Examining the Testing Effect in an Introductory Psychology Course

by

Christopher Ray Howard

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 9, 2010

Key words: testing effect, repeated testing, learning, retrieval

Approved by

William Buskist, Chair, Distinguished Professor in the Teaching of Psychology
Lewis Barker, Professor of Psychology
Aimee Callender, Assistant Professor of Psychology
Chris Correia, Associate Professor of Psychology
Jeffrey Katz, Alumni Associate Professor of Psychology

Abstract

This study examined the effects of repeated testing in an Introductory Psychology course. Student performance on items repeated from quizzes to exams (either unit, cumulative, or both) was compared to items that had not been previously administered. In addition, we examined unit and cumulative exam performance by the format of quiz items (multiple-choice, short-answer, or summary study items) to examine differences resulting from original format. We found that prior exposure to, not prior testing of, items was beneficial for enhanced performance on later assessments. Although students performed better on multiple-choice quiz questions, there were no differences in performance for repeated multiple-choice, repeated short-answer, or study items on either unit or cumulative exams. We performed a series of regression analyses on four individual difference variables: aptitude, academic achievement, learning strategies, and study skills. Academic achievement was the single best predictor for the benefit of testing.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

MC	Multiple-choice
SA	Short-answer
UE	Unit Exam
CE	Cumulative Exam

Introduction

> "Although cognitive and educational psychologists have studied testing off and on over the years, we believe the time is ripe for a dedicated and thorough examination of issues surrounding testing and its application in the classroom." (Roediger & Karpicke, 2006a, p. 206).

Recently there has been considerable scientific interest in the benefits and consequences of educational testing in higher education. Although educational tests have been used primarily as assessment devices, tests may actually serve as a learning opportunity by providing additional exposure to critical course content (Toppino & Luipersbeck, 1993). In some instances, such additional exposure through testing can result in the enhanced retention of previously learned information. This enhancement, termed the *testing effect* or *test-enhanced learning*, is evidenced by increased objective performance on a final assessment at a later point in time (for a review, see Roediger & Karpicke, 2006a).

The enhancement in performance as a result of testing has largely been explained as a benefit of retrieving information from memory (e.g., Roediger & Karpicke, 2006a; Tulving, 1967). Roediger and Karpicke (2006b) conducted two studies examining the effects of testing. In both experiments, students read narrative passages and either took a test immediately after or after some delay. In the first study, participants completed two study trials for half of the passages and completed one study trial and one test trial for the remainder. During the study trial, participants read or reread the passages. During test trials, participants recalled as much

information from the passage as possible. The researchers also manipulated the delay between the initial study or test condition and a final, criterion test, ranging from 5 minutes to 1 week. Participants who completed two study trials performed better on a final test that occurred after a 5-minute delay. Interestingly, participants who completed test trials performed better on the final test after either 2 days or 1 week.

In Roediger and Karpicke's second study, participants either studied the passage four times, studied the passage three times and took one recall test, or studied the passage once and took three recall tests. A final test was completed after either 5 minutes or 1 week. The results for the second study mirrored those found in Study 1: Participants who completed four study trials performed significantly better than the other two experimental conditions. In addition, individuals who completed three study trials and one test trial performed better than those who completed only one study trial and three test trials. However, this pattern was reversed after 1 week. Participants who completed one study trial and three test trials recalled significantly more information than the other two conditions. Taken together, these two studies confirm that retrieval (through testing) can facilitate greater recall after some delay.

Other researchers have suggested that the benefit of testing may be explained by a reduction in forgetting that results from engaging in such tasks (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Wheeler, Ewers, & Buonanno, 2003). Carpenter et al. (2008) conducted several studies examining the rates of forgetting following periods of initial testing or additional study opportunities. Participants in the first two experiments learned trivia facts, while participants in the third experiment learned English-Swahili word pairs. After learning the material, participants either restudied the material or were asked to recall the correct answer during an initial test trial. After variable delays, participants took a cued-recall test over the information. Participants who

took initial tests over the information performed significantly better than participants who restudied the material, in all three experiments. The participants who benefitted the most took three initial tests over the information, as opposed to the participants given a single test or three study periods. Interestingly, Carpenter et al. found that forgetting was significantly reduced following test trials when compared to study trials. Although this effect was small, other researchers have found more significant reduction in rates of forgetting.

Wheeler et al. (2003) conducted two experiments examining forgetting following learning of word lists. Their first experiment examined rates of forgetting after a 48-hour period. After all participants listened to a list of 40 nouns, half of the participants took three initial free-recall tests in succession, while the other half listened to the word lists three additional times. Half of the participants from each condition took a final free-recall test after 5 minutes, while the remainder took the final test after a 48-hour delay. The researchers found a significant interaction between the initial activity type (initial test vs. additional study) and length of retention: Individuals who completed additional study trials performed better than individuals who completed initial test trials after a 5-minute delay, but there was no significant difference in performance after a 48-hour delay. Thus, after 48 hours, participants in the additional study trials forgot more information than individuals who took initial tests.

Wheeler et al.'s (2003) second study examined rates of forgetting following a 7-day delay using a similar procedure to the first study. The researchers, again, found a significant interaction between the initial activity type (test vs. additional study) and delay intervals (5 minutes and 7 days). After a 5-minute interval, participants who engaged in additional study periods performed significantly better than individuals who took initial tests. This pattern was reversed, however, after a 7-day delay. Participants who completed additional study trials performed significantly

worse than individuals who took initial tests. Thus, the researchers found a steeper rate of forgetting for those who took additional study trials. Taken together, these results suggest that, although studying has an immediate benefit, testing may actually reduce the amount of forgetting over longer intervals of time.

Crooks' (1988) also concluded that the testing effect may result because intervening tests allow for self-evaluation of content mastery. This self-evaluation, self-reflection, or awareness of one's own capacity is known as metacognition. Individual learners benefit two-fold from metacognition. First, such awareness can increase studying or learning efficiency by limiting or restricting the amount of time a student studies a given set of material. Once students feel that they have learned the information sufficiently, they can move on to different or more advanced concepts. Second, such awareness can allow students to focus time and attention on concepts that have not been mastered (Glenberg, Sanocki, Epstein, & Morris, 1987). In short, metacognition allows an individual learner to assess both what is known and what remains to the learned. To date, there have been very few studies that have examined the role of metacognition on the testing effect, or studies that have examined the indirect, metacognitive benefits of repeated testing.

Spunzar, McDermott, and Roediger (2007) examined whether expectation of a final test influenced the magnitude of the testing effect. In their experiment, participants studied five different lists of words and took an initial free-recall test for each after a short delay. Half of the participants were told that a final, free-recall test over all the lists would occur at the end of the experimental session, while the remaining half was not. Individuals who were told about the upcoming final test performed significantly better on the final test than individuals who were not. The researchers concluded that individuals who were not expecting a final test did not need to

maintain the lists in memory once the initial test was complete, but individuals expecting a final test may have attempted to maintain the information in memory across trials through a variety of strategies (e.g., rehearsal).

Spunzar et al. (2007) suggested that expectation of a final test may have a pivotal role in classroom learning. If instructors have cumulative assessments across the semester, students may seek to establish links between chapters or units in an effort to retain critical information. Although students have the opportunity to restudy information over the course of the semester, Spunzar et al. found that expectation, in the absence of study opportunities, may be sufficient to enhance the testing effect. Therefore, one would expect that additional study coupled with expectation of a final (or cumulative) test may lead to an even greater magnitude when compared to expectation alone. Although the present study does not directly compare or manipulate the expectation of a final test, all of the assessments included in this study were cumulative.

Theoretical Underpinnings of the Present Research

This study contributes to a growing body of literature that attempts to bridge experimental and educational psychology. Richland, Linn, and Bjork (2007) suggested that greater research collaboration between these fields could be advantageous for both fields and for the understanding of numerous psychological phenomena, including the testing effect. Richland et al. suggested that laboratory researchers could include more educationally relevant dimensions into studies, while applied researchers could examine the phenomena under different, more realistic, levels of motivation and attention. The present study addresses these issues by examining the testing effect in an actual classroom setting while attempting to employ appropriate levels of experimental rigor and control. There are some variables, however, that cannot be controlled in such an applied context (e.g., amount of studying, etc.). Nevertheless,

testing is an important part of pedagogical practice and further research may serve to help both teachers and researchers to understand and bolster its application.

Historical Perspectives on Testing: Twentieth Century Research

Interest in the effects of testing has not been limited to experimental or cognitive psychology. In fact, some of the earliest work on the effects of frequent testing was conducted to understand the factors that influence the retention of classroom information. One of the first classroom investigations of frequent testing (Jones, 1923) found that individuals who took an intervening test over the material performed better on the regularly scheduled exam than individuals who did not receive an intervening test. Other early studies, such as Keys (1934) and Spitzer (1939), also provided empirical support for the benefit of frequent testing.

Keys (1934) conducted an in-class study to determine the consequences of more frequent (weekly) testing compared to less frequent testing (monthly). Students from a large educational psychology course were divided into two smaller instructional sections after the first week of class. The first instructional section, which served as the control condition, took monthly examinations covering lecture material and two required textbooks. The experimental section took weekly tests over the assigned material and lectures. The experimental condition completed weekly quizzes for the first 2 months and only a monthly exam for the last unit. Both groups responded to identical exam items.

Keys (1934) found that participants who completed weekly exams performed significantly better than participants who took monthly exams. Although both the experimental and control conditions took a monthly test for Unit 3, the experimental group performed significantly higher than the control group. Interestingly, on the final exam for the course (which occurred several weeks after Unit 3), there was no significant difference between participants in the weekly and monthly testing conditions. Surveys conducted at both the beginning and the end

of class showed that the majority of students believed that more frequent testing would have "more real and lasting benefit" (Keys, 1934, p. 434). However, this study failed to find an enhancement in long term retention.

Spitzer (1939) investigated the optimal time for initial testing by carefully manipulating testing in a large population of elementary students. His sample of 3,605 Iowan 6th graders was divided into 10 different groups, with each group associated with some delay between the learning phase and the initial testing phase. Children in all groups read a text passage that corresponded to a classroom lecture topic and took an initial test either immediately or after some delay. The delay was variable across groups and ranged from 1 to 63 days after study. All groups took a final test on day 63 to measure retention. Children who were tested relatively early after reading the passage retained the most information at day 63. Children who took the initial test immediately after reading the passage retained more information than all other groups. In addition, children who were initially tested at later delays showed a diminished retention of information at day 63 compared to those who had taken the initial test earlier. Therefore, testing was most beneficial if it occurred closer to the presentation of the material and an initial test with a shortened delay was associated with better performance on a final test. Sones and Stroud's (1940) study largely confirmed Spitzer's findings, by concluding that testing was most beneficial if it occurred within 48-72 hours after the initial study period, while repeated studying was beneficial after a 2-week delay. They also reported no significant difference between testing and restudying conditions for days ranging from approximately 8 to 15 days.

Dustin (1971), while teaching a large developmental psychology course, created two equivalent sections during the second week of the semester, controlling for gender and performance on the first unit exam. For the control group, assessments and exams were given

monthly. The experimental group, however, completed weekly assessments and tests. The questions appearing on the monthly and weekly versions were the same. Thus, any significant effects found between the groups were not attributable to different items or non-equivalent groups. A final, unannounced, retention test was given after the last exam in the course.

