choice test bolstered the performance in the subsequent recall task ($\bar{x}_\text{test-forward + test-backward} = 0.606, s = 0.165$ vs. $\bar{x}_\text{control} = 0.383, s = 0.202$), $t(49) = 11.538, d = 1.536, p$
< 0.001, and the retention enhancement was greater than being re-exposed to the materials for three extra seconds ($\bar{x}_{\text{re-study}} = 0.500$, $s = 0.176$), $t(49) = 5.476$, $d = 0.795$, $p < 0.001$. The results also revealed no significant difference between the cued-recall performance in the test-forward condition ($\bar{x}_{\text{test-forward}} = 0.597$, $s = 0.188$) and the test-backward condition ($\bar{x}_{\text{test-backward}} = 0.615$, $s = 0.173$), $t(49) = -0.776$, $d = -0.117$, $p = 0.442$.

To investigate the effect of response accuracy of the multiple-choice test on memory facilitation and on the symmetry of the testing effect, additional analyses separated the cued-recall task performance based on the correctness of the participants’ responses in the multiple-choice test. This $2 \times 2$ repeated-measure ANOVA was conducted with both Accuracy (correct and incorrect) and Direction (test-forward and test-backward) as within-subject factors and only included the participants who had data in all the cells ($n = 45$). The summary data is presented in Figure 3. The figure suggests the benefit of testing was greater for the correct trials than the incorrect trials in general, and the benefit seemed to be equivalent across the forward and backward directions for the correct trials while for the incorrect trials, it was stronger in the backward than the forward direction.
Figure 3. Experiment 1: Mean proportion recalled for the two cueing directions for the initially correct and incorrect items in the cued-recall task.

In general, the participants recalled more matching word pairs if they responded correctly in the initial multiple-choice test ($\bar{x}_{\text{correct}} = 0.742, s = 0.143$) than if they responded incorrectly ($\bar{x}_{\text{incorrect}} = 0.155, s = 0.137$), $F(1, 44) = 519.753$, $MSE = 0.030$, $\eta_p^2 = 0.922$, $p < 0.001$. 
The results also showed a significant effect of Direction, $F(1, 44) = 5.145$, $MSE = 0.023$, $\eta^2_p = 0.105$, $p = 0.028$, and an Accuracy × Direction interaction, $F(1, 44) = 8.994$, $MSE = 0.034$, $\eta^2_p = 0.170$, $p = 0.004$. For the correct trials, the performance in the forward ($\bar{x}_{test-forward/correct} = 0.758$, $s = 0.165$) and the backward ($\bar{x}_{test-backward/correct} = 0.726$, $s = 0.163$) directions were comparable, $t(44) = 1.312$, $d = 0.196$, $p = 0.196$. However, for the incorrect trials, the recall performance in the backward direction ($\bar{x}_{test-backward/incorrect} = 0.222$, $s = 0.249$) was superior to the forward direction ($\bar{x}_{test-forward/incorrect} = 0.088$, $s = 0.145$), $t(44) = 3.009$, $d = 0.449$, $p = 0.004$.

The performance in both the forward and the backward directions for the correct trials were significantly higher than those for the control and restudy conditions ($p < 0.001$ for all comparisons). In contrast, the performance in the two cueing directions for the incorrect trials were significantly lower than those for the control and re-study conditions ($p < 0.001$ for all comparisons). Prior testing seems to have two distinct effects; it provides positive benefits to the items that were answered correctly in the initial test while it has negative impacts on the items that were answered incorrectly.

Lastly, the negative effects of testing were assessed in two ways: the frequency of the participants recalling 1) the incorrect options they selected in the multiple-choice test and 2) the other unselected lures. The participants were more likely to recall the incorrect options they selected in the multiple-choice test when they were prompted in the forward direction with the cues ($\bar{x}_{incorrect response test-forward} = 0.442$, $s = 0.308$) than when they were prompted in the backward direction with the targets ($\bar{x}_{incorrect response test-backward} = 0.083$, $s = 0.157$), $t(44) = 7.382$, $d = 1.100$, $p < 0.001$. Thus, the high level of persistence of errors observed in the forward direction was all but eliminated when the target was used to prompt retrieval of the cue.
The participants were also more likely to recall unselected lures if they were prompted in the forward direction ($\bar{x}_{\text{unselected lure test-forward}} = 0.097, s = 0.110$) than in the backward direction ($\bar{x}_{\text{unselected lure test-backward}} = 0.060, s = 0.095$) regardless if they selected the correct options in the initial test, $F(1, 44) = 3.958, MSE = 0.015, \eta_p^2 = 0.083, p = 0.053$. However, the likelihood of recalling unselected lures was not related to the accuracy in the initial multiple-choice test ($\bar{x}_{\text{unselected lure correct}} = 0.057, s = 0.054$ vs. $\bar{x}_{\text{unselected lure incorrect}} = 0.100, s = 0.157$), $F(1, 44) = 2.930, MSE = 0.028, \eta_p^2 = 0.062, p = 0.094$.

1.3 Discussion

Experiment 1 investigated whether and to what extent the retention benefit of testing extends to information that is not retrieved in the initial test yet is present in the test; specifically, whether using the targets as prompts enhanced recollection of the cues. The aim of the experiment was to compare the strength of the test-induced retention benefit of the cues relative to that of the targets.

First, a robust testing effect was found. The students showed better retention for items that were tested, regardless of the cueing directions, relative to items that were neither tested nor re-studied. Furthermore, the retention benefit of testing surpassed the retention enhancement that resulted from the extra exposure to the materials (i.e., the re-study condition). These results further support the claim that the testing effect is not merely an artifact of additional exposure to the to-be-tested materials (e.g., Roediger & Karpicke, 2006b; Slamecka & Katsaiti, 1988; Thompson et al., 1978; Wheeler et al., 2003).

The experiment did not find the superior short-term retention benefit of re-studying over testing which was demonstrated in some previous studies (Roediger & Karpicke, 2006b; Wheeler et al., 2003). However, this experiment was not intended or designed to assess the crossover interaction
on the retention performance between testing and re-studying as the retention was only assessed once, after a 15-minute delay. For whatever reason, this delay was sufficient to allow the testing effect to emerge above the benefit of re-studying.

Second, taking a test strengthened the retention of both the targets and the cues, when the students were cued forward with the cues and backward with the targets respectively. The recall performance for both the forward and backward directions was superior to that for items in the control and the re-study conditions. This suggests that the retention benefit of testing is not limited to the targets, and the benefit extends to the cues as well. This result is consistent with Carpenter et al. (2006) who also found that the retention of the cues is enhanced by prior testing when assessed with a cued-recall or a free-recall task. The result is also consistent with other studies which have demonstrated that the retention benefit of testing is not limited to the information that was retrieved in the initial test (Chan, 2009, 2010; Chan et al., 2006; Foos & Fisher, 1988; McDaniel et al., 2007).

Third, the testing effect was found to be generally symmetrical across a cue and its associated target. The experiment found that the cued-recall task performance was similar for the forward and the backward conditions, at least when one looks at the overall performance or at the performance for the items that were correctly responded to in the initial test. This result suggests that taking an initial test enhanced the retention of the cues to a similar extent as the targets, supporting the notion that it is the link between the cue and target that is strengthened by the testing effect.

Fourth, the results observed in the overall accuracy data and in the data related to items responded to correctly in the initial test does not apply to items that were incorrectly responded to during the initial test. The retention of the initially incorrect items was substantially worse
than the initially correct items, consistent with the findings of some previous studies (Butler et al., 2007, 2008; Butler & Roediger, 2008), and it was also found to be significant below the base rate recall performance suggesting a negative testing effect for items that were incorrectly responded to. More importantly, unlike the findings for initially correct items, the negative testing effect was asymmetric across cueing directions. Specifically, the retention performance was dramatically better in the backward condition than the forward condition.

Both the testing benefit and the negative testing effect might indeed be reflections of the same basic process: the strengthening of the link between a cue and whatever response the students choose as their answer. When the answer is correct, the link between the cue and the target is strengthened, and retention in both the forward and backward directions are facilitated. However, when the answer is incorrect, the students effectively associate the cue with one of the incorrect lures. Thus, when they are then presented with the cue during recall, the newly formed association influences them to retrieve that incorrect lure. However, when the students are prompted in the backward direction with the target from the multiple-choice test, they can perhaps figure out that their response in the initial test was incorrect. That is, although the association between cue and target is not strengthened during the test, neither is the association between the target and any other lures. The students could possibly retrieve or recover the cue word, relying on the memory from the study phase or the memory of information provided in the initial multiple-choice test. Such a difference in item structure between the two cueing directions can explain the observed asymmetry.

Consistent with this notion and with Roediger and Marsh (2005), the negative testing effect was mostly a result of the students re-producing the incorrect lures they had chosen in the multiple-choice test (i.e. persistence of errors) with intrusion of unselected lures being rather uncommon.
Nevertheless, the students were still more likely to recall an unselected lure in the forward direction than in the backward direction, regardless of the response accuracy in the initial multiple-choice test. This is likely because the lures, on average, have higher associative strengths with the cue word than the target word. Therefore, when the students are prompted in the forward direction with the cues, they had a greater likelihood of mistakenly recalling one of the lures than when they were prompted backward with the targets.

2 Experiment 2A

Experiment 1 demonstrated that the retention benefit of testing applies not only to the information that was retrieved in the initial multiple-choice test but also the information in the cues. Experiment 1 also demonstrated that the benefit of testing was similar between the retention of the cues and the retention of the targets when the students responded correctly in the initial test; however, for the initially incorrect items, the students showed superior retention of the cues than of the targets. Experiment 1 was conducted in the laboratory using highly controlled word pairs and using a relatively short retention interval between the initial multiple-choice test and the final cued-recall task.

Although laboratory-based studies often allow a high degree of control over stimuli and testing conditions, there always exists the worry that, in the perception of students, a “memory test” for word pairs presented in a laboratory is quite different from a real test performed in a classroom. Given the stakes, the richness of the materials, and the distributed nature of real course-based tests, it is entirely possible that levels of motivation and the specific cognitive processes employed during study and test are quite different from those associated with laboratory testing. Thus, Experiment 2A was conducted to explore the generalizability of the Experiment 1 findings with respect to an in vivo testing context.
Results from the literature suggest that the retention benefit of testing is highly robust with respect to study materials and testing contexts. Studies have found the testing effect using a wide array of materials (e.g., Carpenter & DeLosh, 2005; Fazio, Huelser, et al., 2010; Rohrer et al., 2010; Wheeler et al., 2003) including educationally relevant materials (e.g., Gates, 1917; Marsh et al., 2009; Roediger & Karpicke, 2006b; Spitzer, 1939). Furthermore, not only has the testing effect been demonstrated in laboratory settings (e.g., Carrier & Pashler, 1992; McDaniel & Masson, 1985; Roediger & Marsh, 2005; Wheeler & Roediger, 1992), it has been repeatedly demonstrated in classroom settings as well (e.g., Bangert-drown, Kulik, & Kulik, 1991; Foos & Fisher, 1988; Glover, 1989; Leeming, 2002; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Roediger, Agarwal, McDaniel, & McDermott, 2011). It is of interest then whether the findings from Experiment 1 would also replicate in a realistic educational setting using course relevant materials.

In addition, the retention interval between the initial multiple-choice test and the subsequent cued-recall task in Experiment 1 was very brief (i.e. 15 minutes). It is possible that the results reported could be short-lived and might not endure a longer retention delay. Despite the fact that the testing effect has been generally found to be long-lasting and to persist over a long retention delay (e.g., Butler & Roediger, 2007; Duchastel & Nungester, 1981; McDaniel et al., 2011; Roediger et al., 2011; Spitzer, 1939), Experiment 2A was conducted to insure this generality also applied to the more specific findings of Experiment 1.

Thus, Experiment 2A was an adaptation of Experiment 1 designed to examine the retention benefit of testing in both forward and backward cueing directions in an actual university course. The course midterm was utilized as the initial retrieval test. In order to assess how the midterm test influenced students’ retention, students were invited to the laboratory after various delays to
complete a retention task. The retention task included three types of items. These items were verbatim items from the midterm test, modified items that were based on items on the midterm test, and items that were not included in the midterm test. They corresponded to the forward, backward, and control items in Experiment 1 respectively.

2.1 Method

2.1.1 Participants

The participants were 183 university students who were taking an introductory psychology course at University of Toronto Scarborough. They were recruited in the second half of the semester after the midterm test had been administrated. They participated in the experiment in exchange for course credit. Data from three additional participants were excluded because they took the make-up midterm test rather than the regular midterm test. Also, another participant completed the experiment twice because of a scheduling error; the data from her second session was dropped.