In contrast to Keys (1934), Dustin (1971) found that individuals who were tested weekly performed significantly better on exams than did individuals who were tested monthly. This finding held for both the exams throughout the semester, as well as the final retention test. Dustin also found a significant difference in test anxiety between the two groups: Individuals who were tested monthly reported higher levels of test anxiety than those individuals who were tested weekly.

Nungester and Duchastel (1982) investigated the effects of testing using senior-level high school students. Students assigned to the first experimental condition took an initial test containing both short-answer and multiple-choice items immediately after studying the passage. The second experimental group studied the passage for an extended period instead of completing an initial test to equalize total exposure time. The control participants did not take an initial test or engage in any extended study. All participants completed a final test 2 weeks after studying the passages. The final test contained short-answer and multiple-choice items where half of the items had been used in the initial test for the first experimental condition. The items that appeared on both the initial and final test were presented in alternating formats (multiple-choice to short-answer or short-answer to multiple-choice), but were otherwise identical. Individuals who took an initial test or an additional period of study, as well as participants who did not engage in any type of post-learning activity, did not receive any of the test items previously.

Nungester and Duchastel (1982) found a significant difference between the initial test and no-activity condition. Individuals who took an initial test performed significantly better on the final test when compared with individuals who only studied the passage. There was, however, no significant difference between individuals who took an initial test and those individuals who engaged in extended review on the final test. The authors found that there were differential effects of testing if the item had been previously administered. Participants in the first experimental condition performed significantly better than participants in any other condition if the items appeared on the initial test. If the items did not appear on the initial test, both prior testing and extended review were beneficial. Nungester and Duchastel concluded that this discrepancy would not have occurred if all the items had been included in the initial test. The authors suggested that any difference between individuals in the first experimental condition and those in the second experimental condition must not be attributed to the total exposure time, because exposure was controlled. Performance on both item types—short-answer and multiple-choice—benefited from prior testing even when the format was reversed for previously administered items.

In the studies described above, the researcher randomly assigned or matched participants to either experimental or control conditions. Grover, Becker, and Davis (1989) examined the effects of testing conditions when participants (students) self-selected a particular testing strategy. Grover et al. devised two different assessment options for an introductory psychology course. Students who selected the first option completed unit exams, which covered lectures and readings from four selected chapters, for a total of one exam per unit. Students who selected the second option completed chapter exams, which covered lectures and readings for only one chapter at a time, for a total of four exams per unit. Half of the students in the sample selected

the unit exam option while the remainder selected the exam per chapter option. All questions were identical for both unit and chapter exams. The final exam for the course was comprised entirely of previously administered items; however, the students were not given this information prior to the final exam.

In contrast to Keys (1934), Grover et al. (1989) found no significant difference between the exam scores of individuals who selected the chapter or unit exam option. This finding held for both individual items from each chapter and total performance on the final exam. In an end-of-semester survey, half of the individuals who selected the unit exam option reported that they would switch to the chapter exam option if given the option again. However, none of the individuals who selected the chapter exam option reported a desire to change to the unit exam option.

Conclusions from Historical Perspectives

Early research on the benefits of testing yielded conflicting results. Although several researchers found that initial testing leads to enhanced retention and increased objective performance on the final criterion test (e.g., Dustin, 1971; Jones, 1923; Spitzer, 1939), others failed to find a benefit for testing (e.g., Nungester & Duchastel, 1982). In addition, several researchers found higher performance on the intervening tests, but showed no significant difference on the final criterion test (e.g., Keys, 1934; Grover et al., 1989). The preceding review does not attempt to encapsulate the entire history of the testing effect, but attempts to examine some of the more influential studies from this body of literature. Bangert-Drowns, Kulik, and Kulik (1991) offered a meta-analysis of much of the research done between Spitzer's (1939) classic study and current research on the testing effect. The authors reviewed 35 classroom-based studies on the testing effect and found a positive benefit of testing across 29 of the 35 studies and that only six of the studies found negative effects of testing. Students who took an initial test

performed better than a control condition (taking only the final test). Additionally, performance was better for students who had taken several short tests when compared to one long test. This meta-analysis is important for two reasons. First, the positive effects of frequent testing emerged across many studies that differed with respect to course content (social science, reading, geography, mathematics, etc.), across decades, and with varying methodologies. These findings suggest that the testing effect is reliable and possibly robust. Second, the studies represent historical advancements in testing effect literature. Glover (1989) commented that the testing effect was "not gone, but nearly forgotten" and lamented that the last study of the effects of testing in educational circles was Spitzer (1939). However, Bangert-Drowns et al. offered compelling evidence that classroom-based studies on the effects of testing were being conducted and that the results were largely consistent with earlier researchers (e.g., Jones, 1923; Spitzer, 1939).

Modern Perspectives on Testing: Twenty-first Century Research

McDaniel, Anderson, Derbish, and Morrisette (2007) conducted the most recent classroom investigation of the testing effect using a within-subjects design. In this study, participants completed a 6-week, online "Brain and Behavior" course. Each week, students completed a quiz over assigned readings and received feedback (including the correct answer) concerning their performance. These quizzes contained a combination of multiple-choice, short-answer, and study items (summary statements from the week's readings that served as a baseline for comparison). Students also took two unit exams that covered 3 weeks of assigned readings and a cumulative final exam that covered information from the entire semester. All of the questions on the unit and final exams were in multiple-choice format where half of the questions were repeated from previous assessments.

For unit exams, McDaniel et al. (2007) found that student performance (as measured by proportion correct) was enhanced by both multiple-choice and short-answer quiz items when compared to study items. Although the proportion of correct answers was significantly higher for multiple-choice items than short-answer items on quizzes, this effect reversed for unit exam performance. Items that were initially tested in a short-answer format showed a significant advantage over those tested in multiple-choice format. The advantage for short-answer quiz items held for both unit exams, but largely disappeared on the final exam. Although there was no significant difference between multiple-choice and short-answer quiz items on the final exam, both were significantly better than study items.

Butler and Roediger (2007) conducted a study examining the testing effect using a simulated classroom procedure. In their experiment, a group of students watched three video lectures on art history while taking notes. Participants knew that lecture material would appear on subsequent short-answer or multiple-choice tests. After the completion of the three videos, participants either studied the lecture material (through a summary handout given to them) or took an initial test on information in the video. Participants received feedback concerning their initial test performance. The test questions and the lecture summary contained the same information (in differing formats) to eliminate any discrepancy in content. Participants returned for a final short-answer test 1 month after completing the three lectures and post-lecture activities.

Butler and Roediger (2007) found that the provision of feedback during initial testing did not have a significant impact on the final test, but that the type of question asked during the initial test influenced later scores: Participants who took an initial short-answer test did significantly better on the final test than participants who took an initial multiple-choice test.

Interestingly, the multiple-choice condition did not differ significantly from the study condition. However, all three post activity conditions (short-answer test, multiple-choice test, and study) enhanced retention over the control condition, in which the subjects did not engage in any post-lecture activity. Thus, both testing and repeated study appeared to enhance retention, with added benefit for testing, especially short-answer testing.

Kang, McDermott, and Roediger (2007) conducted two within-subject studies examining the impact of initial test type and the provision of feedback. In their first study, participants read four passages and took an initial multiple-choice test, short-answer test, or read target (to-be-tested) facts. After a 3-day delay, participants completed a multiple-choice or short-answer test for each of the readings. Kang et al. found that an initial multiple-choice test was significantly better than an initial short-answer test for both final test formats. However, an initial multiple-choice test was not significantly different from reading target facts for either final exam type. In their second study, Kang et al. included full feedback into the experimental design. After answering each question on the initial exam, participants received full feedback identifying or disclosing the correct answer to the question. Initial short-answer testing resulted in significantly better performance on a final multiple-choice test when feedback was given, but there was no significant difference between initial test types when the final test was short-answer. There was no differential advantage for taking an initial multiple-choice test or reading target facts.

Free-recall or short-answer tests have frequently been touted as superior methods for enhancing later test performance (e.g., Bjork, 1975). Although Butler and Roediger (2007) found this enhancement in the absence of initial performance feedback, Kang et al. (2007) concluded that feedback was essential for the relative advantage of short-answer tests.

Butler and Roediger (2008) furthered earlier Kang et al.'s (2007) work by examining the role of feedback on multiple-choice tests. During initial multiple-choice testing, students are not only presented with correct, but also incorrect information (in the form of incorrect multiple-choice answers or lures) that may be recalled on subsequent tests. Butler and Roediger predicted that feedback after an initial multiple-choice test would reduce the amount of incorrect information recalled later. In their experiment, participants in the first experimental condition took an intervening test over the material, while participants in the second experimental condition read the passage and studied target facts from the passage. Participants in the control condition did not read or study the passage. After a brief delay, participants in all three conditions took an initial multiple-choice test and received either no feedback, immediate feedback, or delayed feedback. Participants took a final cued recall test 1-week later. Butler and Roediger found that participants who were given delayed feedback performed significantly better than individuals who received immediate feedback. Providing either immediate or delayed feedback reduced the recall of incorrect multiple-choice lures presented during initial testing.

Although the testing effect occurs in the absence of feedback, providing feedback during initial testing appears to have some benefit. For short-answer tests, feedback may enhance performance by exposing individuals to the correct answer. Kang et al. (2007) noted that participants performed better on initial multiple-choice tests than initial short-answer tests. For multiple-choice tests, feedback may reduce the intrusion of incorrect information that occurs because of routine exposure to incorrect information (in the form of multiple-choice lures). The testing effect is not dependent on receiving feedback about initial performance, but such information may increase the magnitude of the effect.

Agarwal, Karpicke, Kang, Roediger, & McDermott (2008) conducted a study of the testing effect using both open- and closed-book test conditions and investigated the impact of feedback for both conditions. In the open-book conditions, participants had access to the studied passage while taking the initial test. In the closed-book condition, participants did not have access to the passage during initial testing. After a 1-week delay, participants returned for a final short-answer test. In alignment with their predictions, the authors found that open-book testing led to superior performance when compared to closed-book testing. It was also determined that the testing effect occurred for both open- and closed-book tests. In their second experiment, the researchers compared testing with three study conditions in which participants restudied the passages one, two or three times, respectively. Participants who studied the passage twice performed significantly better than individuals who studied the passage only once. Participants who studied the passage three times performed significantly better than individuals who studied the passage twice. Although repeated study was beneficial, testing (open- and closed-book) resulted in significantly better performance than studying (at any level). There was no significant difference between the positive effects of testing for open- and closed-book testing on the final assessment. Agarwal et al. also concluded that testing with feedback resulted in significantly higher scores than testing without feedback.

Recently, Marsh, Agarwal, and Roediger (2009) examined the consequences of taking repeated standardized tests. Participants in their first experiment read passages taken from four domains of the Scholastic Aptitude Subject Test (formerly known as the SAT II): biology, chemistry, US history, and world history. After reading a selected passage, participants answered multiple-choice questions from an outdated SAT subject test. After a small delay (approximately 5 minutes), participants completed an 80-item short-answer test over the passages. On the final

test, half of the items were repeated from the initial quiz, while the remainder had not been previously administered. Marsh et al. found a significant effect of testing across all four domains. However, the testing effect was larger for some domains (biology) than others (chemistry). The magnitude of the testing effect also was significantly, and positively, correlated with multiple-choice performance.