2.1.2 Materials

The experimental materials were 24 multiple-choice items covering the materials from the first half of the course. One-third of the items were verbatim items, which means they were the exactly same items that were used in the midterm test. One-third of the items were modified items. Specifically, they were re-worded in such a way that the original correct answers were integrated into the questions while part of the original questions became the correct answers. An example of a pre-modified item might be “Which type of trait is continuous variation characteristic of?” (Answer: “Nonmendelian trait”), an example of a modified item might be “Which of the following is characteristic of a nonmendelian trait?” (Answer: “Continuous
variation”). Appropriate lures were created for the modified items. Lastly, one-third of the items were new items which they were not used in the midterm test.

The three sets of items along with the pre-modified items were generated by the teaching assistant for the course. They had all been used in midterm tests in the previous semesters. Based on the archival data from those tests ($ns > 200$), the difficulty levels of the items and the discriminating factors of the target options were matched across the four sets of items. Furthermore, the item-to-item difficulty level differences between the pre-modified and the modified items were all less than 7%.

2.1.3 Procedure

The participants were presented with the multiple-choice items on a computer screen one at a time. For each trial, they were instructed to select the alternative that represented the best answer to the question or complete the statement. The order of the items and the order of the alternatives for each item were randomized across participants. The participants were given 30 minutes to finish the task. A subset of the participants ($n = 8$) completed a paper-version of the test instead, which was essentially the same as the computer version.

2.2 Results

The accuracy of the 16 previously tested (i.e., verbatim and pre-modified) items in the midterm was 0.723 ($s = 0.151$). There was no significant difference between the verbatim ($\bar{x}_{\text{Midterm\_verbatim}} = 0.721, s = 0.183$) and the pre-modified items ($\bar{x}_{\text{Midterm\_pre-modified}} = 0.725, s = 0.174$), $t(182) = -0.244, d = -0.018, p = 0.807$. 
The performances on the three conditions (i.e., verbatim, modified, and new) in the experimental task are presented in Figure 4. Analysis of the figure suggests a benefit of testing, and the benefit was greater for the modified items than the verbatim items.

*Figure 4.* Experiment 2A: Mean proportion correct for the three experimental conditions in the experiment task.
The data were analyzed with a repeated-measure ANOVA. The results showed a significant effect of Condition, \( F(2, 364) = 39.121, \text{MSE} = 0.023, \eta^2 = 0.177, p < 0.001 \). Further analyses revealed a robust testing effect; the participants performed better on the items that were tested previously in the midterm (\( \bar{x}_{\text{verbatim + modified}} = 0.686, s = 0.176 \)) than the items that were not tested (\( \bar{x}_{\text{new}} = 0.577, s = 0.202 \)), \( t(182) = 7.906, d = 0.570, p < 0.001 \). Additionally, the participants also showed a superior performance on the modified items than on the verbatim items, \( t(182) = 3.967, d = 0.301, p < 0.001 \); however, this result should be interpreted with caution because retention performance for different cueing conditions was previously found to be highly dependent on the performance in the initial test.

The following analyses focused on the effect of response accuracy for specific items in the midterm test on the performance of the verbatim and modified items in the experimental task. Only the participants who had data in all the cells were included in the analysis (\( n = 150 \)). A 2 \( \times 2 \) repeated-measure ANOVA was conducted with both Accuracy (correct and incorrect) and Condition (verbatim and modified) as within-subject factors.

The summary data was presented in Figure 5. Visual examination of the figure suggests that the benefit of testing is greater for the correct trials than for the incorrect trials in general. The benefit is equivalent for the verbatim and modified conditions for the correct trials, illustrating a symmetrical facilitation. However, with respect to items responded to incorrectly in the midterm, the effect is more robust for the modified (i.e., backward cueing) than for the verbatim (i.e., forward cueing) condition.
Figure 5. Experiment 2A: Mean proportion correct of the verbatim and modified conditions for the initially correct and incorrect items in the experimental task.

The participants performed better on the items in the experimental task if they responded to the items correctly in the midterm ($\bar{x}_{\text{correct}} = 0.777, s = 0.166$) than if they responded incorrectly ($\bar{x}_{\text{incorrect}} = 0.386, s = 0.258$), $F(1, 149) = 296.982, MSE = 0.077, \eta^2_p = 0.666, p < 0.001$. The analysis also showed a significant effect of Condition, $F(1, 149) = 52.583, MSE = 0.062, \eta^2_p =$
0.261, \( p < 0.001 \), and Accuracy × Condition interaction, \( F(1, 149) = 57.823, \) \( MSE =0.070, \) \( \eta_p^2 = 0.280, \) \( p < 0.001 \). For the initially correct trials, the performance on the verbatim (\( \bar{x}_{\text{verbatim/correct}} = 0.785, s = 0.222 \)) and the modified (\( \bar{x}_{\text{modified/correct}} = 0.769, s = 0.189 \)) items were similar, \( t(149) = 0.540, d = 0.067, p = 0.590 \), and they were both significantly higher than the performance on the new condition (verbatim/correct vs. new: \( t(149) = 11.905, d = 0.972, p < 0.001 \); modified/correct vs. new: \( t(149) = 11.553, d = 0.943, p < 0.001 \)). For the initially incorrect trials, the recall performance on the modified items (\( \bar{x}_{\text{modified/incorrect}} = 0.542, s = 0.368 \)) was significantly better than the verbatim items (\( \bar{x}_{\text{verbatim/incorrect}} = 0.230, s = 0.315 \)), \( t(149) = 10.214, d =-0.690 , p < 0.001 \). The performance for the verbatim items was significant worse than the performance for the new items, \( t(149) = -10.857, d = -0.887, p < 0.001 \), while the performance for the modified items and the new items were similar, \( t(149) = -0.307, d = -0.025, p = 0.759 \). Prior testing has a negative impact on the verbatim items that were incorrectly answered in the initial test.

Next, the negative testing effect was assessed. The participants persisted with the same incorrect responses 0.540 (\( s = 0.377 \)) of the time on verbatim trials suggesting that the links formed during testing had a strong influence on their choices in the experimental task. Consistent with this idea, when the initial incorrect responses were not among the alternatives in the experimental task as in the modified items, the participants showed a much higher level of performance.

Furthermore, the participants responded in the experimental task with unselected lures from the midterm about 0.207 (\( s = 0.218 \)) of the time if they answered the items in the midterm correctly and 0.209 (\( s = 0.293 \)) otherwise. The difference was not significant, \( t(161) = -0.069, d = -0.005, p = 0.945 \) (only included the participants with data in all the cells). Thus, it seems the different levels of accuracy across the verbatim and modified conditions were due entirely to the persistence of errors occurring on verbatim trials.
This finding also explains why the overall performance was better for the modified than for the verbatim items, as highlighted at the beginning of this results section. Specifically, the associations formed when the participants responded incorrectly during the midterm could not carry over to the modified items, because of the multiple-choice nature of the retention task, but they could for the verbatim items, thereby insulating the modified items from a factor that caused a very strong reduction in accuracy. It is not so much that the participants are better able to remember modified items, but rather that they are spared from persisting to the initial erroneous responses that would otherwise have a large effect on memory performance.

Forgetting functions

Previous research has suggested that tested items are forgotten at a slower rate than are non-tested items suggesting that testing produce some sort of insulation from forgetting (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b; Runquist, 1983, 1986b; Slamecka & Katsaiti, 1988; Spitzer, 1939; Thompson et al., 1978; Wheeler et al., 2003). Although not specifically designed to do so, our data allows us to also examine the testing effect as a function of retention interval. That is, although retention delay was not manipulated experimentally, the participants did come to the laboratory after various numbers of days had passed since the midterm. This is not an optimal way to assess delay effects as some third variable related to forgetting might confound the results. For example, it might be possible that certain participants were keen to attend the experiment whereas others signed up later, primarily to insure they got credit. Nonetheless, it was a relatively large dataset and therefore worth examining. To the extent previous results are replicated, that would be the case despite potential confounds. To the extent they were not replicated, more specific ideas for further experimentation might become apparent.
Linear regression analyses were conducted using the number of days that had elapsed between the midterm and when participants arrived at the laboratory (linear and inverse) to predict the performance within each condition. The analyses with the inverse transformation were reported here as they provided the best-fitting models overall. The results showed that this inverse function was correlated with the performance on the verbatim items, $r = 0.222$, $r^2 = 0.049$, $p = 0.002$, the modified items, $r = 0.250$, $r^2 = 0.063$, $p = 0.001$, and the new items, $r = 0.230$, $r^2 = 0.053$, $p = 0.002$. Interestingly though, despite there being a clear testing effect throughout, the slopes were not significantly different across the three models: the verbatim items, $b = 0.681$, $\beta = 0.222$, $p = 0.002$; the modified items, $b = 0.696$, $\beta = 0.250$, $p = 0.001$; and the new items, $b = 0.666$, $\beta = 0.230$, $p = 0.002$. The regression lines for the three conditions were presented in Figure 6. Thus, the suggestion was approximately equal forgetting rates for all conditions.
However, returning to the third variable possibility, the slopes from the linear regressions might have been contaminated if there was a tendency for poorer performing students to sign up for the experiment latter during the semester. That is, a negative correlation was found between the midterm performance (excluding the verbatim and pre-modified items) and the number of days elapsed after the midterm, $r = -0.166$, $r^2 = 0.028$, $p = 0.025$ (the correlation between the midterm...
In order to partial out the effect of participants’ midterm performance, three multiple regressions were conducted using the midterm performance and the inverse function of the day elapsed to predict the performance on the three experimental conditions separately. All three regression models were found significant (verbatim: $R = 0.593$, $R^2 = 0.352$, $p < 0.001$; modified: $R = 0.594$, $R^2 = 0.352$, $p < 0.001$, new: $R = 0.566$, $R^2 = 0.320$, $p < 0.001$), and both factors significantly contributed to the regression models with tolerance = 0.969. Similar to the results of the linear regressions, the slopes for the inverse function of the retention interval were not significantly different across the three models: the verbatim items, $b = 0.378$, $\beta = 0.123$, $p = 0.045$; the modified items, $b = 0.427$, $\beta = 0.153$, $p = 0.013$; and the new items, $b = 0.397$, $\beta = 0.137$, $p = 0.029$. Thus, once again, it appeared that forgetting rates were similar for all three experimental conditions.

2.3 Discussion

Experiment 2A examined the generalizability of the findings of Experiment 1 by assessing how well those findings hold in a more realistic educational context with longer retention delays. Experiment 2A incorporated the materials and the midterm component from the introductory psychology course as the source of the testing effect. The students were recruited after the midterm test had been administrated, and they came in to the laboratory to complete a retention task. This setup allowed an assessment of the impacts of the midterm test on the retention of the course materials. The retention task included three types of items which they were corresponding to the forward (i.e. verbatim), the backward (i.e., modified) and the control (i.e., new) conditions in Experiment 1.
Again, testing was found to improve the retention of the tested materials, and its effect extends over long retention intervals. The students showed better retention performance on the items that were tested in the midterm test than the items that were not included in the midterm. This result is consistent with a number of studies looking at the testing effect using educational relevant materials in laboratories and in classroom settings (Glover, 1989; Leeming, 2002; Roediger & Marsh, 2005). Experiment 2A also showed that the retention benefit of testing is not short-lived. The benefit was found with a retention delay of three days to six weeks. This result provided an extra piece of evidence to support the long-lasting nature of the testing effect (Butler & Roediger, 2007; Duchastel & Nungester, 1981; McDaniel et al., 2011; Roediger et al., 2011; Spitzer, 1939).

Importantly, the students not only performed better with the verbatim items, they also performed better on the modified items. The modified items were modified and re-worded so that the original answers were integrated to the questions, while the part of the original questions became the answers. The retention enhancement of the modified items suggests two important points. First, prior testing promotes the retention of information embedded in the cues. This is consistent with the findings of McDaniel et al. (2007) in which the researchers asked students to retrieve a missing word in a statement in the initial test and found that prior testing improved the retention of other words in the statement that were not retrieved in the initial test. Second, the information retained from prior testing is transferable and can be used to respond to questions that are different from the ones in the initial test. This is consistent with findings by Rohrer et al. (2010) when they showed that the name of regions and cities retained through testing can be utilized to respond to different geographic transfer questions. Overall, the results align with a number of studies that have shown that memory for information related to the retrieved information is strengthened through testing and this information can be applied to novel transfer questions.
(Butler, 2010; Johnson & Mayer, 2009; McKenzie, 1972). These results also provide an ecologically valid replication of Experiment 1 supporting the assertion that the testing effect enhances retention regardless of the directions of cueing, forward or backward.

Experiment 2A also once again demonstrated that the overall testing effect is actually composed of a positive effect for items originally responded to accurately and a negative effect for items originally responded to inaccurately. The retention enhancement is closely related to performance in the initial test. Once again, the retention benefit was found to be symmetric across a cue and its associated target if students answered the initial item correctly. That is, the retention performance of the initially correct items was similar for the verbatim and modified conditions. On the other hand, for the initially incorrect items, the pattern was distinctly different. The retention performance was drastically better for the modified items than for the verbatim items, consistent with the findings of Experiment 1 in which the students performed better in the backward direction than the forward direction for initially incorrect items. This poorer retention for verbatim items again appears due to the persistence of errors biasing students to perseverate on incorrect answers when tested in a forward (i.e., verbatim) direction.

The procedure of this experiment did not allow quantification of the persistence of errors for the modified items. However, for the verbatim items consistent with Experiment 1, the large proportion of the erroneous responses did reflect the persistence of initial errors. Yet, Experiment 2A found a more severe level of intrusion of unselected lures which is likely due to the forced multiple-choice format of the retention task. The intrusion of unselected lures was again not affected by the accuracy in the initial test, consistent with Experiment 1.

In conclusion, Experiment 2A replicated the findings of Experiment 1 suggesting that the interaction of the retention performance between initial accuracy and cueing direction is not an
artifact of either the materials used or the short retention delay. The testing effect was observed; it was symmetrical between the verbatim and modified conditions for the items that were responded to accurately in the initial test, but it was asymmetrical for the items that were responded to inaccurately. The testing effect actually worked against the accurate retrieval of verbatim items that were responded to inaccurately in the initial test. That is, it supported perseveration of incorrect responses. Critically then, this effect is not due to any sort of laboratory artifact related to materials, motivation or the cognitive approach of students in an experiment as it occurs in a virtually identical form in the context of a real high-stakes exam.

Experiment 2A also added another piece of information to the understanding of the mechanism underlying the testing effect. It has been widely believed that testing improves retention by slowing down the rate of forgetting. This belief originated from Spitzer (1939) who found that taking a test completely changed the trajectory of the rate of forgetting. Without testing, retention drops in an inverse function of time, dropping rapidly at first, but then with a systematic slowing of the rate of forgetting, hitting asymptote at around three weeks. Spitzer (1939) also found that the slope of forgetting drastically decreases at the point when students take a test. Based on these previous results, it follows that the retention benefit of testing should increase over time, presumably until the retention level of the non-tested condition reaches the asymptotic level at which the remaining materials seem to have a high resistance against forgetting, and then the benefit of testing would attenuate gradually which a number of studies seem to suggest (Duchastel & Nungester, 1981; Fazio, Agarwal, et al., 2010; McDaniel et al., 2011; Roediger et al., 2011; Spitzer, 1939).

This view has been fortified by findings from studies that examined the crossover interaction between testing and re-studying (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b;
Thompson et al., 1978; Wheeler et al., 2003). These studies showed that re-studying enhances retention to a greater extent than does testing in a short-term. Yet, the level of retention of the re-studied materials drops at a much faster rate than that for the tested materials. Over time, the benefit of re-studying over testing gradually disappears, and eventually, the pattern reverses favoring testing over re-studying. These studies usually used a delay of five minutes to a week.

Experiment 2A, however, found no indication that the students forgot the non-tested items at a faster rate than the tested items. Instead, the retention benefit of testing remained constant over a period of three days to six weeks. However, because of the shape of the inverse function, if one only focused on a short period of time after the initial test, for instance, the first 10 days, the forgetting rate of the tested items indeed appeared to be slower than the forgetting rate of the non-test items. This might explain the inconsistent findings between the current study and the previous research. Nevertheless, Experiment 2A demonstrated that the benefit of testing remained constant over time and did not increase infinitely.

While it is true that Experiment 2A was not directly designed to assess forgetting rates, and that there does remain the potential of some confounding variable that might explain the findings. The most obvious confounding variable was ruled out, and no other obvious explanations for our observed results are apparent. Rather, it is our belief that the forgetting curves we reported are valid, and that they provide the most accurate depiction of the testing effect across long retention intervals.

3 Experiment 2B

Experiment 1 and 2A found that testing improved the retention of the information that was retrieved in the initial test along with information that was not retrieved but was integrated in the questions. The two studies also found that the level of retention for these two types of
information are equivalent when students responded correctly in the initial test, while the retention of the latter information was superior when students responded incorrectly in the initial test. Experiment 1 demonstrated these results in the laboratory setting, while Experiment 2A showed that these results are generalizable to educationally relevant materials and to long delay intervals. Experiment 2A also found that the benefit of testing remained constant over a six-week period despite overall forgetting of all items.

Experiment 2B was a follow-up of Experiment 2A. The aim of Experiment 2B was to explore whether the findings of Experiment 2A hold over a very long delay interval, even after the end of the introductory psychology course. Only a handful of experiments have examined testing effect over a retention interval of more than six weeks (Duchastel & Nungester, 1981; McDaniel et al., 2011; Roediger et al., 2011; Spitzer, 1939).

Spitzer (1939) demonstrated that the testing effect remains robust even after a delay of about two months. Yet, the data also suggested that the testing effect diminished as the retention performance of the students who did not take an initial test reached the asymptotic level while the retention performance of the students who took the test continue to drop gradually. Similarly, Duchastel and Nungester (1981) found a marginally significant effect of testing five months after the initial test (four and a half months after the first retention test), and they also found that the benefit of testing dropped when compared to the performance in the first retention test that was administrated two weeks after the initial test.

McDaniel et al. (2011) and Roediger et al. (2011) examined the effect of taking quizzes on retention performance on subsequent exams at Columbia Middle School. They found that taking quizzes did not only promote students’ performance on the chapter exams a few days after teacher’s lessons, it also improved students’ performance on the end-of-semester and end-of-
school-year exams which these exams took place months after the initial lessons. The researchers also found the magnitude of the testing effect diminishing over time although the effect remained significant after eight months.

3.1 Method

3.1.1 Participants

The participants were 27 university students who took the introductory psychology course in the previous semester at University of Toronto Scarborough. They were recruited four to five and a half months after the midterm test of the course. All participants took the regular midterm test rather than the make-up test. They participated in the experiment in exchange for course credit. Data from an additional participant was excluded due failure to complete the study appropriately.

3.1.2 Materials

The materials were the same as in Experiment 2A.

3.1.3 Procedure

All the participants took the computer version of the task, and the procedure was the same as in Experiment 2A.

3.2 Results

The accuracy of the 16 tested (verbatim and pre-modified) items in the midterm was 0.718 ($s = 0.166$). There was no significant difference between the verbatim ($\bar{x}_{Midterm\_verbatim} = 0.732, s = 0.201$) and pre-modified items ($\bar{x}_{Midterm\_pre-modified} = 0.704, s = 0.197$), $t(26) = 0.656, d = 0.126, p = 0.518$.

The experimental task performance for the three conditions (verbatim, modified, and new) are plotted in Figure 7, and were analyzed with a repeated-measures ANOVA. The analysis showed
a significant effect of Condition, $F(2, 52) = 4.353$, $MSE = 0.020$, $\eta_p^2 = 0.143$, $p = 0.018$. After the long delay, the performance on the tested items ($\bar{x}_{\text{verbatim + modified}} = 0.563$, $s = 0.169$) was only marginally better than the new items ($\bar{x}_{\text{new}} = 0.500$, $s = 0.177$), $t(26) = 1.871$, $d = 0.325$, $p = 0.073$. The participants still performed significantly better on the modified items ($\bar{x}_{\text{modified}} = 0.606$, $s = 0.213$) than on the verbatim items ($\bar{x}_{\text{verbatim}} = 0.519$, $s = 0.165$), $t(26) = 2.281$, $d = 0.499$, $p = 0.031$, and the new items, $t(26) = 2.761$, $d = 0.464$, $p = 0.010$. However, the difference between verbatim items and new items were no longer significant, $t(26) = 0.480$, $d = 0.096$, $p = 0.635$. Thus, the testing effect was disappearing, at least in terms of overall accuracy.
The following analysis focused on the effect of response accuracy of the midterm test on the performance of the verbatim and modified items. Only the participants who had data in all the cells were included in the analysis ($n = 23$). A $2 \times 2$ repeated measures ANOVA was conducted with both Accuracy (correct and incorrect) and Condition (verbatim and modified) as within-
subject factors. The summary data are presented in Figure 8. Visual examination of the figure suggests the benefit of testing is greater for the correct trials than for incorrect trials in general, and the benefit is equivalent for the verbatim and modified conditions for the correct trials while the effect is more robust for the modified than the verbatim condition for the incorrect trials.

Figure 8. Experiment 2B: Mean proportion correct of the verbatim and modified conditions for the initially correct and incorrect items in the experimental task.
The participants performed better in the experimental task if they responded correctly \( \bar{x}_{\text{correct}} = 0.618, s = 0.170 \) in the initial midterm test than if they responded incorrectly \( \bar{x}_{\text{incorrect}} = 0.366, s = 0.283 \), \( F(1, 22) = 13.998, MSE = 0.104, \eta^2_p = 0.389, p = 0.001 \). The analysis also showed a significant effect of Condition, \( F(1, 22) = 6.327, MSE = 0.055, \eta^2_p = 0.223, p = 0.020 \) and an Accuracy \( \times \) Condition interaction, \( F(1,22) = 5.467, MSE = 0.038, \eta^2_p = 0.199, p = 0.029 \). For the initially correct items, performance for the verbatim \( \bar{x}_{\text{verbatim/correct}} = 0.603, s = 0.227 \) and modified \( \bar{x}_{\text{modified/correct}} = 0.632, s = 0.197 \) items were similar, \( t(22) = -0.750, d = -0.111, p = 0.461 \), and again they were both significantly higher than the performance on the new condition (verbatim/correct vs. new: \( t(22) = 2.314, d = 0.482, p = 0.030 \); modified/correct vs. new: \( t(22) = 3.003, d = 0.626, p = 0.007 \)). For the initially incorrect items, the performance for the verbatim items \( \bar{x}_{\text{verbatim/incorrect}} = 0.257, s = 0.276 \) was significantly worse than the modified items \( \bar{x}_{\text{modified/incorrect}} = 0.475, s = 0.381 \), \( t(22) = -5.771, d = -0.626, p < 0.001 \). The performance for the verbatim items was also significantly worse than the performance for the new items, \( t(22) = -3.691, d = -0.770, p = 0.001 \), while the performance for the modified items and the new items were similar, \( t(22) = -0.109, d = -0.023, p = 0.914 \).

Again, the effect of incorrect responses and unselected lures from the midterm on performance on the verbatim items was analyzed. The persistence of errors was again influencing participants to pick the same incorrect response as in the midterm 0.319 \( s = 0.356 \) of the time. The size of this effect was significant less than that observed in Experiment 2A \( \bar{x} = 0.540, s = 0.377 \), \( t(184) = 2.693, d = 0.589, p = 0.008 \). Thus, the potency of the links formed during the midterm, links essentially representing misconceptions formed during testing, are weakening but are still fairly robust even after such a long interval. When the initial incorrect responses were not among the alternatives in the experimental task as in the modified items, the participants showed a much higher level of performance.
The participants responded with unselected lures from the midterm about 0.387 (s = 0.227) of the time if they answered the items in the midterm correctly, which was significantly higher than in Experiment 2A (x̄ = 0.207, s = 0.218), t(184) = 3.767, d = 0.824, p < 0.001. While participants picked unselected lures from the midterm 0.434 (s = 0.428) of the time if they answered the items incorrectly previously, which was also significantly higher than in Experiment 2A (x̄ = 0.209, s = 0.293), t(184) = 3.300, d = 0.722, p = 0.001. However, the difference between the two accuracy conditions remained insignificant, t(23) = -0.449, d = -0.092, p = 0.658 (only included the participants with data in all the cells). Again, it seems the different levels of accuracy across the verbatim and modified conditions were due entirely to the persistence of errors occurring on verbatim trials.

3.3 Discussion

As a follow-up study, Experiment 2B examined whether the findings of Experiment 2A are stable over a much longer retention interval. It is of particular interest whether the level of the retention benefit from testing and the asymmetry of the effect between the forward and the backward cueing directions would change over time.

The testing effect on overall accuracy was disappearing after the long retention delay. The result of Experiment 2B showed that the retention performance on the verbatim items diminished and no longer significantly exceeded the retention performance on the non-tested items. This result converged with the previous research (Duchastel & Nungester, 1981; McDaniel et al., 2011; Roediger et al., 2011; Spitzer, 1939), although some of the previous studies did find significant testing effects over very long retention intervals. All these studies suggest that the testing effect attenuate over time. The attenuation might be to be due to the retention of the non-tested materials reaching the asymptotic level while the retention of the tested materials continues to
Fazio, Agarwal, et al. (2010) also found that the effect of testing dropped with a longer retention interval; however, they found the drop occurred within a week. It is unclear whether their results stemmed from the retention of the non-tested materials reaching asymptotic level. Yet, their results did show that the rate of retention drop for the tested materials was substantially faster than the drop for the non-tested materials (19% vs. 8%).

In contrast, while the retention benefit of the retrieved information attenuated, the retention enhancement of the tested but non-retrieved information was still present. The students continued to perform better on the modified items than the non-tested items, primarily because of the immunity those items from the participants persisting to the initial erroneous responses.

Experiment 2B found that the persistence of errors, which drives the negative testing effect, was dwindling over time. The students showed a lower level of persistence of errors in Experiment 2B compared to 2A. They were less likely to select the incorrect responses they had chosen in the midterm test. This result is in line with the findings of Fazio, Agarwal, et al. (2010) which also suggests the negative effects of testing decrease as the retention interval increases. However, there was an increase in the intrusion of previous unselected lures. Yet, it was probably an artifact of the forced multiple-choice nature of the retention test.