In their second experiment, Marsh et al. (2009) examined the impact of free and forced responding to items appearing on the initial quiz. In forced responding, participants selected an answer for each question, whereas in the free responding condition participants could skip a question without any penalty. The researchers found no significant differences in the positive effects of testing between the forced and free responding conditions. The pattern of results across domains was consistent with their first experiment. Interestingly, the researchers compared the top 25% of the students in the experiment with the bottom 25%. Individuals who achieved higher scores on the final test had significantly larger increases in scores (from 31% on the initial test to 65% on the final test) than did individuals who had lower scores on the final test (from 16 to 27%). These findings suggest that the magnitude of the testing effect may be mediated by the participants' academic ability.

Rohrer, Taylor, and Sholar (2010) examined the effects of transfer on the magnitude of the testing effect. They argued that many of the studies conducted on the testing effect use identical or nearly identical questions during the initial testing and final testing phases. In their study, fourth graders learned locations on a series of fictional maps using either a test-study strategy or a study only strategy. After a 1-day delay, participants returned and completed a standard final test and a transfer-based final test. For the standard final test, participants were given an unlabeled map accompanied by a list of place names. For the transfer-based final test,

participants were given an unlabeled map, but were not given the list of place names. Based on the proportion correct for each test, the researchers concluded that there was a testing effect present for both the standard final test and the transfer-based final test. However, the magnitude of the testing effect was larger for the transfer-based final test.

Most recently, Sumowski, Chiaravalloti, and DeLuca (2010) conducted the first clinical application of the testing effect by comparing the effects of repeated testing between individuals who suffer from multiple sclerosis (MS) and healthy controls. Memory impairment is a common feature of MS. Participants learned verbal paired associates either using massed study, spaced study, or spaced testing. After a 45-minute delay participants completed a cued-recall test over the word lists. Sumowski et al. determined that performance was significantly higher for items that were initially presented in the spaced testing condition when compared to the other two conditions. There was no significant difference between the final performance of patients suffering from MS (with no minimal memory impairment) and healthy controls. Patients suffering from MS with significant memory impairment performed significantly lower than MS patients with minimal memory impairment and healthy controls. This study provides additional support for the utility of repeated testing in an another type of applied setting.

Conclusions from Modern Perspectives

Recent studies in experimental and applied psychology have overwhelmingly found that testing can be a powerful tool to promote retention for information (e.g., Butler & Roediger, 2007; McDaniel et al., 2007). The testing effect has been found for both multiple-choice and short-answer items. Short-answer items seem to promote better retention, in part, because they require deeper levels of processing and retrieval (Kang et al., 2007). Although the testing effect occurs in the absence of feedback about initial performance (e.g., Butler & Roediger, 2007), feedback may modulate the effects of prior testing (e.g., Kang et al., 2007; Butler & Roediger,

2008). Feedback may be more beneficial for short-answer items, because of lower performance scores on initial short-answer items when compared to initial multiple-choice items. However, feedback on initial multiple-choice tests appears to reduce the recall of incorrect information (or lures; Butler & Roediger, 2008).

Finally, testing effect research has practical significance for pedagogical practice (Roediger & Karpicke, 2006a). Many researchers have noted the significant advantage of more or frequent testing over traditional or less frequent methods of assessment (e.g., Jones, 1923; Spitzer, 1939; Grover, et al., 1989; McDaniel, et al., 2007). Seidel, Benassi, and Lewis (2008) suggested that research on the testing effect promotes test-enhanced learning as an evidence-based pedagogical strategy—a technique or practice that has been empirically verified to facilitate learning or improve retention (Svinicki & McKeachie, 2010).

The pedagogical implications of frequent testing have been discussed in several diverse literatures including marketing courses (e.g., Kling, McCorkle, Miller, & Reardon, 2005), health administration courses (e.g., Seidel et al., 2008), medical education (e.g., Larsen, Butler, & Roediger, 2008), and psychology (e.g., Grover, et al., 1989; McDaniel, Anderson, Derbish, & Morrisette, 2007). Also, as new technologies and innovative strategies for learning and testing are developed, researchers must assess their value in the classroom. For example, Mayer, et al. (2009) examined whether frequent in-class assessments using personal response systems (or clickers) had any benefit over traditional methods of quizzing. They found those students who were frequently assessed using clickers performed about 3% higher on exams across the semester when compared to students in control groups (no clickers and no quiz groups).

Overview of the Present Research

This study examined the effects of repeated testing in a large naturally occurring Introductory Psychology class. According to the regularly scheduled activities for the class, students took quizzes, tests, and cumulative tests during the semester. Quiz items were presented in various formats (multiple-choice and short-answer) and test items were either repeated from previous quizzes or were previously untested items. The present research both extended earlier research on the testing effect and examined several novel dimensions that have not been previously addressed by other studies.

The present study extended previous studies by examining the testing effect using educationally relevant stimuli. Other researchers (e.g., Butler & Roediger, 2007; McDaniel et al. 2007; Marsh et al., 2009) have used a variety of stimuli, ranging from art history videos to SAT subject test passages. Several researchers (e.g., Dustin, 1971; McDaniel et al., 2007) have used content from psychology lectures or texts as stimuli. The present study examined how testing might enhance the learning of psychology course content. This study also attempted to confirm the robustness of the testing effect by examining it in an actual classroom context. Although such applied contexts cannot control for every variable (e.g., amount of studying), a high degree of experimental control was achieved.

In order to control for as many individual factors as possible, we chose to use a within-subjects approach. There are several advantages to examining the effects of testing this design: First, there is a higher degree of control because each participant completes all of the study and test trials. The effects of many personal or individual variables are held constant across

participants. Second, most of the recent research has focused on smaller samples from simulated educational contexts(e.g., McDaniel et al.,2007). Although students in this study were enrolled in a "Brain and Behavior" course, their participation and the grades from the assessments were not used to determine a course grade (see McDaniel et al., 2007, p. 501, footnote 2). Student behavior and motivation may be significantly different from that of students in real educational contexts, where quiz and exam performance are directly related to course grades.

Although many studies have been conducted on the testing effect, only a few studies have examined individual difference variables that may influence the magnitude of the testing effect. The most recent study to examine individual difference variables was conducted by Marsh et al. (2009). Using a post hoc analysis, Marsh et al. (2009) concluded that students who performed in the top 25% on SAT subject test questions showed a greater advantage of testing than students who performed in the lower 25%. This diminished return for lower achieving students was found in both high school and undergraduate college students. Graham (1999) examined the impact of announced and unannounced quizzes in two psychology courses. He incorporated both quiz types into the structure of both courses across several semesters. After controlling for several factors, he found that students performed significantly better on course exams when they were preceded by unannounced quizzes rather than by announced quizzes. Interestingly, the significant effect was largely accounted for by higher exam averages for mid-range students (those who were earning Cs) who took unannounced quizzes.

The present research drew from a larger sample, with greater diversity (or variation), allowing for a more focused examination of factors that may affect the magnitude of the testing effect. The present research also examined individual and academic differences among participants that may impact the overall benefit of testing.

Hypotheses for the Present Research

Testing effect. We hypothesize that students will score higher on exam questions that were initially quizzed when compared to exam items that have not been previously administered. In addition, we predict that students will score higher on cumulative exam questions that were previously administered on both quizzes and a unit exam when compared to novel items on the cumulative exam. Finally, we believe that students will perform better on items that appear on both the unit exam and the cumulative exam when compared to items appearing only on the cumulative exam.

These predictions result from many studies (for a review, see Roediger & Karpicke, 2006a) which concluded that performance on a final, criterion test is enhanced by intervening test trials when compared to intervening study trials. Experiments examining the effects of multiple intervening trials have found that enhanced performance is a function of the number of intervening test trials (e.g., Agarwal et al., 2008).

Item Format. We hypothesize that students will perform better on multiple-choice quiz items than short-answer quiz items. However, on unit and cumulative exams, we predict students will perform better on items that were initially quizzed in a short-answer format when compared with those quizzed in a multiple-choice format. This prediction is based earlier findings (e.g., Roediger & Karpicke, 2006) and supports the effortful retrieval theory of the testing effect. We also predict that this relationship will be maintained even when study time has been statistically controlled.

Aptitude. We predict that students with higher aptitude scores will benefit more from repeated testing than lower scoring students. This hypothesis is based on the findings of Marsh et al. (2009) who found that participants with higher American College Test (ACT) scores performed better on final retention tests than participants with lower ACT scores. Also, we

predict that aptitude, along with academic achievement and learning and study skills, will be a significant predictor for the incidence and magnitude of the testing effect.

Academic achievement. We hypothesize that prior academic achievement will impact the incidence and magnitude of the testing effect. Although there have been no studies that specifically address this question, we believe that above average students (A and B students) will show a greater benefit for repeated testing than average students (C students and below). This prediction rests on Marsh et al. (2009) findings showing that higher performing students had a more significant increase in final, criterion performance when compared to lower performing students. In addition, we hypothesize that academic achievement will be a significant predictor for the presence of the testing effect.

Learning and study skills variables. We predict that individuals who have developed better strategies for learning and studying will benefit more from repeated testing than those who have not developed superior learning and studying strategies. There have been no studies, to our knowledge, that specifically address learning and study skills. The predictions that we make here are, again, based on findings from Marsh et al. (2009). Students with better learning and study skills should have a better record of academic achievement. We also believe that learning and study skills will be a significant predictor for the presence of the testing effect.

Additional Considerations Regarding the Present Research

Examining the testing effect in the classroom presents both unique challenges and opportunities. McDaniel, Roediger, and McDermott (2007) addressed the lack of control and inherent difficulty of investigating the testing effect in natural contexts. The authors determined that the lack of control results from several key differences between laboratory and classroom settings, namely the inability to control for study time and the inability to distribute tests evenly across time. Although the timing of testing in classroom studies can be controlled with

scheduling, studying behavior cannot be regulated across participants. There may be considerable variation in the amount of study prior to the intervening tests, between intervening tests, and prior to the final assessment all of which may influence the effects of testing. The present study did not attempt to manipulate or control the amount of studying that students engaged in prior to each assessment because the amount of studying will vary by individuals and experimental manipulation of study habits may hinder the educational progress of the students. However, the present study measured and statistically controlled for the amount of studying that each student engaged in prior to each assessment. Measures of studying were gathered via self-report and entered as a covariate during statistical analyses.

In addition, the delay between the presentation of the to-be-learned information and intervening tests or the intervening tests and the final assessment is not often consistent throughout any college. For example, participants in McDaniel et al.'s (2007) study completed quizzes and unit tests at a self-determined pace. Therefore, there was some variation in the amount of time that elapsed between completion of the readings, the initial tests, and the unit exams. The study and test conditions in Agarwal et al.'s (2008) study also were self-paced. The present study was designed around an actual classroom setting in which the delay between the presentation of the information and the initial test varied between assessments. The delay between initial quizzes, unit exams, and cumulative exams also varied. Although these limitations are generally considered undesirable, there is no a priori reason to believe that they would negatively impact the results of the experiment (McDaniel et al., 2007).

Other difficulties in applied research on the testing effect include: (a) the feasibility of giving numerous tests over the course of the semester and (b) the change in complexity of course content as the semester progresses. This particular Introductory Psychology class was structured

to have recurring and frequent assessments. Introductory Psychology at Auburn University is taught in two distinct parts: larger, combined lectures for two class periods followed by smaller, discussion sections once per week. Typically, the discussion sections are led by graduate teaching assistants (GTAs) who teach for part of the class and do some form of course assessment for the remainder. The content for the course varies considerably in its scope and depth throughout the course—ranging from biological to psychosocial mechanisms that underlie behavior and mental processes. The design of the course and assessment devices allowed for statistical comparisons between different topics. For example, Marsh et al. (2009) compared the testing effect across several distinct domains (e.g., biology and chemistry) and directly compared the benefits of testing within each domain.

General Methodology

This study examined the effects of repeated testing in a large naturally occurring Introductory Psychology course. The study's design minimized disruptions in student's learning process and avoided hindering the educational objectives of the instructor. The course instructor developed course objectives, content, and student assessment formats. The researcher worked with the instructor to incorporate the multiple-choice, short-answer, and study conditions, as well the procedures for repeating items across the semester. Grading of assignments and awarding of grades for the course occurred independently of the researcher's conclusions about assessment performance. The researcher received ungraded copies of all quizzes and exams and graded them for the purposes of this study. Students identified themselves using a unique code name and neither the instructor for the course nor the researcher knew any identifiable information about any students.

Course Specifics

The instructor planned and delivered all of the course content for the Monday and Wednesday lectures. GTAs planned and delivered course content for the smaller Friday discussion sections. The instructor also developed the schedule for quizzes, exams, and cumulative exams. A copy of the calendar for the course is found in Appendix A.

Participants

Participants were recruited from two large sections of Introductory Psychology at Auburn University. The number of participants providing complete assessment and survey data varied

over the duration of the study. Overall, 174 participants completed the demographic survey, the LASSI subscales, and a majority of the assessments and preparation questionnaires.

The majority of participants in this study were female (69.7%). Most of the participants were entering (or first-semester) freshmen (54.6%). The remaining participants were sophomores (25.3%), juniors (10.9%), seniors (2.9%), and non-entering freshmen (1.1%). For entering freshmen, the average high school grade point average upon entering college was 3.64 out of 4.0 (SD = .32). For non-entering freshmen and upper-level students, the average college grade point average was 3.04 (SD = .60). The average ACT score for all participants was 25.77 (SD = 3.99). Most participants (59%) rated their academic capabilities as "good," while the remainder rated their academic abilities as "very good" (27.7%), "okay" (12%) and "fair" (1.2%).

Materials

Demographic Questionnaire. Participants completed a demographic questionnaire that asked questions about their gender, class standing, and aptitude scores upon entering college (ACT or SAT scores). To provide a common index to compare across students, SAT scores were converted to comparable ACT scores using concordance statistics for both aptitude tests (American College Test, 2009). In addition to these demographics, students reported either (a) their overall grade point average from high school or (b) their overall college grade point average to date. On the final Assessment Preparation survey, students reported each class they were taking during the current semester, along with estimates of the anticipated grade for each course. From this information, we calculated a rough estimate of the student's grade point average for the semester. The average, self-reported semester GPA for all participants was 3.18 (SD = .48). Although some researchers have concluded that students cannot accurately predict their academic success (Glenberg, Sanocki, Epstein, & Morris, 1987), this self-report procedure provided a rough estimate of academic achievement while maintaining participants' anonymity.

Students also rated their own academic ability ("Overall, how would you rate your academic ability?") using a Likert-type item (1=very strong, 5=poor). A copy of the demographic questionnaire found in Appendix B.

Learning and Study Strategies Inventory, 2nd edition (LASSI-II). The LASSI-II is an 80-item standardized measure that assesses student learning and study skills on 10 components: attitude, motivation, time management, anxiety, concentration, information processing, selecting main ideas, use of study aids, self-testing, and test strategies. We scored each subscale using the grading criteria established by Weinstein and Palmer (2002). In addition, we also compared the participants' calculated score with percentile rankings based on the norm-referenced criteria reported by the publisher. The reliability for each LASSI-II subscale is shown in Table 1.

Table 1

Reliability Scores for the LASSI-II Subscales in the Present Study

Subscale	Cronbach's Alpha
Anxiety and worry	.879
Attitude and interest	.718
Concentration and attention	.886
Information processing	.830
Motivation	.858
Self-testing, reviewing	.835
Selecting main ideas	.894
Support techniques and materials	.691

Time management	.873
Test strategies	.794

These results are comparable to the reliability reported by Weinstein and Palmer (2002). A copy of the LASSI II is found in Appendix C.

Quizzes. There were eight quizzes during the semester that each corresponded to particular chapters in the text. Each quiz contained eight multiple-choice and seven short-answer questions. The researcher carefully selected multiple-choice questions from a publisher-provided test bank using two criteria. First, the questions had to be of moderate difficulty. Although the publisher did not provide item validation, the level of difficulty was determined by criteria provided in the test bank. Second, the questions had to be worded in a way that minimized alterations to the question stem when generating short-answer items from them.

In most cases, the researcher developed short-answer questions by removing the answer alternatives from the multiple-choice questions. In other cases, minor alterations to sentence structure or sentence arrangement were necessary to maintain the central idea of the question. With the exception of several items on the first quiz, short-answer questions required the correct recall of a single term or concept. Therefore, student responses were graded using a binary grading procedure (0=wrong, 1=correct). Other researchers who compared results from a three-point grading system (0=wrong, 1=partially correct response, 2=correct) and a binary grading procedure concluded that there was no significant difference the two grading methods (Marsh et al. 2009).

In addition to multiple-choice and short-answer items, each quiz contained seven study items. The researcher generated these items by modifying selected multiple-choice questions into

summary statements that included the correct answer. Study items served as a control comparison for the multiple-choice and short-answer formats. Examples of each question type are found in Appendix D.

The presentation of multiple-choice, short-answer, and study items were counterbalanced across participants. Table 2 presents an example of the counterbalancing procedure used for all quizzes. The sample question numbers in the table represent individual items selected for inclusion on the study assessments.

Table 2

An Example of the Counterbalancing Procedure for Quizzes

Question Type	Sample student A	Sample student B	Sample student C
Multiple-choice	101, 102, 103, 104, 105, 106, 107, 108	116, 117, 118, 119, 120, 121, 122, 108	109, 110, 111, 112, 113, 114, 115, 108
Short-answer	109, 110, 111, 112, 113, 114, 115	101, 102, 103, 104, 105, 106, 107	116, 117, 118, 119, 120, 121, 122
Study question	116, 117, 118, 119, 120, 121, 122	109, 110, 111, 112, 113, 114, 115	101, 102, 103, 104, 105, 106, 107

Note. Question 108 was repeated in all three multiple-choice conditions to maintain a 15-item quiz for the instructor's grading purposes. This question was not repeated during any future assessment.

Student assessment included three versions of each quiz (termed Forms A, B, and C). Forms A, B, and C of each week's quiz were randomly distributed to participants during class sessions. Although in some cases there were significant differences in student performance among forms, we had no *a priori* reason to anticipate such effects.

Unit Exams. There were four unit exams during the semester that each covered two chapters of classroom readings and lectures. Each exam consisted of 45 multiple-choice questions that primarily covered information presented in the text and two essay questions that primarily covered material from lectures. Although there was some overlap between the text and lectures, all multiple-choice questions were drawn from the textbook test bank. Unit exams were composed of four categories of items: (a) those previously administered as multiple-choice items on a preceding quiz, (b) those previously administered as short-answer items on a preceding quiz, (c) those previously administered as study items on a preceding quiz, and (d) novel items that had not been previously administered on a preceding quiz.

Five multiple-choice questions, five short-answer, and five study items from each of the two quizzes prior to the unit exam were randomly selected for inclusion on each unit exam. Table 3 presents an example of the selection procedure for a unit exam. The sample question numbers in the table represent individual items selected for inclusion on the study assessments.

Table 3

An Example of the Selection Procedure for Unit Exams

	Original Item Type		
Origin	Multiple-choice	Short-answer	Study items

Repeated items			
First quiz	**101**, 102, **103**, **104**, **105**, 106, **107**, 108	**109, 110, 111**, 112, 113, **114, 115**	116, **117, 118**, 119, **120, 121, 122**
Second quiz	**201**, 202, **203**, **204**, **205**, 206, **207**, 208	**209, 210, 211**, 212, 213, **214, 215**	216, **217, 218**, 219, **220, 221, 222**
Novel questions	523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537		

Note. Bolded numbers represent the randomly selected items that were repeated from an earlier quiz.

Each exam contained 10 multiple-choice items, 10 short-answer items, and 10 study items repeated from the two quizzes corresponding to the chapters covered on the unit exam. In addition, students also completed 15 novel items (8 from one chapter, 7 from another). The novel items were used a control condition when comparing performance on these items with preceding study items.

Cumulative Exams. There were two cumulative exams during the semester that each covered four chapters of classroom readings and lectures. Each cumulative exam contained 45 multiple-choice items that primarily cover material presented in the text and two essay questions that primarily cover the lecture material. Although there was some overlap between the text and lectures, all multiple-choice questions were drawn from the textbook test bank. Cumulative exams were composed of five categories of items: (a) items that were previously administered as

multiple-choice items on a preceding quiz and a preceding unit exam, (b) items that were previously administered as short-answer items on a preceding quiz and a preceding unit exam, (c) items that were previously administered as study items on a preceding quiz and a preceding unit exam, (d) items that had been previously administered as 'novel items' on a preceding unit exam, and (e) items that had not been previously on any preceding quiz or unit exam.

Three multiple-choice questions, three short-answer, and three study items from each of the four quizzes prior to the cumulative exam were randomly selected for inclusion on the cumulative exam. In addition, there were three novel items on the cumulative exam. Table 4 presents an example of the selection procedure for a cumulative exam. The sample question numbers in the table represent individual items selected for inclusion on the study assessments.

Table 4

An Example of the Selection Procedure for Cumulative Exams

	Original Item Type		
Origin	Multiple-choice	Short-answer	Study items
Repeated items			
First quiz	101, 102, **103**, **104**, **105**, 106, 107, 108	109**, 110, 111**, 112, 113, **114,** 115	116, 117**, 118**, 119, **120, 121,** 122
Second quiz	201, 202, **203**, **204**, **205**, 206, 207, 208	209**, 210, 211**, 212, 213, **214**, 215	216, 217**, 218**, 219, **220, 221,** 222
Unit 1 Novel items	**523**, 524, 525, 526, 527, **528**, 529, 530,		

	531, 532, **533**, 534, 535, 536, 537		
Third quiz	301, 302, **303**, **304**, **305**, 306, 307, 308	309**, 310, 311**, 312, 313, **314,** 315	316, 317**, 318**, 319, **320, 321**, 322
Fourth quiz	401, 402, **403**, **404**, **405**, 406, 407, 408	409**, 410, 411**, 412, 413, **414,** 415	416, 417**, 418**, 419, **420, 421,** 422
Unit 2 Novel items	**623**, 624, 625, 626, 627, **628**, 629, 630, 631, 632, **633**, 634, 635, 636, 637		
Novel items	**701, 702, 703**		

Note. Bolded numbers represent the randomly selected items that were repeated from an earlier quiz.