The pattern of retention benefit was again found to be dependent on the performance in the initial test. The retention of the initially correct items was superior to the initially incorrect items. For the initially correct items, the level of retention was similar across the verbatim and modified items suggesting a symmetric retention benefit. However, for the initially incorrect items, the retention performance was significantly better for the modified items than the verbatim items, while the influence of persistence of errors only affect the verbatim items but not the modified
items. Additionally, intrusion of unselected lures was not related to the accuracy in the initial test. These results were consistent with the findings of Experiment 1 and 2A.

The overall story suggested by these experiments then is the following. Students are forming powerful associations between the questions and the answers they choose while writing a test. Correct choices lead to a positive testing effect while incorrect choices lead to a negative testing effect. Both of these can be explained by a simple strengthening of the association between question and chosen answer. The power of these associations is apparent even as long as four to five and a half months after testing, but their influence does weaken as evidenced by an increasing tendency to choose incorrect lures that were not chosen during the initial test. Clearly tests do not just tap learning, they also support the learning of new associations that, if allowed to occur unchecked, have both positive and negative effects that persist long after a test is completed.

One important point of all this is that it is not sufficient to assess the longevity of the testing effect by looking at overall accuracy alone given the two faces, that is the positive and negative, effects of testing. If one only looked at overall accuracy for verbatim items, it would appear as if the testing effect was non-existent after four to five and a half months. However, when one looks only at the items responded to correctly during the test, the testing effect is clearly still very robust, however so too is the negative effect of testing that is associated with initially incorrect responses. The two effects are simply more balanced after the long delay, leading to an illusion of no overall effect.
Chapter 3
Cue-Target Asymmetry and Prior Study

The experiments in the previous chapter demonstrated that the extent of the test-induced retention benefit is highly dependent on the accuracy in the initial test. The retention patterns for the items that were answered correctly in the initial multiple-choice test are substantially different from the patterns for the items that were answered incorrectly. Other than retaining considerably fewer items that were answered incorrectly initially, the test-induced retention benefit was also found to be asymmetric across the two cueing directions among the initially incorrect items. For those items, students showed a better retention performance and were influenced by the errors committed in the initial test to a lesser extent in the backward direction when prompting with the target words from the initial test than in the forward direction when prompting with the cue words from the initial test.

As argued, the positive and negative effects of testing might, in fact, result from the same process: testing strengthens the link between a cue and a response regardless of the objective correctness of the response. When the response is correct, the link between the cue and the target is strengthened, and retention in both the forward and backward directions are facilitated. However, when the response is incorrect, an asymmetrical effect is observed which appears to be attributable to some form of insulation to the persistence of errors that occurs when backward prompts are used. For example, when students are prompted backward with the target word from the initial multiple-choice test, if they remembered the information in the study phase or in the multiple-choice item, they could possibly come to realization of their error in the initial test. They could then escape from the persistence of errors and deduce the cue word based on either information from the study phase or from the initial multiple-choice test. This process is not
possible for the forward condition hence might have led to the asymmetry in the retention performance.

In this chapter and the next chapter, a series of experiments were conducted to further examine the retention asymmetry for the initial incorrect items. Particularly, the experiments focused on investigating various factors that might contribute or eliminate the retention asymmetry.

1 Experiment 3

The theoretical account provided above assumes that when students inaccurately associate the cue with a lure in the test phase by choosing it as the answer, and when they are subsequently provided with the correct target as a prompt for recall, they can somehow escape the association formed at the test and recall the correct cue. The aim of Experiment 3 was to examine whether students rely on the information from the study phase or the information from the initial multiple-choice test, or both, to arrive at the cue.

Specifically, Experiment 3 assessed the relevance of the study phase to this asymmetry by manipulating whether or not some of the word pairs are presented in the study phase, creating a study and a no study condition. As a result, some pairs were both studied and tested before the retention of these pairs was assessed in the cued-recall task, while other pairs were only tested before the assessment. If the ability to escape the association formed at test has its locus in the testing phase, the groups should not differ in terms of the retention asymmetry. But if the presentation of the “correct” association between cue and target during the study phase supports the ability to escape the preservation error, then we may see a greater asymmetry in the study versus no-study condition.
1.1 Method

1.1.1 Participants

The participants were 46 university students who were taking an introductory psychology course at University of Toronto Scarborough. They participated in the experiment in exchange for course credit.

1.1.2 Materials

The materials were the same as in Experiment 1.

1.1.3 Procedure

The experiment consisted of three phases: the study phase, the multiple-choice test phase, and the recall phase. In the study phase, the participants were given an opportunity to study the word pairs. The word pairs were presented at the center of a computer screen one pair at a time for three seconds each. Half of the experimental pairs were presented in a random order along with five of the fillers presented at the beginning of the sequence before the experimental pairs and the rest of the fillers presented at the end after all the experimental pairs. In total, there were 40 trials in the study phase. The left-right order of the words was randomized.

In the multiple-choice test phase, all 60 experimental word pairs were tested in a multiple-choice format. For each trial, one of the words in a word pair was randomly selected and appeared as a cue, and the participants were instructed to select the matching target, from four alternatives. The participants were informed that they might not have seen some of the pairs in the study phase, and they were instructed to study all the options and select the best option appeared to them. The pairs from the study and no study conditions were presented in a random order. Upon finishing, the participants were allowed to play Tetris for 15 minutes.
Lastly, in the recall phase, all 60 experimental word pairs were tested in a cued-recall format. In each trial, one of the words from an experimental pair was presented, and the participants were instructed to recall and type out the matching word. Half of the pairs from study and no study conditions were tested in a forward manner, while the rest was tested in a backward manner.

1.2 Results

The accuracy on the multiple-choice test was 0.471. The participants performed significantly better if they studied the pairs in the study phase ($\bar{x}_{MC, study} = 0.679, s = 0.149$ vs. $\bar{x}_{MC, no study} = 0.263, s = 0.084$), $F(1, 45) = 276.953, MSE = 0.029, \eta^2_p = 0.860, p < 0.001$. There was no significant performance difference between the pairs that were subsequently tested in the forward manner ($\bar{x}_{MC, forward} = 0.478, s = 0.089$) and the ones that were tested in the backward manner ($\bar{x}_{MC, backward} = 0.464, s = 0.116$), $F(1, 45) = 0.764, MSE = 0.013, \eta^2_p = 0.017, p = 0.387$. The interaction between study conditions and recall directions was not significant either, $F(1, 45) = 0.006, MSE = 0.013, \eta^2_p = 0.000, p = 0.938$.

The previous experiments found that the cued-recall performance for the items which students responded correctly in the initial multiple-choice test was distinctly different from those items students got incorrect initially. Therefore, subsequent analyses separated the cued-recall task performance based on the correctness of participants’ responses in the multiple-choice test. The data was analyzed using a $2 \times 2 \times 2$ repeated-measure ANOVA with Study (study and no study), Accuracy (correct and incorrect), and Direction (forward and backward) as within-subject factors. The analysis only included participants who had data in all the cells ($n = 43$). The data are presented in Figure 9.
Figure 9. Experiment 3: Mean proportion recalled for the two cueing directions for the no study and study conditions separately for the initially correct and incorrect items in the cued-recall task.

Again, the participants recalled more matching pairs if they responded correctly in the initial multiple-choice test ($\bar{x}_{\text{correct}} = 0.588, s = 0.152$) than if they responded incorrectly ($\bar{x}_{\text{incorrect}} = 0.174, s = 0.102$), $F(1, 42) = 316.873, MSE = 0.046, \eta^2_p = 0.883, p < 0.001$. In term of the impact
of studying, the participants performed better if they were given chance to study the materials ($x_{study} = 0.450, s = 0.129$ vs. $x_{no \ study} = 0.312, s = 0.130$), $F(1, 42) = 35.063, MSE = 0.046, \eta_p^2 = 0.455, p < 0.001$. The Study conditions also interacted with Accuracy, $F(1, 42) = 17.746, MSE = 0.041, \eta_p^2 = 0.297, p < 0.001$. Prior studying led to a better retention for the correct items from the multiple-choice test, $t(42) = 5.260, d = 0.932, p < 0.001$. Note that correctness of a response for the study and no study conditions might have different meanings to the participants; in the no study condition, correctness is arbitrary as the participants just picked an answer in the initial multiple-choice test with no tangible basis.

Although Study condition did not interact with Direction in any way according to the ANOVA analysis, neither Study × Direction, $F(1, 42) = 1.984, MSE = 0.040, \eta_p^2 = 0.045, p = 0.166$, nor the three way interaction, $F(1, 42) = 0.087, MSE = 0.020, \eta_p^2 = 0.002, p = 0.770$, was significant, a few apriori contrasts were conducted. The results for which the participants were given chance to study the pairs in the study phase are consistent with Experiment 1. For the correct trials, the cued-recall performance in the forward ($x_{study/forward/correct} = 0.719, s = 0.161$) and the backward directions ($x_{study/backward/correct} = 0.686, s = 0.208$) were similar and symmetrical, $t(42) = 1.101, d = 0.158, p = 0.277$, while, for the incorrect trials, the recall performance in the backward direction ($x_{study/backward/incorrect} = 0.317, s = 0.276$) was better than the forward direction ($x_{study/forward/incorrect} = 0.077, s = 0.132$), $t(42) = 7.889, d = 0.807, p < 0.001$.

However, the results for the no study conditions were a little bit different. For the incorrect trials, the asymmetry between the two cueing directions was still present; the recall performance in the backward direction ($x_{no \ study/backward/incorrect} = 0.246, s = 0.184$) was better than the forward direction ($x_{no \ study/forward/incorrect} = 0.057, s = 0.072$), $t(42) = 6.185, d = 0.946, p < 0.001$. Surprisingly, an asymmetry in an opposite direction was found for the correct trials; the cued-
recall performance in the forward direction ($\bar{x}_{\text{no study}/\text{forward/correct}} = 0.524, s = 0.241$) showed a superior performance than the backward direction ($\bar{x}_{\text{no study}/\text{backward/correct}} = 0.421, s = 0.313$), $t(42) = 3.394, d = 0.327, p = 0.002$.

Next, the level of asymmetry (performance difference between the forward and the backward directions) between the study and the no study conditions was compared. The level of asymmetry was similar for the initially incorrect trials, $t(42) = -1.705, d = -0.159, p = 0.096$, suggesting prior studying did not mitigate the performance difference between forward and backward conditions. But, for the initially correct trials, there was a significantly bias toward the forward direction for the no study condition, $t(42) = -2.293, d = -0.191, p = 0.027$.

In terms of the negative impacts of testing, the participants were more likely to recall the incorrect options they selected in the multiple-choice test when they were prompted in the forward direction ($\bar{x}_{\text{incorrect response/forward}} = 0.473, s = 0.200$) compared to the backward direction ($\bar{x}_{\text{incorrect response/backward}} = 0.085, s = 0.110$), $F(1, 42) = 136.019, MSE = 0.048, \eta_p^2 = 0.764, p < 0.001$. The likelihood of recalling the incorrect response from the multiple-choice was not affected by the Study conditions ($\bar{x}_{\text{incorrect response/study}} = 0.292, s = 0.175$ vs. $\bar{x}_{\text{incorrect response/no study}} = 0.266, s = 0.124$), $F(1, 42) = 0.822, MSE = 0.035, \eta_p^2 = 0.019, p = 0.370$.

The participants were again more likely to recall an unselected lures if they were prompted in the forward direction ($\bar{x}_{\text{unselected lure/forward}} = 0.099, s = 0.064$) than in the backward direction ($\bar{x}_{\text{unselected lure/backward}} = 0.045, s = 0.054$), $F(1, 42) = 22.108, MSE = 0.011, \eta_p^2 = 0.345, p < 0.001$. The main effect of Accuracy ($\bar{x}_{\text{unselected lure/correct}} = 0.091, s = 0.0797$ vs. $\bar{x}_{\text{unselected lure/incorrect}} = 0.053, s = 0.0394$), $F(1, 42) = 8.047, MSE = 0.015, \eta_p^2 = 0.161, p = 0.007$, and Study ($\bar{x}_{\text{unselected lure/study}} = 0.046, s = 0.041$ vs. $\bar{x}_{\text{unselected lure/no study}} = 0.098, s = 0.070$), $F(1, 42) = 22.188, MSE = 0.010, \eta_p^2 = 0.346, p < 0.001$, along with their interaction, $F(1, 42) = 4.318, MSE = 0.010, \eta_p^2 = 0.093, p =$
0.044, were all significant. These results seem to be driven by a considerably higher level of unselected lure recall for the initially correct items in the no study condition ($\bar{x} = 0.128, s = 0.132$).

1.3 Discussion

Experiment 3 investigated whether the retention asymmetry of the initially incorrect items found in previous experiments is a product of the knowledge acquired from the study phase or the information obtained during the initial multiple-choice testing or both. Experiment 3 was designed to isolate the effects of these two components by manipulating the possibility of information being acquired during the study phase.

Overall, Experiment 3 essentially replicated the findings of Experiment 1. The retention benefit of testing was found to be strongly dependent on the performance in the initial test. The students were more likely to recall the items that they got correct in the initial test. When given a chance to study the materials, the retention benefit was found to be symmetric across a cue and its target for the initially correct items. Yet, for the initially incorrect items, there was a strong bias towards more accurate retention in the backward direction. The students were able to escape from the persistence of errors and re-discover the correct answers in the backward direction while they were performing at a very poor level in the forward direction.

The manipulation of study list presence gave rise to several effects. Without studying, the students performed worse in the initial multiple-choice test as would be expected. Experiment 3 also showed that having no prior study reduces the effect of testing on retention. For the initially correct items, the retention performance was significantly better if the students studied the items previously than if they did not see those items in the study phase. Such a difference was not
found in the initially incorrect items. These findings are consistent with Butler and Roediger (2008).

The retention asymmetry between the two recall directions for the initially incorrect items was found for both study and no study conditions. The extent of the asymmetry was similar which suggests that the asymmetry stems primary from the multiple-choice testing phase. When the students were prompted in the backward direction with the target from the multiple-choice test, if they remembered the multiple-choice item and realized they made an error in the initial test, the findings of Experiment 3 seem to suggest that the students rely mostly on the memory of information available in the initial multiple-choice test to recover the correct answer. For example, they may use associations between the actual target and lures to retrieve the lure that they chose, and then use this lure to recall the associated cue. This mechanism might be similar to how students correct errors committed in an initial multiple-choice test, with the aid of verification feedbacks according to Marsh et al. (2012).

Surprisingly, Experiment 3 also found the retention performance on the forward and the backward directions were slightly asymmetrical for the initially correct items that were never presented in the study phase. The effect was small but significant. This might be because, in absence of prior studying, the word pairs were only presented in one possible spatial configuration with the cue presented on the left with the response presented on the right during the multiple choice test. The students might have encoded the positional information associated with each word. These positional associations might have caused the bias on the recall task performance in favor of the forward direction over the backward direction.

In terms of the negative testing effect, Experiment 3 again found that most of the intrusions came from the students re-producing the incorrect responses from the multiple-choice test, and the
intrusion of the unselected lures was relatively rare. Students showed a much higher level of
persistence of errors from the initial test in the forward condition than in the backward condition.
This was true regardless of the present of prior study. Experiment 3 also found that students were
more likely to recall unselected lures in the forward direction than in the backward direction.
Interestingly, having no prior studying increased the likelihood that students would recall
unselected lures for the initially correct items.
Chapter 4
Cue-Target Asymmetry and Feedback

The previous experiments reported in this thesis described the positive and negative effects that occur when students perform a typical multiple-choice test, and the processes they might use to escape the negative effects of testing on incorrect trials. However, in all of the previous studies the students did not know the accuracy of their responses during the test, and this lack of knowledge seems to explain why strong associations were formed for both accurate and inaccurate choices.

An optimal test would be one that allows the positive effects of testing but prevents the negative effects, and providing feedback seems like an obvious mechanism for achieving this goal. Thus, the experiments reported subsequently in this thesis will focus on the impact of various forms of feedback on the positive and negative effects of testing.

One major distinction that one can make in terms of feedback is the distinction between verification versus answer feedback. Verification feedback involves informing the students whether they are correct or incorrect, but goes no further. Answer feedback involves showing students the correct answer, thereby not only informing the student of their accuracy but also informing them of which answer is accurate.

1 Experiment 4A

So far, it was found that the test-induced retention benefit is asymmetric across the two cueing directions if students answered the initial multiple-choice items incorrectly. Specifically, students showed a better retention performance and a lower level of influence from the persistence of errors for the initially incorrect items in the backward direction when prompted with the target
words from the initial test than in the forward direction when prompted with the cue words. It was speculated that when students made a mistake in the initial test and were subsequently prompted in the backward direction in the cued-recall task, they come to realize their mistake and they attempt to recover the correct answer based on the information available in the initial multiple-choice test. This mechanism might be similar to how students correct errors committed in an initial multiple-choice test, with the aid of verification feedback according to Marsh et al. (2012). However, when students were prompted in the forward direction, they were provided with no information to judge their response accuracy in the initial test, making error correction very difficult, if not impossible. Whether students are able to recognize an error they have committed might be a critical factor that sets apart the cued-recall performance in the forward and backward directions.

The goal of Experiment 4A was to examine whether students’ knowledge of the accuracy of their own responses in the initial multiple-choice test is a contributing factor that leads to the retention asymmetry between the forward and backward directions for the initially incorrect items. To make the information about accuracy available for both cueing directions, verification feedback was implemented. Students were provided with a verification feedback after completing a multiple-choice item for half of the trials. It is of interest whether providing verification feedback would lessen level of asymmetry or eliminate it all together.

1.1 Method

1.1.1 Participants

The participants were 40 introductory psychology students from University of Toronto Scarborough. They participated in the experiment in exchange for course credit.
1.1.2 Materials

The materials were the same as in Experiment 1.

1.1.3 Procedure

The experiment consisted of three phases: the study phase, the multiple-choice test phase, and the recall phase. The study phase was the same as in Experiment 1.

In the multiple-choice test phase, all 60 experimental word pairs were tested in a multiple-choice format. For each trial, one of the words in a word pair was randomly selected and appeared as a cue, and the participants were instructed to select the matching target, from four alternatives. Then, for half of the trials, the participants were given feedback indicating if their response was correct or incorrect. The feedback was presented at the center of the screen for two seconds. To control for the length of the trials across conditions, in the no feedback condition, the participants were presented with a blank screen for two seconds instead. The feedback and no feedback conditions were administrated in a random presentation order. Upon finishing, the participants were allowed to play Tetris for 15 minutes.

Finally, in the recall test phase, all 60 experimental word pairs were tested in a cued-recall format. On each trial, one of the words from an experimental pair was presented, and the participants were instructed to recall and type out the matching word. Half of the pairs from feedback and no feedback conditions were test in a forward manner, while the rest was tested in a backward manner.

1.2 Results

The accuracy on the multiple-choice test was 0.717 ($s = 0.178$). There was no significant difference in accuracy among the conditions (Feedback: $F(1, 39) = 0.244, MSE = 0.013, \eta^2_p =$
0.006, \( p = 0.624 \); Direction: \( F(1, 39) = 1.912, MSE = 0.014, \eta^2_p = 0.047, p = 0.174 \); Feedback ×
Direction: \( F(1, 39) = 0.001, MSE = 0.013, \eta^2_p = 0.000, p = 0.978 \).

The previous experiments demonstrated that the cued-recall performance for the items which
students responded correctly in the initial multiple-choice test was distinctly different from those
which students initially got wrong. Therefore, subsequent analyses separated the cued-recall task
performance based on the correctness of the participants’ responses in the multiple-choice test.
The data was analyzed using a \( 2 \times 2 \times 2 \) repeated-measure ANOVA with Accuracy (correct and
incorrect), Direction (forward and backward) and Feedback (feedback and no feedback) as
within-subject factors. The analysis only included the participants who had data in all the cells (\( n = 32 \)). The resulting data is presented in Figure 10.
Figure 10. Experiment 4A: Mean proportion recalled for the two cueing directions for the no feedback and feedback conditions separately for the initially correct and incorrect items in the cued-recall task.

The ANOVA revealed that the participants recalled more matching pairs if they responded correctly ($\bar{x}_{\text{correct}} = 0.686, s = 0.186$) in the initial multiple-choice test than if they responded incorrectly ($\bar{x}_{\text{incorrect}} = 0.219, s = 0.129$), $F(1, 31) = 270.046$, $MSE = 0.052$, $\eta_p^2 = 0.897$, $p <$
The results also showed a significant effect of Direction, \( F(1, 31) = 18.553, \text{MSE} = 0.021, \eta_p^2 = 0.374, p < 0.001 \), and the expected Accuracy \times Direction interaction, \( F(1, 31) = 19.492, \text{MSE} = 0.036, \eta_p^2 = 0.386, p < 0.001 \). For the correct trials, the retention performance in the forward (\( \bar{x}_{\text{forward/correct}} = 0.699, s = 0.234 \)) and the backward (\( \bar{x}_{\text{backward/correct}} = 0.673, s = 0.170 \)) directions were similar, \( t(31) = 0.551, d = 0.154, p = 0.586 \), while, for the incorrect trials, the retention performance in the backward direction (\( \bar{x}_{\text{backward/incorrect}} = 0.310, s = 0.175 \)) was significantly better than the forward direction (\( \bar{x}_{\text{forward/incorrect}} = 0.128, s = 0.131 \)), \( t(31) = 3.864, d = 1.095, p < 0.001 \).

The results also showed that providing verification feedback in the multiple-choice test did not affect the results in any systematical way. Neither the main effect of Feedback, \( F(1, 31) = 0.005, \text{MSE} = 0.040, \eta_p^2 = 0.000, p = 0.943 \), nor any of the interactions including Feedback as a factor was significant (Accuracy \times Feedback, \( F(1, 31) = 0.660, \text{MSE} = 0.032, \eta_p^2 = 0.021, p = 0.423 \); Direction \times Feedback, \( F(1, 31) = 0.415, \text{MSE} = 0.052, \eta_p^2 = 0.013, p = 0.524 \); Accuracy \times Direction \times Feedback, \( F(1, 31) = 0.379, \text{MSE} = 0.061, \eta_p^2 = 0.012, p = 0.543 \)).

The participants were more likely to recall the incorrect options they selected in the multiple-choice test when they were prompted in the forward direction (\( \bar{x}_{\text{incorrect response-forward}} = 0.357, s = 0.237 \)) compared to the backward direction (\( \bar{x}_{\text{incorrect response-backward}} = 0.058, s = 0.087 \)), \( F(1, 31) = 52.219, \text{MSE} = 0.055, \eta_p^2 = 0.627, p < 0.001 \). The likelihood of recalling the incorrect response from the multiple-choice was not affected by the verification feedback (\( \bar{x}_{\text{incorrect response-no feedback}} = 0.196, s = 0.170 \) vs. \( \bar{x}_{\text{incorrect response-feedback}} = 0.218, s = 0.174 \)), \( F(1, 31) = 0.329, \text{MSE} = 0.046, \eta_p^2 = 0.010, p = 0.570 \).

The participants were also more likely to recall unselected lures if they were prompted in the forward direction (\( \bar{x}_{\text{unselected lure-forward}} = 0.084, s = 0.080 \)) than in the backward direction (\( \bar{x}_{\text{unselected}} \)).
$lure_{backward} = 0.046, s = 0.057), F(1, 31) = 3.675, MSE = 0.024, \eta_p^2 = 0.106, p = 0.065$. While the likelihood of recalling an unselected lure was not related to the accuracy in the initial test ($\bar{x}\_unselected\_lure\_correct = 0.050, s = 0.041$ vs. $\bar{x}\_unselected\_lure\_incorrect = 0.081, s = 0.088$), $F(1, 31) = 2.662$, $MSE = 0.023, \eta_p^2 = 0.079, p = 0.113$, or the present of verification feedback, ($\bar{x}\_unselected\_lure\_no\_feedback = 0.067, s = 0.073$ vs. $\bar{x}\_unselected\_lure\_feedback = 0.063, s = 0.062$), $F(1, 31) = 0.040, MSE = 0.021, \eta_p^2 = 0.001, p = 0.843$.

1.3 Discussion

Experiment 4A investigated whether the knowledge of the accuracy of individual responses in the initial multiple-choice test modulates the retention asymmetry between the forward and the backward directions for the initially incorrect items. Because this information is only implicitly available for the backward cueing direction, in this experiment, I attempted to equate the two cueing conditions by making the information explicitly available for both cueing directions through the provision of verification feedback. For half of the trials in the multiple-choice test phase, the students were informed whether their responses were accurate after they had made their responses.

The results of the no feedback condition replicated the basic findings of previous experiments. The future retention is highly dependent on the accuracy in the initial test. The retention was significantly better for the items that were successfully retrieved in the multiple-choice test. Additionally, the retention benefit was found to be symmetric across a cue and its pairing target for the initially correct items, while there was a strong bias towards the backward direction for the initially incorrect items. Furthermore, the students recalled more erroneous responses and unselected lures from the initial test in the forward condition compared to the backward
condition. The likelihood to recall unselected lures was again not related to the accuracy in the initial test.

Experiment 4A did not find the verification feedback provided any extra retention benefit to the items that were initially tested in the multiple-choice test. The cued-recall task performance for the feedback condition and no feedback condition were essentially the same. This result was inconsistent with the recent study conducted by Marsh et al. (2012). Marsh et al. (2012) found that providing verification feedbacks on a multiple-choice test, although not as efficient as answer feedback, did have a positive impact on the future retention. They found that verification feedback can help to maintain correct responses and correct errors. Neither of these was evidenced in Experiment 4A. There was no difference between the feedback and no feedback conditions in term of retention performance (for both forward and backward directions), persistence of errors, and intrusion of unselected lures regardless of the accuracy in the initial multiple-choice test.