Each cumulative exam repeated 12 multiple-choice items, 12 short-answer items, and 12 study items from previous quizzes, and also repeated six multiple-choice items from previous unit exams. Each cumulative examination also contained three novel items, which were used a control condition when comparing performance on these items with preceding study items and novel items from unit exams.

Assessment Preparation Questionnaire. After each assessment (quiz, exam, or cumulative exam) participants completed a brief survey concerning their preparation for the assessment. Students reported how many hours they studied for the assessment, as well as

subjective ratings concerning their own performance (i.e., “How would you rate your performance on today’s quiz/exam?,” “What grade do you expect on this quiz/exam?”).

Procedure

After participants consented to join this study, they provided a unique alphanumeric codename to the researcher, which they used for submitting all items throughout the semester. Therefore, student participation was anonymous. Participants completed the demographic and LASSI questionnaires as an online survey at the beginning of the semester. The class was structured to have regularly occurring quizzes and examinations during regularly scheduled Friday discussion sections. A copy of the ungraded quiz or exam for each participating student was collected from the instructor or graduate teaching assistant. Student responses were graded using a binary grading procedure (0=wrong, 1=correct) and the complied grades were collated with study times, grade expectation, and confidence estimates. In addition, participants completed and returned the Assessment Preparation Questionnaire after each quiz or exam.

Results

Quiz Performance By Chapter

Although units in of Introductory Psychology often contain overlapping or related content, each chapter covers distinct concepts. Some chapter concepts or content may be inherently more difficult than others, thus resulting in differential student performance across content domains. The means and standard deviations for each chapter are included in Figure 1.

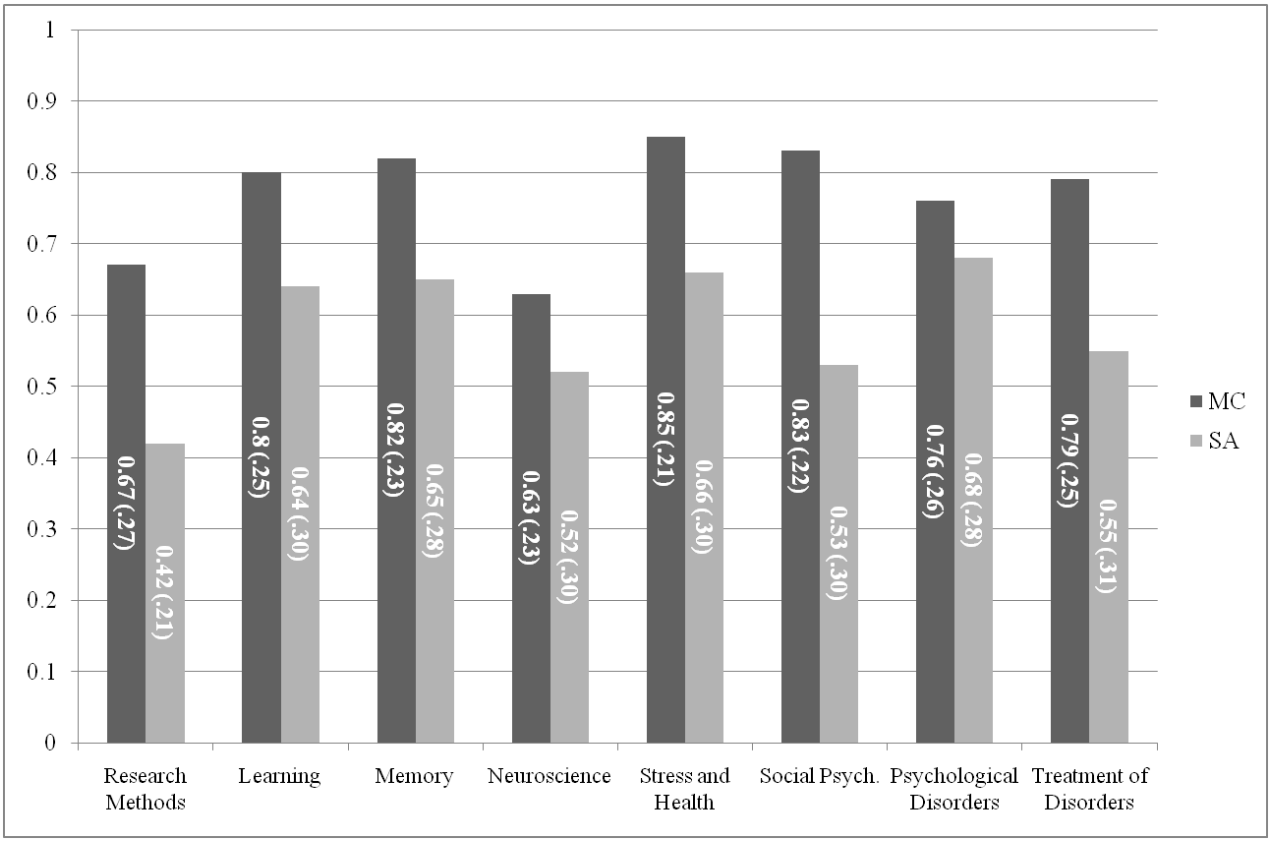

Figure 1. Means and standard deviations for quiz performance. Standard deviations are included in parentheses. MC = multiple-choice. SA = short-answer.

A 2 (format) X 8 (chapter) repeated measures ANOVA was performed for participants with complete data. There was a significant effect of format, F (1, 33) = 140.73, $p < .001$. Participants performed significantly better on multiple-choice items than short-answer items across chapters. Further, there was a significant main effect of chapter, F (7, 231) = 6.12, $p < .001$. Post hoc tests (adjusted by a Bonferroni correction procedure) revealed that performance on Chapter 1 (Research Methods) was significantly lower than performance on Chapter 2 (Learning), Chapter 3 (Memory), Chapter 5 (Stress and Health), and Chapter 7 (Psychological Disorders). There were no other significant different between chapters. There was no significant interaction between format and chapter, F (7, 231) = 1.65, *ns.*

Exam Performance By Chapter

Unit exams served as both a criterion test for previously administered quiz items and as an intervening trial for the upcoming cumulative exam. The means and standard deviations for student performance are included in Figure 2. Performance is displayed by initial quiz question type. Study items were first assessed on unit exams. Novel items were those that did not appear on any previous assessment (i.e., a quiz).

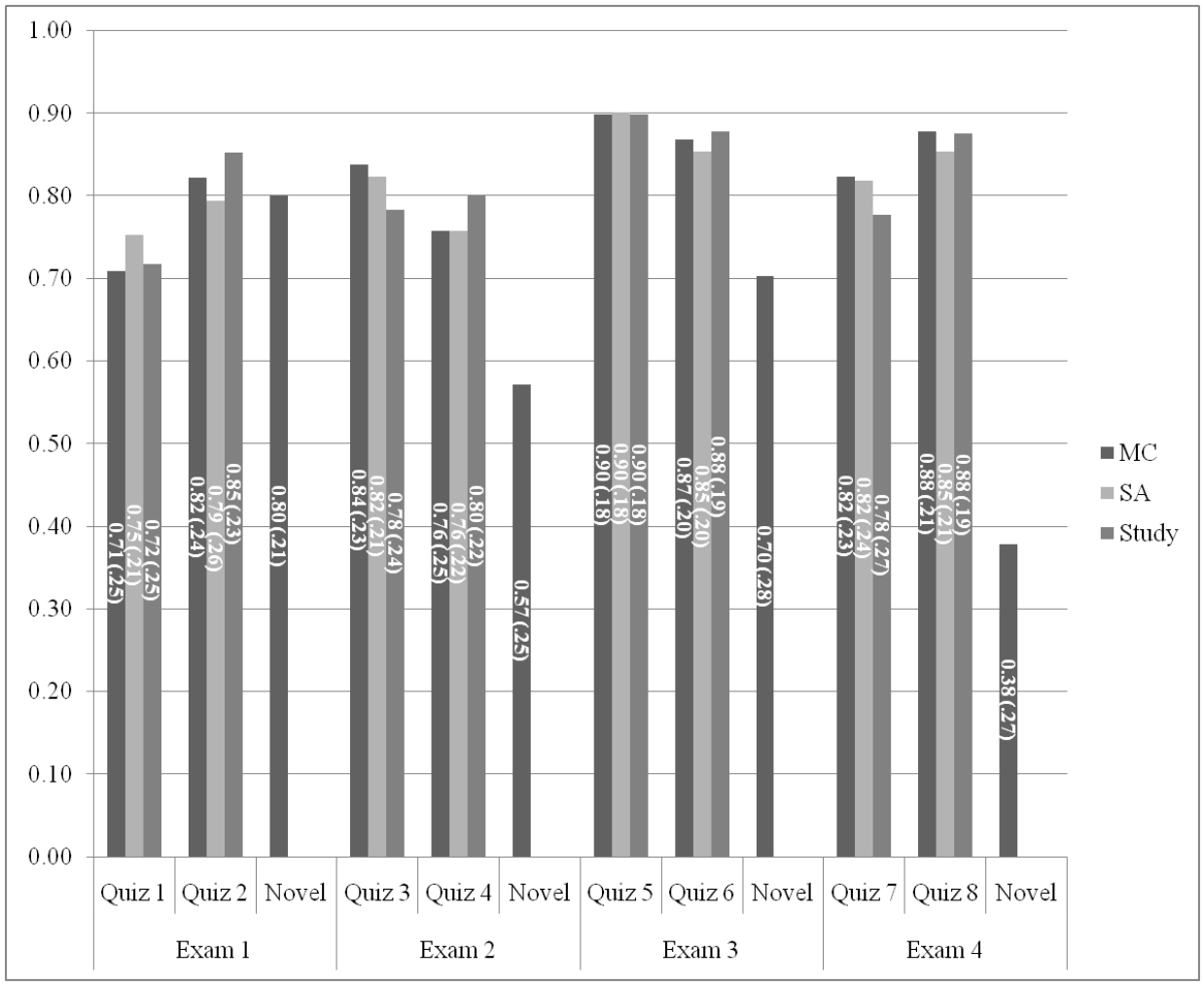

Figure 2. Means and standard deviations for unit exam performance. Standard deviations are included in parentheses. MC = multiple-choice. SA = short-answer.

A 2 (format) X 8 (chapter) repeated measures ANOVA was performed for participants with complete data. There was a significant effect of chapter, F (7, 203) = 3.16, p =.01. Post hoc tests (adjusted by a Bonferroni correction procedure) revealed that multiple-choice and short-answer performance on items repeated from Chapter 1 (Research Methods) was significantly lower than performance on items from Chapter 5 (Stress and Health) and Chapter 8 (Treatment of Disorders). No other comparisons were significant. There was no significant main effect for format, F (1, 29) = .43, *ns*, nor an interaction between format and chapter, F (7, 203) = .66, *ns*.

Later analyses compare multiple-choice and short-answer performance with the read-only and novel conditions.