The discrepancy of the findings from Experiment 4A and Marsh et al. (2012) could be due to the way the feedback was presented. In Marsh et al. (2012), the verification feedback was presented for five seconds and was presented along with the multiple-choice item, the question and the alternatives, after students had made a response, while in Experiment 4A, the verification feedback was presented only for two seconds, and it was presented alone without the multiple-choice item. Marsh et al. (2012) argued that it is crucial to present multiple-choice alternatives along with the verification feedback; otherwise, the usefulness of the feedback would largely depend on students’ ability to remember the multiple-choice alternatives. It is also possible that students, in fact, needed more than two seconds to process the verification feedback.
2 Experiment 4B

Experiment 4A found that providing verification feedback does not induce any extra retention benefit to the information that was initially tested in the multiple-choice test nor did feedback decrease or eliminate the retention asymmetry between the forwarding and backward cueing direction for the initially incorrect items. Experiment 4B explored whether the null results were due to a sub-optimal implementation of the verification feedback. Experiment 4B adopted the methodologies used in Marsh et al. (2012). In this experiment, the verification feedback was presented for a longer duration (i.e., five seconds) and was presented along with the original multiple-choice items.

2.1 Method

2.1.1 Participants

The participants were 42 introductory psychology students from University of Toronto Scarborough. They participated in the experiment in exchange for course credit. Data from an additional participant was excluded because of an experimental error.

2.1.2 Materials

The materials were the same as in Experiment 1.

2.1.3 Procedure

The procedure was almost the same as in Experiment 4A, except that the feedback was presented along with the original multiple-choice item, the question and the alternatives, for five seconds. By contrast, in the no feedback condition, only the multiple-choice item was presented on the screen for five seconds.
2.2 Results

The accuracy on the multiple-choice test was 0.736 (s = 0.151). The participants performed better on the pairs that were subsequently tested in the forward manner ($\bar{x}_{MC\_forward} = 0.751$, s = 0.141) than the pairs that were tested in the backward manner ($\bar{x}_{MC\_backward} = 0.718$, s = 0.173), $F(1, 41) = 4.961$, $MSE = 0.009$, $\eta_p^2 = 0.108$, $p = 0.031$. None of the other effects was significant (Feedback: $F(1, 41) = 0.329$, $MSE = 0.012$, $\eta_p^2 = 0.008$, $p = 0.569$; Feedback × Direction: $F(1, 41) = 0.820$, $MSE = 0.015$, $\eta_p^2 = 0.020$, $p = 0.371$).

The previous experiments demonstrated that the cued-recall performance for the items which students responded correctly in the initial multiple-choice test was distinctly different from those which the students initially got wrong. Therefore, subsequent analyses separated the cued-recall task performance based on the correctness of the participants’ responses in the multiple-choice test. The data was analyzed using a $2 \times 2 \times 2$ repeated-measure ANOVA with Accuracy (correct and incorrect), Direction (forward and backward) and Feedback (feedback and no feedback) as within-subject factors. The analysis only included the participants who had data in all the cells ($n = 35$). The data is presented in Figure 11.
Figure 11. Experiment 4B: Mean proportion recalled for the two cueing directions for the no feedback and feedback conditions separately for the initially correct and incorrect items in the cued-recall task.

The ANOVA revealed that the participants recalled more matching pairs if they responded correctly ($\bar{x}_{correct} = 0.737, s = 0.151$) in the initial multiple-choice test than if they responded incorrectly ($\bar{x}_{incorrect} = 0.229, s = 0.140$), $F(1, 34) = 391.181$, $MSE = 0.046$, $\eta_p^2 = 0.920$, $p <$
The results also showed a significant effect of Direction, $F(1, 34) = 6.663$, $MSE = 0.058$, $\eta_p^2 = 0.164$, $p = 0.014$, and the expected Accuracy $\times$ Direction interaction, $F(1, 34) = 13.941$, $MSE = 0.048$, $\eta_p^2 = 0.291$, $p = 0.001$. For the correct trials, the retention performance in the forward ($\bar{x}_{\text{forward/correct}} = 0.749$, $s = 0.164$) and the backward ($\bar{x}_{\text{backward/correct}} = 0.726$, $s = 0.165$) directions were similar, $t(34) = 0.443$, $d = 0.181$, $p = 0.661$, while for the incorrect trials, the retention performance in the backward direction ($\bar{x}_{\text{backward/incorrect}} = 0.315$, $s = 0.257$) was significantly better than the forward direction ($\bar{x}_{\text{forward/incorrect}} = 0.143$, $s = 0.314$), $t(34) = 3.291$, $d = 0.575$, $p = 0.002$.

The results again suggested that providing verification feedback in the multiple-choice test did not affect the results in any systematical way. Neither the main effect of Feedback, $F(1, 34) = 0.858$, $MSE = 0.034$, $\eta_p^2 = 0.025$, $p = 0.361$, nor any of the interactions including Feedback as a factor was significant (Accuracy $\times$ Feedback, $F(1, 34) = 0.256$, $MSE = 0.037$, $\eta_p^2 = 0.007$, $p = 0.616$; Direction $\times$ Feedback, $F(1, 34) = 0.000$, $MSE = 0.030$, $\eta_p^2 = 0.000$, $p = 1.000$; Accuracy $\times$ Direction $\times$ Feedback, $F(1, 31) = 0.012$, $MSE = 0.024$, $\eta_p^2 = 0.000$, $p = 0.915$).

The participants were more likely to recall the incorrect options they selected in the multiple-choice test when they were prompted in the forward direction ($\bar{x}_{\text{incorrect response forward}} = 0.341$, $s = 0.223$) compared to the backward direction ($\bar{x}_{\text{incorrect response backward}} = 0.049$, $s = 0.094$), $F(1, 34) = 57.726$, $MSE = 0.052$, $\eta_p^2 = 0.629$, $p < 0.001$. The likelihood of recalling the incorrect response from the multiple-choice was not affected by the verification feedback ($\bar{x}_{\text{incorrect response no feedback}} = 0.191$, $s = 0.128$ vs. $\bar{x}_{\text{incorrect response feedback}} = 0.199$, $s = 0.205$), $F(1, 34) = 0.047$, $MSE = 0.051$, $\eta_p^2 = 0.001$, $p = 0.830$.

The participants were more likely to recall unselected lures if they were prompted in the forward direction ($\bar{x}_{\text{unselected lure forward}} = 0.108$, $s = 0.082$) than in the backward direction ($\bar{x}_{\text{unselected lure backward}} = 0.114$, $s = 0.089$), $t(34) = 2.003$, $d = 0.283$, $p = 0.050$.
lure\_backward = 0.033 \pm 0.041), \ F(1, 34) = 25.376, MSE = 0.015, \ \eta^2_p = 0.427, \ p < 0.001. \ They were also more likely to recall an unselected lure if their response in the initial test was incorrect (\ \bar{\chi}_{\text{unselected lure\_correct}} = 0.048, \ s = 0.044 \ \text{vs.} \ \bar{\chi}_{\text{unselected lure\_incorrect}} = 0.092, \ s = 0.075), \ F(1, 31) = 11.728, MSE = 0.012, \ \eta^2_p = 0.256, \ p = 0.002. \ However, the likelihood of recalling an unselected lure was not related to the present of the verification feedback, (\ \bar{\chi}_{\text{unselected lure\_no\ feedback}} = 0.076, \ s = 0.063 \ \text{vs.} \ \bar{\chi}_{\text{unselected lure\_feedback}} = 0.065, \ s = 0.052), \ F(1, 34) = 0.932, MSE = 0.009, \ \eta^2_p = 0.027, \ p = 0.341.

2.3 Discussion

As in Experiment 4A, Experiment 4B was conducted to investigate whether the knowledge of the accuracy of individual responses in the initial multiple-choice test contributes to the retention asymmetry between the forward and the backward cueing directions for the initially incorrect items. Because this information is absent for the forward cueing direction, verification feedback was provided in an attempt to equate the two cueing conditions in this aspect. Specifically, Experiment 4B explored whether the null findings in Experiment 4A was a result of a sub-optimal implementation of the verification feedback.

The results of Experiment 4B were essentially the same as those of Experiment 4A. Again, Experiment 4B failed to replicate the findings of Marsh et al. (2012), verification feedback did not provide any extra retention benefit to the items that were initially tested in the multiple-choice test. Additionally, providing verification feedback did not improve the retention performance for the forward direction or the backward direction, and they did not decrease or eliminate the retention asymmetry for the initially incorrect items.

The difference in study materials used could be a possible cause for the result discrepancies. Marsh et al. (2012) used prose materials in their experiments rather than paired associates.
Compared to paired associates, prose materials provide more contextual information and might provide more indirect retrieval routes, allowing a better chance of recovering the correct answer when students realized their initial responses were incorrect.

While this reason might be valid, it could not account for the complete ineffectiveness of the verification feedback. To be able to recover the correct answers in the backward cueing direction, the students have to be able to remember information from the multiple-choice items as well. Given students were able to recover some of the correct answers in the backward cueing direction, it is extremely surprising that the verification feedback did not provide any observable benefit at all in the current study.

The complete failure of verification feedback for correcting errors in the forward cueing direction suggests that the asymmetry between the forward and backward direction, or the task demand required to recover correct answers in the two directions cannot be solely explained by the knowledge of the accuracy of the responses in the initial test.

On a more theoretical level, it appears the associations that are formed when a student chooses an incorrect response are relatively impervious to change, even when students are immediately informed they responded incorrectly and when they are allowed to see other possible responses. Perhaps knowing one is wrong is simply not enough; perhaps one must also know what the correct answer is.

3 Experiment 5

Experiment 4A and 4B showed that verification feedback was not effective in promoting the retention benefit and reducing the negative effects of testing. It neither helped to maintain the correct responses nor correct the erroneous responses. Furthermore, it did not eliminate or lessen
the retention asymmetry for the initial incorrect items between forward and backward cueing direction. The results of Experiment 4A and 4B suggest that the retention asymmetry cannot be completely accounted by the knowledge of the accuracy in the initial test. Experiment 5 explored if the retention asymmetry can be eliminated or lessened in any way after it has been induced by providing precise answer feedback.

Answer feedback, in which the feedback message includes the correct answer, has been found to be extremely effective in promoting future retention (e.g., Butler et al., 2007, 2008). If the asymmetry between the two cueing directions is a result of recognizing the errors committed in the initial multiple-choice test and using the limited information from the initial test to deduce and re-discover the correct answers in the backward direction, re-presenting the word pairs should provide enough information to eliminate the advantage of the backward condition over the forward condition. Experiment 5 examined whether providing answer feedback would eliminate the retention asymmetry of the initially incorrect items. The methodology of Experiment 5 was identical to Experiment 4A except answer feedbacks were presented instead of verification feedback.

3.1 Method

3.1.1 Participants

The participants were 58 introductory psychology students from University of Toronto Scarborough. They participated in the experiment in exchange for course credit. Data from an additional participant was dropped due to failure to complete the experiment.

3.1.2 Materials

The materials were the same as in Experiment 1.
3.1.3 Procedure
The procedure was the same as in Experiment 4A, except there was a modification for the feedback condition in the multiple-choice test phase. Rather than indicating the correctness of the response, the actual word pair was presented for two seconds in the middle of the screen as the feedback.

3.2 Results
The accuracy on the multiple-choice test was 0.726 (s = 0.165). The participants performed significantly better on the feedback conditions ($\bar{x}_{MC\_feedback} = 0.745, s = 0.172$) than the no feedback conditions ($\bar{x}_{MC\_no\_feedback} = 0.706, s = 0.179$), $F(1, 57) = 6.227$, $MSE = 0.014$, $\eta_p^2 = 0.098$, $p = 0.016$, while the performance did not differ between the forward ($\bar{x}_{MC\_forward} = 0.734, s = 0.170$) and backward directions ($\bar{x}_{MC\_backward} = 0.717, s = 0.178$), $F(1, 57) = 1.465$, $MSE = 0.012$, $\eta_p^2 = 0.025$, $p = 0.231$, nor did the two factors interact, $F(1, 57) = 0.389$, $MSE = 0.014$, $\eta_p^2 = 0.007$, $p = 0.535$.

Given the cued-recall performance largely depended on the performance in the initial multiple-choice test according to the previous experiments, subsequent analyses separated the cued-recall task performance based on the correctness of the participants’ responses in the multiple-choice test. Only the participants who had data in all the cells were included in the analysis ($n = 46$). A $2 \times 2 \times 2$ repeated-measure ANOVA was conducted with Accuracy (correct and incorrect) and Direction (forward and backward) and Feedback (feedback and no feedback) as within-subject factors. The summary data was presented in Figure 12.
Figure 12. Experiment 5: Mean proportion recalled for the two cueing directions for the no feedback and feedback conditions separately for the initially correct and incorrect items in the cued-recall task.

The participants recalled more matching pairs if they responded correctly ($\bar{x}_{correct} = 0.712, s = 0.173$) in the initial multiple-choice test than if they responded incorrectly ($\bar{x}_{incorrect} = 0.365, s = 0.191$), $F(1, 45) = 222.799, MSE = 0.050, \eta^2_p = 0.832, p < 0.001$. 
In terms of the effect of feedback, the results showed that providing answer feedback improves subsequence recall performance, $F(1, 45) = 61.186$, $MSE = 0.049$, $\eta^2_p = 0.576$, $p < 0.001$, and the improvement was dependent on the accuracy in the initial test, $F(1, 45) = 36.351$, $MSE = 0.046$, $\eta^2_p = 0.447$, $p < 0.001$. Answer feedback largely improved the recall performance on the items that were previously answered incorrectly ($\bar{x}_{feedback/incorrect} = 0.522$, $s = 0.290$ vs. $\bar{x}_{no\ feedback/incorrect} = 0.207$, $s = 0.164$), $t(45) = 7.067$, $d = 1.134$, $p < 0.001$, while the improvement for the items that the participants got correct initially was not significant ($\bar{x}_{feedback/correct} = 0.735$, $s = 0.180$ vs. $\bar{x}_{no\ feedback/correct} = 0.689$, $s = 0.190$), $t(45) = 1.038$, $d = 0.349$, $p = 0.305$.

Although the three way interaction was not significant, $F(1, 45) = 2.954$, $MSE = 0.039$, $\eta^2_p = 0.062$, $p = 0.093$, further analyses were conducted to examine apriori predictions. The results for which the participants were given no feedback on their responses in the multiple-choice test are consistent with the previous experiments. For the correct trials, the cued-recall performance in the forward ($\bar{x}_{forward/no\ feedback/correct} = 0.683$, $s = 0.238$) and the backward directions ($\bar{x}_{backward/no\ feedback/correct} = 0.694$, $s = 0.183$) were similar and symmetrical, $t(45) = -0.253$, $d = -0.055$, $p = 0.801$, while for the incorrect trials, the recall performance in the backward direction ($\bar{x}_{backward/no\ feedback/incorrect} = 0.338$, $s = 0.309$) was better than the forward direction ($\bar{x}_{forward/no\ feedback/incorrect} = 0.077$, $s = 0.137$), $t(45) = 6.326$, $d = 0.751$, $p < 0.001$.

The answer feedback eliminated the asymmetry between the forward ($\bar{x}_{forward/feedback/incorrect} = 0.482$, $s = 0.362$) and backward directions ($\bar{x}_{backward/feedback/incorrect} = 0.562$, $s = 0.334$) for the items that were previously answered incorrectly, $t(45) = -1.940$, $d = -0.209$, $p = 0.059$. For the initially correct items, the recall performance for the two cueing directions remained similar ($\bar{x}_{forward/feedback/correct} = 0.749$, $s = 0.192$ vs. $\bar{x}_{backward/feedback/correct} = 0.721$, $s = 0.216$), $t(45) = 0.696$, $d = 0.148$, $p = 0.490$. 
The participants were again more likely to recall the incorrect options they selected in the multiple-choice test when they were prompted in the forward direction ($\bar{x}_{\text{incorrect response forward}} = 0.255, s = 0.200$) compared to the backward direction ($\bar{x}_{\text{incorrect response backward}} = 0.045, s = 0.062$), $F(1, 45) = 48.573, MSE = 0.042, \eta^2_p = 0.519, p < 0.001$. Presence of feedback decreased the likelihood of recalling the same incorrect response from the multiple-choice test when they were cued in the forward direction ($\bar{x}_{\text{incorrect response forward/no feedback}} = 0.366, s = 0.298$ vs. $\bar{x}_{\text{incorrect response forward/feedback}} = 0.144, s = 0.236$), $t(45) = 5.697, d = 0.621, p < 0.001$, while the feedback had no effect when participants were cued in backward direction ($\bar{x}_{\text{incorrect response backward/no feedback}} = 0.040, s = 0.071$ vs. $\bar{x}_{\text{incorrect response backward/feedback}} = 0.050, s = 0.107$), $t(45) = 0.280, d = 0.082, p = 0.781$.

The participants were also more likely to recall unselected lures if they were prompted in the forward direction ($\bar{x}_{\text{unselected lure forward}} = 0.082, s = 0.070$) than in the backward direction ($\bar{x}_{\text{unselected lure backward}} = 0.022, s = 0.035$) regardless if they selected the correct option in the initial test, $F(1, 45) = 24.725, MSE = 0.013, \eta^2_p = 0.355, p < 0.001$. While the likelihood of recalling an unselected lure was not related to the accuracy in the initial test ($\bar{x}_{\text{unselected lure correct}} = 0.053, s = 0.050$ vs. $\bar{x}_{\text{unselected lure incorrect}} = 0.051, s = 0.049$), $F(1, 45) = 0.051, MSE = 0.008, \eta^2_p = 0.001, p = 0.822$, or the presence of feedback, ($\bar{x}_{\text{unselected lure no feedback}} = 0.055, s = 0.052$ vs. $\bar{x}_{\text{unselected lure feedback}} = 0.049, s = 0.053$), $F(1, 45) = 0.265, MSE = 0.011, \eta^2_p = 0.006, p = 0.609$.

### 3.3 Discussion

Experiment 5 investigated whether providing answer feedback would eliminate the asymmetry between the forward and the backward conditions for the initially incorrect items. If the asymmetry is a result of recognizing the initial incorrect responses and deducing the correct answers in the backward direction, a possibility that is absent in the forward direction, then
answer feedback should provide enough information to eliminate the differences. To allow the investigation of this premise, in Experiment 5, the students were re-presented the word pairs as feedbacks for half of the trials in the multiple-choice test phase after they had made their responses.

For the no feedback condition, the results of Experiment 5 replicated the basic findings of the previous experiments. The accuracy in the initial test was an important determinant of the retention performance. The retention was dramatically better for the items which the students successfully retrieved the correct answers during the initial multiple-choice test phase. The retention benefit was found to be symmetric between the forward and the backward cueing directions for the initially correct items, while there was a strong bias towards the backward direction for the initially incorrect items. Additionally, the students also recalled more erroneous responses and unselected lures from the initial test in the forward condition compared to the backward condition. The intrusion of unselected lures was not related to the accuracy in the initial test.

Answer feedback was found to be effective in improving retention performance. The students’ retention performance was significantly better for the items that they received feedback towards. It appears the students were able to correct and reverse many of the errors they committed previously if they were given feedback. This error correction function of answer feedback has been widely documented in the literature (e.g., Butler et al., 2007; Butler & Roediger, 2008; Guthrie, 1971; Pashler et al., 2005). In contrary, the current study did not find students more likely to maintain correct responses in the feedback condition as they did in Butler et al. (2008) although there was a trend in this direction.
For the items the students responded incorrectly in the multiple-choice test, providing answer feedback promoted the retention in both cueing directions but to different extents, thus it altered the level of retention asymmetry between the two cueing directions. The experiment found the students showed better retention performance for both the forward and the backward cueing directions if they were given feedback after they made an incorrect response in the initial test. Moreover, the presence of answer feedback after an incorrect response improved the retention performance and provided a greater retention benefit for the items in the forward condition than for those in the backward condition. It brought the retention performance of the forward direction on par with the backward direction, eliminating the retention asymmetry. These findings suggest that re-presenting the word pairs provided enough information to eliminate the retention differences between the forward and backward cueing directions for the initially incorrect items.

However, answer feedback did not erase all traces of the asymmetry. Although the students were less likely to recall the incorrect responses they selected in the initial multiple-choice test in general, they were still more likely to recall the incorrect responses in the forward direction than in the backward direction. On the other hand, answer feedback did not suppress the occurrence of intrusions of unselected lures. The students were more likely to recall unselected lures in the forward direction than in the backward direction.
Chapter 5
General Discussion

This project employed both purely laboratory-based studies and hybrid studies combining classroom experiences with controlled assessments in the laboratory to investigate mnemonic effects of multiple-choice testing. The main goal of this project was to gain a better understanding of the processes underlying the testing effect by investigating the symmetricity of the retention benefit of testing across a cue and its associated target, and the related issues. At the same time, I also gathered data to advance the knowledge and understanding of the testing effect more generally.

1 Symmetricity of testing effect

It has been established that the facilitation of retention attributed to testing is not specific to the retrieved information in the initial test; it extends to non-retrieved information that was presented in the test as well (Carpenter et al., 2006; McDaniel et al., 2007). The main question that inspired this study was how robust is the test-induced retention facilitation for the information that was not retrieved but was presented in the initial test (i.e., the cues) relative to the information that was retrieved in the initial test (i.e., the targets). The experiments in this thesis demonstrate that the pattern is highly dependent on whether the initial items were answered correctly.

1.1 Initially correct items

If the items were answered correctly in the initial test, the magnitude of the testing effect is equivalent for the retrieved and the non-retrieved information in the initial test. Experiment 1 showed that taking a test strengthens the memory of the cues to a similar extent as the targets when prompted with the counterparts. Similarly, Experiment 2A and 2B found testing improves the retention of the information embedded within the cues to the same degree as it improves the
retention of the targets with educationally relevant materials and with longer delay intervals. These results suggest that the test-induced strengthening is symmetric across the cues and their associated targets.

These findings of bi-directional symmetry seem inconsistent with one of the major theoretical accounts of memory enhancement effects: the transfer-appropriate processing hypothesis. The transfer appropriate processing theory assumes that encoding is not a uni-dimensional process. For instance, Craik and Tulving (1975) argued that encoding can be done at different levels of depth, while Hunt and McDaniel (1993) differentiated item-specific processing and relational processing in encoding. Transfer-appropriate processing suggests that memory performance is primarily dependent on the overlap between encoding and retrieval processes. In other words, performance is assumed to reflect the degree to which the processing required to perform well in a memory task matches the processing that occurred when encoding during the study phase (Morris, Bransford, & Franks, 1977). Therefore, the same study strategy at encoding process might enhance performance on one type of test, yet may have no effect, or even an opposite effect, on a different type of test that required different type of processing (Blaxton, 1989; Fisher & Craik, 1977).

The transfer-appropriate processing concept has been adapted to explain the testing effect (Kang, McDermott, & Roediger, 2007). Recall that the basic testing effect is the finding that performance on some final test is often improved more by performing an earlier test than it is by simply studying. The similarity in terms of the cognitive processing that occurs on the initial and final test might be the reason for the enhanced performance for the previously tested item. That is, when people study they might not engage the same cognitive processes that they engage while being tested, reducing the transfer relative to a test-test context.
Similarly, Duchastel and Nungester (1982) argued that the testing effect might in fact stem from a combination of the consolidation of learning and practice with the test itself. Thus, taking an initial test does not only provide an opportunity to review the materials, it also provides an opportunity to practice on that particular test format. Therefore, taking a retrieval test in general can boost memory retention, but the practice effect especially enhances retention performance if the performance is assessed with the same test format.

With respect to the current study, clearly, there is a higher level of dissimilarity between the initial test and final test if students are asked to retrieve the cues from the initial test than if they are asked to retrieve the targets. Given the transfer-appropriate hypothesis suggests that the similarity between the initial test and the final test determines the level of facilitation in the retention, it would predict that the retention of the targets would be better than the retention of the cues. However, the results of the current study did not align with this prediction.

On the other hand, the results of this study were more compatible with the retrieval hypothesis, the competing explanation of the testing effect. The retrieval hypothesis suggests that the testing effect is a result of retrieval rather than encoding. It suggests that the act of retrieval increases the elaboration of the memory traces and the number of retrieval routes, which increases the likelihood that the information can be retrieved again in the future (e.g., Bjork, 1975, 1988; McDaniel, Kowitz, & Dunay, 1989; McDaniel & Masson, 1985).

The retrieval hypothesis suggests the link or association (direct or indirect) between a cue and the paired target are strengthened through the process of retrieval during testing. It does not imply any directionality in term of the strengthening. Therefore, the retention of the cues and the targets would be expected to be strengthened to a similar extent in general. This prediction is supported by the results of the current study.
What might be going on during the retrieval process itself that accounts for the retention benefits? The elaborative retrieval hypothesis suggests that testing may activate elaborative memory traces between the cue and the target that facilitate later retrieval of the information (Carpenter, 2009). According to this view, retrieval leads to the formation of elaborative memory traces because searching through memory for the correct answer activates other related information, and connections are formed between the cue, the target answer, and other related information in memory. This search is not necessary during restudying because the target information is explicitly presented. The elaborative retrieval hypothesis also suggests that retrieval that induce more elaborative search would provide a greater benefit to retention.

Pyc and Rawson (2010) proposed the mediator effectiveness hypothesis which further specifies the nature of these elaborative information. The mediator effectiveness hypothesis argues that testing enhance retention more than restudying because testing are more likely than restudying to enhance the link between a cue and target via mediating information (i.e., a word or concept that links a cue to a target). Carpenter (2011) and her semantic mediator hypothesis suggests that this mediating information can be a semantic in nature.