Cumulative Exam Performance By Chapter

Cumulative exams served as the final criterion test for items repeated on both quizzes and unit exams, as well as for items initially tested on unit exams (novel items). The means and standard deviations for student performance are included in Figure 3. Performance is displayed by initial quiz question type. Study items were first assessed on unit exams and again on cumulative exams. Novel items that appeared previously on unit exams are notated by their unit number. Items appearing only on the cumulative exam are labeled as novel.

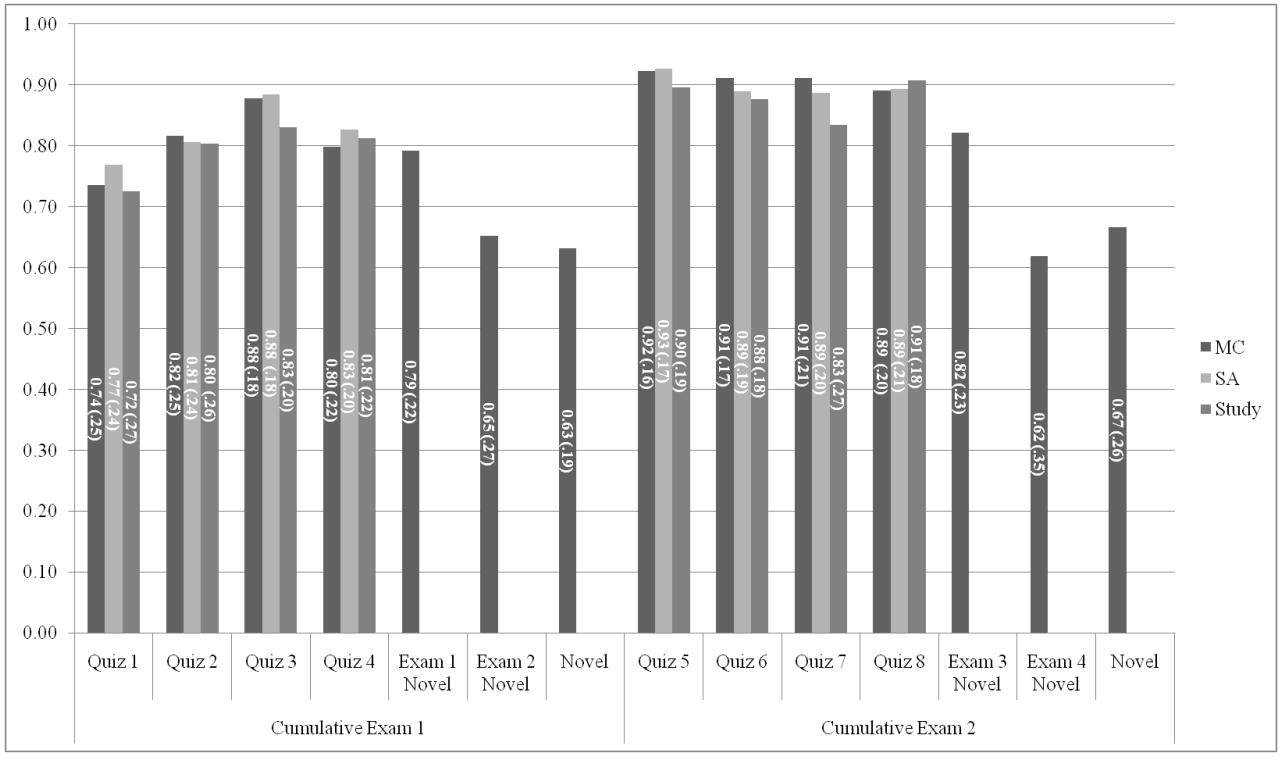

Figure 3. Means and standard deviations for cumulative exam performance. Standard deviations are included in parentheses. MC = multiple-choice. SA = short-answer.

A 2 (format) X 8 (chapter) repeated measures ANOVA was performed for participants with complete data. There was a significant main effect for chapter, F (7, 196) = 6.51, p < .001. Post

hoc tests (adjusted by a Bonferroni correction procedure) revealed that multiple-choice and short-answer performance on items repeated from Chapter 1 (Research Methods) were significantly lower than performance for Chapter 5 (Stress and Health), Chapter 6 (Social), Chapter 7 (Disorders), and Chapter 8 (Treatment). Also, performance on Chapter 2 (Learning) was significantly lower than performance on Chapter 5 (Stress and Health). No other comparisons were statistically significant. Further, there was no significant main effect for format or an interaction between chapter and format ($Fs \leq 1$).

The Effects of Delay on the Testing Effect

Because this study was conducted in a classroom setting, the delay between the initial quiz over the material and the subsequent exams varied throughout the semester. On average, quizzes 1, 3, 5, and 7 occurred 2 weeks before the unit exam and quizzes 2, 4, 6, and 8 occurred the week prior to the unit exam. Exams 2 and 4 were administered closer to the midterm and final than were exams 1 and 3. Therefore, there may be some impact of the delay between initial testing and criterion testing. We conducted several analyses to investigate any effects of delay.

In order to maximize the number of participants with complete data at each point in time, we divided the analyses into four segments. The first segment compared student performance for material from Chapter 1 (Research Methods) with material from Chapter 2 (Learning). The second segment compared student performance for material from Chapter 3 (Memory) with material from Chapter 4 (Neuroscience). The third segment compared student performance for material from Chapter 5 (Stress and Health) with material from Chapter 6 (Social). Finally, segment four compared student performance from material from Chapter 7 (Disorders) with material from Chapter 8 (Treatment).

Four 2 X 2 X 3 repeated measures ANOVA were performed, with the factors of delay (long or short), initial quiz format (multiple-choice or short-answer) and repeated exposure (initial quiz performance, unit exam performance, or cumulative exam performance) being analyzed. In addition, effects of class section was examined as a between subjects factor.

Segment 1. The means and standard deviations for quiz performance, unit exam performance, and cumulative exam performance are displayed in Table 5.

Table 5

Means and Standard Deviations for Segment 1

	Long-delay		Short-delay	
Assessment	MC	SA	MC	SA
Quiz Performance	.67 (.27)	.42 (.21)	.80 (.25)	.64 (.30)
Unit Exam Performance	.71 (.25)	.75(.21)	.82 (.24)	.79 (.26)
Cumulative Exam Performance	.74 (.25)	.77(.24)	.82 (.25)	.81 (.24)

Note. N = 62. MC = Multiple-choice. SA = Short-answer. Standard deviations are included in parentheses.

There were no significant differences between any of the class sections, F (10, 51) = .926, *ns*, nor did it interact with any other variable (*ps* > .05). There were significant main effects for delay, F (1, 51) = 34.515, $p < .001$, format, F (1, 51) = 4.77, $p = .03$, and exposure, F (2, 102) = 35.32, $p < .001$. These main effects were qualified by a significant two-way interaction between delay and exposure, F (2, 102) = 6.11, $p = .003$, a significant two-way interaction between format and exposure, F (2, 102) = 21.457, $p < .001$, and a significant three-way interaction between delay, format, and exposure, F (2, 102) = 4.43, $p = .014$.

The three-way interaction between delay, format, and exposure resulted from significantly lower performance on short-answer quiz items when compared to multiple-choice quiz items across both delay conditions. For the long delay condition, students performed significantly better on initial multiple-choice quiz items when compared to initial short-answer quiz items. There was no significant difference in performance for repeated multiple-choice or short-answer items on either the unit exam or the cumulative exam. The pattern for the short-delay condition was similar.

The two-way interaction between exposure and format resulted because multiple-choice performance was significantly higher than short-answer performance on the initial quiz. There was, however, no significant difference between the two format conditions at either the unit exam or the cumulative exam. The two-way interaction between exposure and delay followed the same pattern: Participants scored significantly better on multiple-choice quiz items when compared to short-answer quiz items. There was no difference between the format conditions at either the unit or cumulative exam.

The main effect of delay resulted from significantly higher performance for the short-delay condition. On average, student performance was 10.5% higher in the short-delay condition when compared to the long-delay condition. Participants also scored significantly higher on multiple-choice items when compared with short-answer items. On average, participants scored 6.1% better on multiple-choice items. Finally, there was a significant increase in performance when comparing quiz performance with either unit exam performance or cumulative exam performance. On average, participants scored 13.6% better on items repeated from the quiz to the unit exam and 13.7% better on items repeated from the quiz to the cumulative exam. There was no significant difference between performance on the unit and cumulative exam.

Segment 2. The means and standard deviations for quiz performance, unit exam performance, and cumulative exam performance are displayed in Table 6.

Table 6

Means and Standard Deviations for Segment 2

	Long-delay		Short-delay	
Assessment	MC	SA	MC	SA
Quiz Performance	.82 (.23)	.65 (.28)	.63 (.23)	.52 (.30)
Unit Exam Performance	.84 (.23)	.82 (.21)	.75 (.25)	.76 (.22)
Cumulative Exam Performance	.92 (.16)	.88 (.19)	.80 (.22)	.83 (.20)

Note. MC = Multiple-choice. SA = Short-answer. Standard deviations are included in parentheses.

There were no significant differences between any of the class sections, F (14, 60) = 1.04, *ns*, nor was there an interaction with any other variable (*ps* > .05). There were significant main effects for delay, F (1, 60) = 25.95, $p < .001$, format, F (1, 60) = 14.73, $p < .001$, and exposure, F (2, 120) = 55.11, $p < .001$. Student performance for the long-delay condition was 9.8% higher than performance in the short-delay condition. Student performance on multiple-choice items also was 7% higher when compared to short-answer questions. Finally, performance was significantly higher on the unit and cumulative exams when compared to initial quiz performance. In addition, performance was significantly higher on the cumulative exam when compared to the unit exam.

These main effects were qualified by a significant two-way interaction between delay and exposure, F (2, 120) = 4.55, $p = .012$ and a significant two-way interaction between format and exposure, F (2, 120) = 12.42, $p < .001$. No other interactions were significant. The two-way

interaction between delay and exposure resulted because there was a significant difference in performance between the long-delay and short-delay conditions on the initial quiz and the cumulative exam. In both conditions, performance for the long-delay condition was significantly higher than for the short-delay condition. There was no significant difference in performance for either delay condition on the unit exam.

The two-way interaction between format and exposure resulted because of a significant difference in performance between multiple-choice and short-answer items on the initial quiz. Participants scored significantly higher on the multiple-choice when compared to the short-answer. There were no significant differences between format conditions on either the unit or cumulative exam.

Segment 3. The means and standard deviations for quiz performance, unit exam performance, and cumulative exam performance are displayed in Table 7.

Table 7

Means and Standard Deviations for Segment 3

	Long-delay		Short-delay	
Assessment	MC	SA	MC	SA
Quiz Performance	.85 (.21)	.66 (.30)	.83 (.23)	.53 (.30)
Unit Exam Performance	.90 (.18)	.90 (.18)	.87 (.20)	.85 (.20)
Cumulative Exam Performance	.91 (.21)	.93 (.17)	.91 (.17)	.89 (.19)

Note. MC = Multiple-choice. SA = Short-answer. Standard deviations are included in parentheses.