1.2 Initially incorrect items
For the initially incorrect items, the retention performance is not symmetric across the cues and the targets; the retention of the cues was, in fact, better than the retention of the targets.

Experiment 1 showed that if students answered an item incorrectly in the initial test, they had a better chance to recall the cue from the initial test when prompted with the target in a future cued-recall task than if they were asked to recall the target when given the cue. Again, Experiment 2A and 2B replicated these results with educationally relevant materials and with a longer delay.
These results are in line with the notion of the retrieval hypothesis, which suggests the link between a cue and the retrieved information is strengthened through the process of retrieval during testing. When students respond incorrectly in a multiple-choice item, an association is formed between the cue and the incorrect lure, in addition to the association formed between the cue and the target during the initial study. So, when students are prompted forward and presented with the cue during cued-recall, competition can occur between retrieval of the target and retrieval of the incorrect lure. Students show a tendency persisting to the previous erroneous responses and perform in a lower rate of accuracy. Yet, when students are prompted in the backward direction with the target, there would be no competition because the correct target has only ever been associated with the cue. Such difference in item structure might be the basis of the retention asymmetry between the two cueing directions for the initially incorrect trials.

So how do students recognize the error they made and how do they recover the correct cue? When students attempt to recover the correct answers after realizing a previous mistake, according to Experiment 3, students rely on the memory of information from the initial multiple-choice test rather than tracing back to the information from the study phase. The mechanism of recovering the correct answer based on the limited information from the initial test after recognizing an error seems to be similar to how students use verification feedbacks to correct errors committed in an initial multiple-choice test (Marsh et al., 2012). However, Experiment 4A and 4B showed that the process might, in fact, be more complex. When the accuracy of the initial multiple-choice test was made available to both cueing conditions by providing verification feedback, the performances were not modified and the asymmetry did not eliminate or reduce at all.
Finally, Experiment 5 showed that it is possible to eliminate the asymmetry induced by prior multiple-choice testing by providing answer feedback. Answer feedback provides enough information to eliminate the differences between the two cueing directions. It improved the retention for both cueing directions, but provided a greater retention benefit when students were prompted with the cues and required to respond with the targets than the opposite.

2 Transfer of knowledge

The current study also provided additional data to support the idea that information retained through the process of testing is transferable. Whether the knowledge retained through testing can be used flexibly and applied to scenarios that are different from the initial test is an important theoretical and practical question. The relevance of the testing effect would be limited if the effect is restricted to learning a specific response to a specific question in a specific context. The current study showed that the transferability of the retained information is not restricted by any of these constraints.

For example, this study demonstrated that the mnemonic facilitation of testing is not specific to the response or information retrieved in the initial test. Experiment 1 found that the test-induced retention benefit extends to the information that was presented but not retrieved in the initial test, (i.e. the cues). Experiment 2A and 2B further demonstrated that the benefit applies to the information that was embedded in the cues too. These findings converge with Carpenter et al. (2006) and McDaniel et al. (2007) who have also reported similar results. Furthermore, other studies have demonstrated that the retention of information simply related to the retrieved information can be facilitated as well, even if this information was only presented in the study phase but not in the initial test (Chan, 2009, 2010; Chan et al., 2006; Foos & Fisher, 1988).
These findings suggest that the locus of memory strengthening induced by testing is not localized or restricted. Rather, the entire memory network that is built upon the retrieved information seems to be facilitated as well which leads to the strengthening of the memory of the related information. This view is also consistent with the retrieval hypothesis in that it suggests that testing promotes subsequent retention by increasing the elaboration of a memory trace and the number of retrieval routes (e.g., Bjork, 1975, 1988; McDaniel et al., 1989; McDaniel & Masson, 1985). This process promotes the inter-connection of information that is directly and indirectly related to the retrieved information, which increases the likelihood of successful future retrieval of these related materials.

In addition, the information retained through the process of testing is not restricted to that being used to respond to a specific question. Experiment 2A and 2B demonstrated that testing enhanced the students’ memory of the information embedded in the cues in the initial test, and students can use this information to answer questions that are different from the ones in the initial test. The current study extended the findings of Rohrer et al. (2010) in which the researchers demonstrated that the information retrieved in the initial test can be used to respond to different transfer questions. In addition, other studies have also shown that memory of information related to the retrieved information is strengthened and this information can be applied to novel transfer questions (Butler, 2010; McKenzie, 1972).

These findings suggest that any information that is retained through the process of testing even if it was not retrieved in the initial test is transferable and can be used flexibly to answer questions that are different from the ones in the initial test. According to Barnett and Ceci (2002), memory demands of a transfer task involve three components: recognition, recall, and execution. Students have to be able to recognize the appropriate situation to apply previously learned knowledge,
they have to be able to recall the knowledge, and finally, they have to be able to apply the knowledge and execute the transfer. Since testing promotes the retention of the retrieved information along with information related to it, taking a test prepares students better for the recall demands of the transfer process. Therefore, it puts students in a better position for completing a transfer task.

Finally, the use of retained information is not restricted to a specific context. Aligning with numerous studies reported in the field in which they found that taking a test in one format can promote the retention performance even if it is assessed by a test in a different format (e.g., Duchastel & Nungester, 1982; Hogan & Kintsch, 1971; Kang et al., 2007; Lockhart, 1975; Thompson et al., 1978), the laboratory experiments of this study showed that taking a multiple-choice test can promote the retention performance on a cued-recall task. In a recent study, Butler (2010) also found that repeated testing produced superior transfer on the final test compared to repeated studying, even the final test was composed of new inferential questions from a different knowledge domain.

All these studies converge and suggest that taking a test provides enormous mnemonic effects. Testing enhanced the retention of the whole memory network that is built upon the retrieved information. The information retained can be used flexibly to answer different questions and in different context. These properties make testing a very powerful tool to promote memory retention, at least when the negative effects of testing are contained.

3 Temporal dimension of testing effect

The findings of the current study have also illustrated a more complex temporal picture of the effect of testing than what has been commonly believed. It was believed that testing improves memory retention by simply slowing down the rate of forgetting. This view has been mainly
supported by two lines of evidence. Spitzer (1939) illustrated a typical forgetting curve in which retention drops at a rate that is an inverse function of time delay. He found that taking a test completely changed the trajectory of the rate of forgetting, and that the slope of the forgetting curve drastically decreases at the point when students take a test. The second line of evidence came from studies that examined the crossover interaction between testing and re-studying (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b; Thompson et al., 1978; Wheeler et al., 2003). These studies found that although re-studying provide a superior retention enhancement over testing in the short-term, the level of retention of the re-studied materials drops at a much faster rate than the tested materials. Over time, the benefit of re-studying over testing gradually diminishes, and eventually, the pattern reverses favoring testing over re-studying. Both of these results suggest that taking a test reduces the rate of forgetting.

Given the different rates of forgetting for the tested materials and non-tested materials, the assumption is that the retention difference between the tested and the non-tested items or the test-induced retention benefit would increase gradually over time. Experiment 2A, however, showed that the benefit of testing does not actually increase infinitely. Instead, the retention benefit of testing remained a constant over a period of three days to six weeks. However, because the forgetting curves have the shape of inverse functions, the forgetting rate of the tested items appears to be slower than the forgetting rate of the non-test items over a short delay (about ten days) after the initial test, and subsequently, the forgetting rate of the tested and non-tested materials seemed to stabilize and the retention of these two types of items dropped at roughly the same rate up to six weeks.

Rather than the conventional view that testing slows down the forgetting rate, testing seems to provide a protection against the rapid plummet of retention that occurs soon after study ends, and
while the protection is not long-lasting, the advantage gained from this protection is maintained over time. It seems that, after taking a test, those soon-to-be-forgotten materials are no longer forgotten rapidly. Testing might have strengthened the memory of these otherwise soon-to-be-forgotten materials, and it might have pushed the memory of these materials over the threshold of irretrievability under the given retention test format or demand. Testing might have strengthened the retention of all the materials, but perhaps, the retention enhancement of the materials that are vulnerable to forgetting might indeed be the most critical contribution of testing.

These results are consistent with the bifurcation model (Kornell, Bjork, & Garcia, 2011). The bifurcation model builds on the assumption that memory tests do not measure memory strength of an item; instead, they measure whether the memory strength of an item is above a particular retrieval threshold. Items with memory strengths above the threshold are retrieved in a test, and the act of retrieval further enhances the memory strengths of these items. Whereas, items with memory strengths below the recall threshold are not retrieved in the test, and the memory strengths of these items are not boosted at all. The differential effects of testing towards these two classes of items results in a gap between them and bifurcates the distribution of item strengths. Over time, items gradually lose memory strengths. But even if the tested items are losing memory strengths at the same rate as the non-tested items or the restudied items, because of the gap in the bifurcated distribution, the tested items would appear to be forgotten (fall below the retrieval threshold) in a slower rate in the short delay. This prediction is supported by the data of the current study.

Another issue related to the temporal aspect of the benefit of testing is the durability of the effect. Studies have shown that the effect of testing can last well over a month (Duchastel & Nungester,
1981; McDaniel et al., 2011; Roediger et al., 2011; Spitzer, 1939). One possibility is that the retention benefit would maintain permanently. The forgetting rates of the tested and the non-tested materials would slow down gradually and simultaneously. The retention of these materials might reach two different asymptotes at which the remaining materials have high resistance against forgetting. As a result, testing leads to a greater proportion of the materials that have a high resistance against forgetting, and the retention of the test materials would always be superior to the retention of the non-tested materials. However, the results of Experiment 2B suggest that it is not the case. Experiment 2B found that the overall benefit of testing diminished after four to five and a half months, and the retention of the tested materials was no longer significantly different from the retention of the non-tested materials. Other studies have also documented similar attenuation (Duchastel & Nungester, 1981; Fazio, Agarwal, et al., 2010; Spitzer, 1939). It appears that the benefit of testing would maintain at a fairly constant level until the retention level of the non-tested materials approaches the asymptotic level. At that point, the forgetting rate of non-tested materials would slow down substantially. While the retention of the tested materials continues to drop, the level of retention benefit would gradually diminish, and the benefit would attenuate eventually.

Yet, the overall accuracy data does not provide an accurate picture of the whole story given the two opposite effects of testing. When one looks only at the items responded to correctly during the test, the testing effect is clearly still very robust after the long delay, however so too is the negative effect of testing that is associated with initially incorrect responses. The two effects are simply more balanced after the long delay, leading to an overall null effect.
4 Final Summary

To summarize, the study replicated the testing effect phenomenon. It shows that the accuracy in the initial test along with the cueing direction in the retention task was an important determinant of the retention performance. The symmetry observed for items responded accurately suggests that the basis of the testing effect might be a simple strengthening of connections between items as predicted by the retrieval hypothesis, not a transfer of processing as suggested by the transfer appropriate processing theory. The asymmetry observed for items responded inaccurately along with the tendency of students persisting to previously selected lures is also consistent with the retrieval hypothesis.

This study demonstrated that the testing effect is not specific to the information retrieved in the initial test but extends to the information that was presented but not retrieved in the initial test. Also, the knowledge retained through testing can be applied flexibly beyond specific questions and specific contexts. Finally, the previous notion that testing slows forgetting seems to be incorrect. Instead testing seems to provide a short-term insulation against immediate forgetting, but then the memory of the tested materials decay in the same way as the non-tested materials.

5 Implications for education

This thesis along with other studies in the testing effect literature has demonstrated that testing does not only assess students’ learning, it also enhances students’ learning. The effect is robust and long-lasting. Additionally, the information retained through testing can be used flexibly to answer different questions and in different contexts. These properties make testing a very powerful teaching and learning aid. The use of testing as a formative learning device should be encouraged.
This thesis also showed that students form powerful associations between the questions and the answers they choose while writing a test regardless of the objective correctness of their answers. Correct choices lead to a positive testing effect while incorrect choices lead to a negative testing effect. Both of these effects persist for months after testing. In order to maximize the efficiency of testing, it is critical to provide students with the correct answers to reverse the negative testing effect and unlearn the incorrect associations formed during the test. This is an important practice regardless of whether a test is meant to be summative or formative.

When traditional methods are used (i.e., pen and paper multiple-choice test), it is not trivial to find ways that students can be given immediate feedback (but see DiBattista, Gosse, Sinnige-Egger, Candale, & Sargeson, 2009). However, as technology plays a larger role in our assessment of students, it may allow feedback in ways traditional assessments do not. In fact, one of the tangible results of the work described in this thesis was the development of mTuner, an online assessment tool that utilizes feedback and formative assessment to maximize the positive effects of testing while preventing the negative effects (Cheng, Joordens, & Walker, 2013).
References


Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. 

*Contemporary Educational Psychology, 6*(3), 217-226. doi: 10.1016/0361-476x(81)90002-3


*Psychological Reports, 49*(2), 470.