There were no significant differences between any of the class sections, F (14, 51) = 1.73, *ns*, nor did class section interact with any other variable (*ps* > .05). There were significant main effects for format, F (1, 51) = 9.83, p = .003, and exposure, F (2, 102) = 33.67, p < .001. Overall, participants scored 7.3% better on multiple-choice questions when compared short-answer questions. Also, performance was significantly lower on the initial quiz when compared to performance on either the unit or cumulative exam. There was no significant difference between overall performance on the unit exam when compared to the cumulative exam. There was also a marginally significant main effect for delay, F (1, 51) = 3.97, p = .052. Overall performance for the long-delay condition was 4.5% higher than overall performance for short-delay condition.

These main effects were qualified by a significant two-way interaction between format and exposure, F (2, 102) = 23.55, p < .001. The interaction between format and exposure resulted because participants scored significantly lower on short-answer quiz questions when compared to multiple-choice quiz questions. There were, however, no significant differences between format conditions on either the unit or cumulative exam. No other interactions between study variables were significant.

Segment 4. The means and standard deviations for quiz performance, unit exam performance, and cumulative exam performance are displayed in Table 8.

Table 8

Means and Standard Deviations for Segment 4

	Long-delay		Short-delay	
Assessment	MC	SA	MC	SA
Quiz Performance	.76 (.26)	.68 (.28)	.79 (.25)	.55 (.31)
Unit Exam Performance	.82 (.23)	.82(.24)	.88 (.21)	.85 (.21)

Cumulative Exam Performance	.88 (.18)	.89 (.20)	.89 (.20)	.89 (.21)

Note. MC = Multiple-choice. SA = Short-answer. Standard deviations are included in parentheses.

There were no significant differences between any of the class sections, F (1, 39) = 1.05, *ns*, nor did it interact with any other variable (*ps* > .05). There were significant main effects for format, F (1, 39) = 19.26, $p < .001$, and exposure, F (2, 78) = 22.38, $p < .001$. On average, participants scored 6.9% better on multiple-choice items when compared to short-answer items. Also, participants scored significantly higher on both the unit and cumulative exams when compared with quiz performance. There was no significant difference between performance on the unit and cumulative exams.

These main effects were qualified by a significant two-way interaction between delay and exposure, F (2, 78) = 5.81, $p = .004$, a significant two-way interaction between format and exposure, F (2, 78) = 10.02, $p < .001$, a significant two-way interaction between delay and format, F (1, 39) = 6.74, $p = .013$, and a significant three-way interaction between delay, format, and exposure, F (2, 78) = 9.4, $p < .001$. The three-way interaction resulted from a significant difference between multiple-choice quiz performance and short-answer quiz performance for the short-delay condition. Participants scored significantly higher on multiple-choice quiz items than short-answer quiz items but only for the short-delay condition. There was no significant difference between multiple-choice and short-answer quiz performance for the long-delay condition. Also, there was no significant difference between repeated multiple-choice or repeated short-answer performance on either the unit or cumulative exam for both the long and short-delay conditions.

The two-way interaction between delay and format resulted from a significant difference between multiple-choice and short-answer performance for the short-delay condition. Participants scored significantly higher on multiple-choice items than short-answer items for the short-delay condition. There was, however, no significant difference in performance between formats for the long-delay condition.

The two-way interaction between delay and exposure resulted because performance in the long-delay condition was significantly higher for the initial quiz, but the short-delay condition was significantly higher for both the unit and cumulative exam. The two-way interaction between format and exposure resulted because of a significant difference between formats on the initial quiz. Performance for multiple-choice items on the initial quiz was significantly higher than performance on short-answer quiz items. There was, however, no significant difference between formats on either the unit exam or the cumulative exam.

Conclusions concerning the effects of delay on the testing effect. The naturally occurring schedule of assessments for the course resulted in having half of the quizzes closer to the unit exam. Also, half of the unit exams were closer to the cumulative exam than the rest. Because other researchers have concluded that delay between initial and subsequent assessments may influence the incidence and magnitude of the testing effect (e.g., Spitzer, 1939; Wheeler et al., 2003), we decided to perform a series of analyses to examine whether any effects in our results could be explained by the delay variable. We divided the data into four distinct segments to facilitate more data points for the analyses. Each segment examined performance over either a long-delay (approximately 2 weeks from initial quiz to first criterion test) or a short-delay (approximately 1 week from initial quiz to first criterion test). These delay conditions also

reflected the proximity of initial quizzes, as well as unit exams, in relation to the cumulative exam.

There was a significant effect of format in all four segments. Overall, participants performed better on multiple-choice questions than short-answer questions. There was also a significant effect of delay. For three of the four segments, participants performed better on both unit and cumulative exams when compared to initial quiz performance. For these three segments, there was no significant difference between unit and cumulative exam performance. For the fourth segment (Segment 2), participants performed better on the cumulative exam compared to the unit exam. Finally, two of the four segments had a significant effect of delay. For Segment 1, performance was better when it was associated for a short-delay, but for Segment 2 this pattern was reversed. This contradictory finding, along with a failure to find an effect of delay for Segments 3 and 4, does not provide conclusive evidence for a consistent pattern of delay on the testing effect. Also, these main effects were qualified by numerous interactions across conditions.

Overall, the interactions present across the four segments resulted from two main sources. First, multiple-choice quiz performance was significantly higher than short-answer quiz performance. For most interactions between other study variables and format, there was a significant difference between formats on the initial quiz, but not for either the unit or cumulative exam. Second, the source for most of the interactions presented here was exposure. Overall, participants performed significantly lower on the initial quiz when compared to either the unit or cumulative exam.

In conclusion, there was no clear or consistent pattern for the effects of delay across the four segments. This finding, in conjunction with significant differences between chapters, made

it necessary to analyze each chapter of data separately. Analyzing separate chapters allowed us to examine trends and patterns stable across differing content without undue influence of delay and initial performance.

Performance for Items Repeated From Quiz 1

The results for the eight sections that follow (Performance for Items Repeated from Quiz 1 – Quiz 8) are structured identically. First, we present the means and standard errors in graphical form. Performance is displayed by initial quiz question type. Study items were first assessed on unit exams and again on cumulative exams. Both the unit exam and the cumulative exam contained novel items. Novel items from the unit exam were repeated on the cumulative exam. A concise summary of the significant effects are presented in Appendix E. Second, statistical results from repeated-measures ANOVAs for proportion correct, along with an interpretation of the findings, are reported. We used a Bonferroni correction procedure for all multiple-comparisons to protect against Type I errors. Third, statistical results from a repeated-measures ANOVCA controlling for study time are reported. Finally, we present conditional analyses for proportion correct by initial item accuracy.

The means and standard errors for material from the Research Methods chapter are presented in Figure 4.

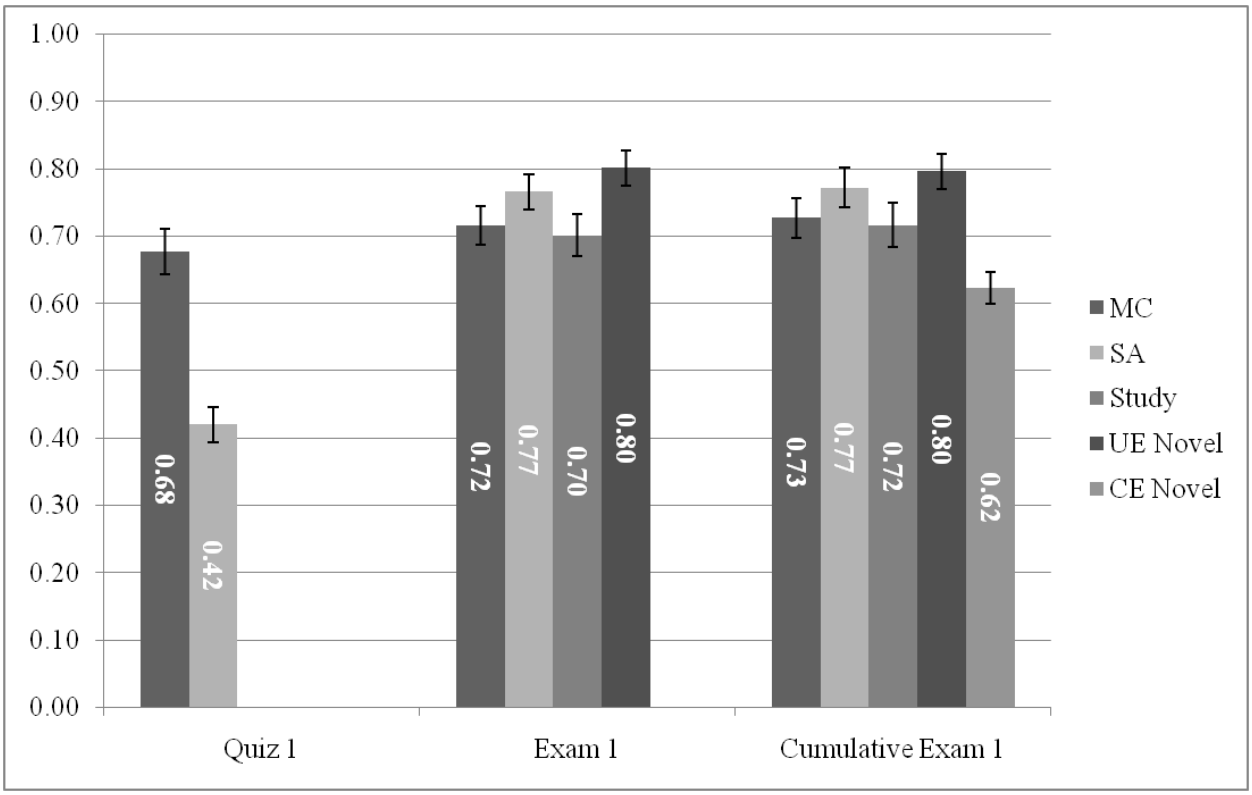

Figure 4. N = 67. Mean performance across assessments by original item type. Error bars represent standard error. MC = multiple-choice. SA = short-answer. UE Novel = unit exam novel. CE Novel = cumulative exam novel.

Repeated-measures ANOVA for proportion correct. We conducted a repeated-measures ANOVA to compare performance across assessments. There was a significant main effect of assessment time, F (10, 499) = 15.47, $p < .001$.

***Quiz performance*.** Overall, participants performed significantly higher on multiple-choice items when compared to short-answer items. This finding supports the notion that multiple-choice questions have higher initial performance relative to short-answer because a recognition task is easier than a recall task.

***Exam performance*.** There was no significant difference for performance when multiple-choice items were repeated from Quiz 1 to Exam 1. There was, however, a significant increase in

performance when short-answer items were repeated on Exam 1. On average, participants' scores increased by 35 percent. On Exam 1, there was no significant difference among performance on multiple-choice, short-answer, read-only, and novel items. In summary, student performance on short-answer items increased significantly from Quiz 1 to Exam 1, but there were no significant differences among any of the groups on Exam 1. Any benefit of short-answer quizzing over multiple-choice quizzing was not evident on Exam 1.

***Cumulative exam performance*.** There was no significant difference for performance when initial multiple-choice, short-answer, or read-only items were repeated from Exam 1 to Cumulative Exam 1. There also was no significant difference for performance on novel items that appeared on Exam 1 and subsequently on Cumulative Exam 1. In addition, there were no significant differences found among performance on initial multiple-choice, short-answer, read-only, and novel items repeated from Exam 1 on the cumulative exam. Participants performed better on initial short-answer questions and novel items repeated from Exam 1 compared with novel items on the cumulative exam.

In summary, there was no significant difference in performance on items for which students took an initial multiple-choice test, an initial short-answer test, or read summary statements. In addition, performances in those conditions were not significantly different from items that were first assessed on the unit exam.

Repeated-measures ANCOVA controlling for study time. To control for any effects of study time on student performance, we conducted a repeated-measures ANCOVA. We found a significant main effect for assessment time, F (10, 480) = 3.24, p = .002. There were no significant interactions of study time for Quiz 1, Exam 1, or Cumulative Exam 1 with assessment time (ps >.05).

Conditional analyses for item accuracy. We conducted a 2 (correct, incorrect) X 2 (multiple-choice, short-answer) within-subjects repeated-measures ANOVA to examine differential effects that may have resulted from getting an initial item correct or incorrect. The means and standard errors for Exam 1 and Cumulative Exam 1 by Quiz 1 accuracy are presented in Table 9.

Table 9

Mean Performance for Correct and Incorrect Items from Quiz 1

	Exam 1		Cumulative Exam 1	
Quiz Accuracy	MC	SA	MC	SA
Correct	.84 (.03)	.86 (.03)	.88 (.05)	.89 (.04)
Incorrect	.64 (.04)	.63(.05)	.59 (.07)	.66 (.06)

Note. N = 37 for Exam 1 and 35 for Cumulative Exam 1. MC = Multiple-choice. SA = Short-answer. Standard error is included in parentheses.

For Exam 1, there was a significant main effect for initial accuracy, F (1, 36) = 11.54, p = .002. On average, participants performed 21.6% better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was no significant main effect for format, F (1, 36) = .24, *ns.* Also, there was no significant interaction between initial accuracy and format, F (1, 36) = .22, *ns.*

For Cumulative Exam 1, there was a significant main effect for initial accuracy, F (1, 34) = 27.77, $p < .001$. On average, participants scored 26.4% better on items that were initially answered correctly when compared to items that were initially answer incorrectly. There was no

significant main effect for format, F (1, 34) = .35, *ns.* Also, there was no significant interaction between initial accuracy and format, F (1, 34) = .42, *ns.*

Conclusions from Quiz 1 content. On the initial quiz, participants performed worse on short-answer quiz items than for multiple-choice quiz items. On the unit exam, however, there was no difference in performance between multiple-choice and short-answer. Also, there was no significant difference between multiple-choice, short-answer, read-only and novel questions appearing on the unit exam. Student performance did not increase from the unit exam to the cumulative exam for any item type. On the cumulative exam, students performed significantly better on repeated short-answer items and novel items repeated from the unit exam when compared to novel items appearing only on the cumulative exam.

We found that participants performed better on questions that were initially answered correctly when compared to questions that were initially answered incorrectly. When initial accuracy was considered, any differences between multiple-choice and short-answer vanished. These findings held for performance on both the unit and cumulative exam.

The present part of the study supports the hypothesis that multiple-choice quiz performance would be higher than short-answer quiz performance. There was, however, no increase in performance for repeated short-answer questions when compared with multiple-choice questions on either the unit or cumulative exam. The results presented here suggest that any prior exposure (either through initial quizzing or a read-only study condition) were equally beneficial on subsequent exam performance.

Performance for Items Repeated From Quiz 2

The means and standard errors for material from the learning chapter are presented in Figure 5.

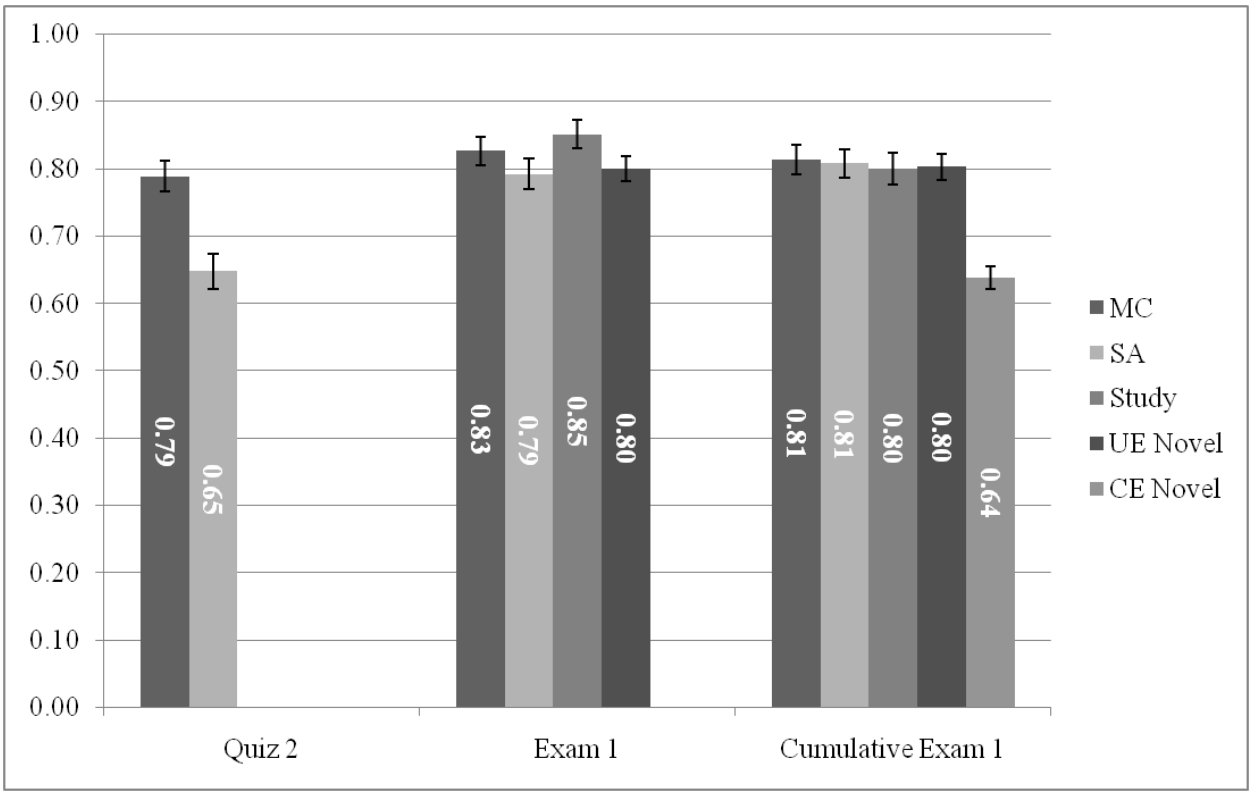

Figure 5. N = 125. Mean performance across assessments by original item type. Error bars represent standard error. MC = multiple-choice. SA = short-answer. UE Novel = unit exam novel. CE Novel = cumulative exam novel.

Repeated-measures ANOVA for proportion correct. We conducted a repeated-measures ANOVA to compare performance across assessments. There was a significant main effect of assessment time, F (10, 798) = 11.28, $p < .001$.

***Quiz performance*.** There was a significant difference between performance on multiple-choice and short-answer quiz questions. Overall, participants scored 14% higher on multiple-choice questions.

***Exam Performance*.** There was no significant difference in performance for multiple-choice items that were repeated from Quiz 2 to Exam 1. There was, however, a significant increase in performance for short-answer questions that were repeated from Quiz 2 to Exam 1.

Performance on initial short-answer items increased by 14.4% on Exam 1. On Exam 1, there was no significant difference between performance on initial multiple-choice, initial short-answer, read-only, and novel items. Therefore, there was no differential effect of short-answer at this level of analysis. In summary, short-answer performance increased on Exam 1, but the short-answer condition did not produce better overall retention when compared with other experimental and control conditions.

***Cumulative Exam Performance*.** There was no significant difference in performance scores for multiple-choice, short-answer and read-only items from Exam 1 to Cumulative Exam 1. In addition, there was no significant difference in performance between novel items that appeared on the unit exam and their subsequent presentation on the cumulative exam. The initial multiple-choice condition, initial short-answer condition, the read-only condition, and novel items from the unit exam were significantly higher than novel items that appeared on the cumulative exam.

Repeated-measures ANCOVA controlling for study time. To control for any effects of study time on student performance, we conducted a repeated-measures ANCOVA. We found a significant main effect main effect for assessment time, F (10, 765) = 3.07, p = .005. There were no significant interactions of study time for Quiz 2, Exam 1, or Cumulative Exam 1 with assessment time (ps >.05).

Conditional analyses for item accuracy. We conducted a 2 (correct, incorrect) X 2 (multiple-choice, short-answer) within-subjects repeated-measures ANOVA to examine differential effects that may have resulted from getting an initial item correct or incorrect. The means and standard errors for Exam 1 and Cumulative Exam 1 by Quiz 2 accuracy are presented in Table 10.

Table 10

Mean Performance for Correct and Incorrect Items from Quiz 2

	Exam 2		Cumulative Exam 1	
Quiz Accuracy	MC	SA	MC	SA
Correct	.94 (.03)	.83 (.05)	.88 (.04)	.89 (.04)
Incorrect	.61 (.08)	.58(.08)	.77 (.07)	.70 (.07)

Note. N = 33. MC = Multiple-choice. SA = Short-answer. Standard error is included in parentheses.

For Exam 1, there was a significant main effect for initial accuracy, F (1, 32) = 26.86, $p < .001$. On average, participants performed 29.5% better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was no significant main effect for format, F (1, 32) = .87, *ns.* Also, there was no significant interaction between initial accuracy and format, F (1, 32) = .40, *ns.*

For Cumulative Exam 1, there was a significant main effect for initial accuracy, F (1, 32) = 8.84, $p = .006$. On average, participants scored 15.2% better on items that were initially answered correctly when compared to items that were initially answer incorrectly. There was no significant main effect for format, F (1, 32) = .22, *ns.* Also, there was no significant interaction between initial accuracy and format, F (1, 32) = .74, *ns.*

Conclusions from Quiz 2 content. The results from Quiz 2 are similar to those from Quiz 1. Participants performed better on multiple-choice quiz questions when compared to short-answer quiz questions. And, although there was a significant increase in short-answer performance from Quiz 2 to Exam 1, the performance level for this group was not significantly

different from repeated multiple-choice items, read-only items, or novel question appearing on Exam 1. Also, there was no significant difference between initial question type, read-only questions, or novel items repeated from the unit exam on the cumulative exam. There was, however, a significant difference between these conditions and novel questions appearing only on the cumulative exam. These effects held even when study time was statistically controlled.

Conclusions from the conditional analyses for material from Quiz 2 largely mirrored material from Quiz 1. Participants performed significantly better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was no differential effect of format or an interaction between these two variables.

This part of the study confirms that prior exposure to quiz material (either through initial testing or study) is equally beneficial for performance on future assessments. There was no added benefit of short-answer when compared to other exposure types. These findings are consistent with the findings from the previous chapter.

Performance for Items Repeated From Quiz 3

The means and standard errors for material from the Memory chapter are presented in Figure 6.

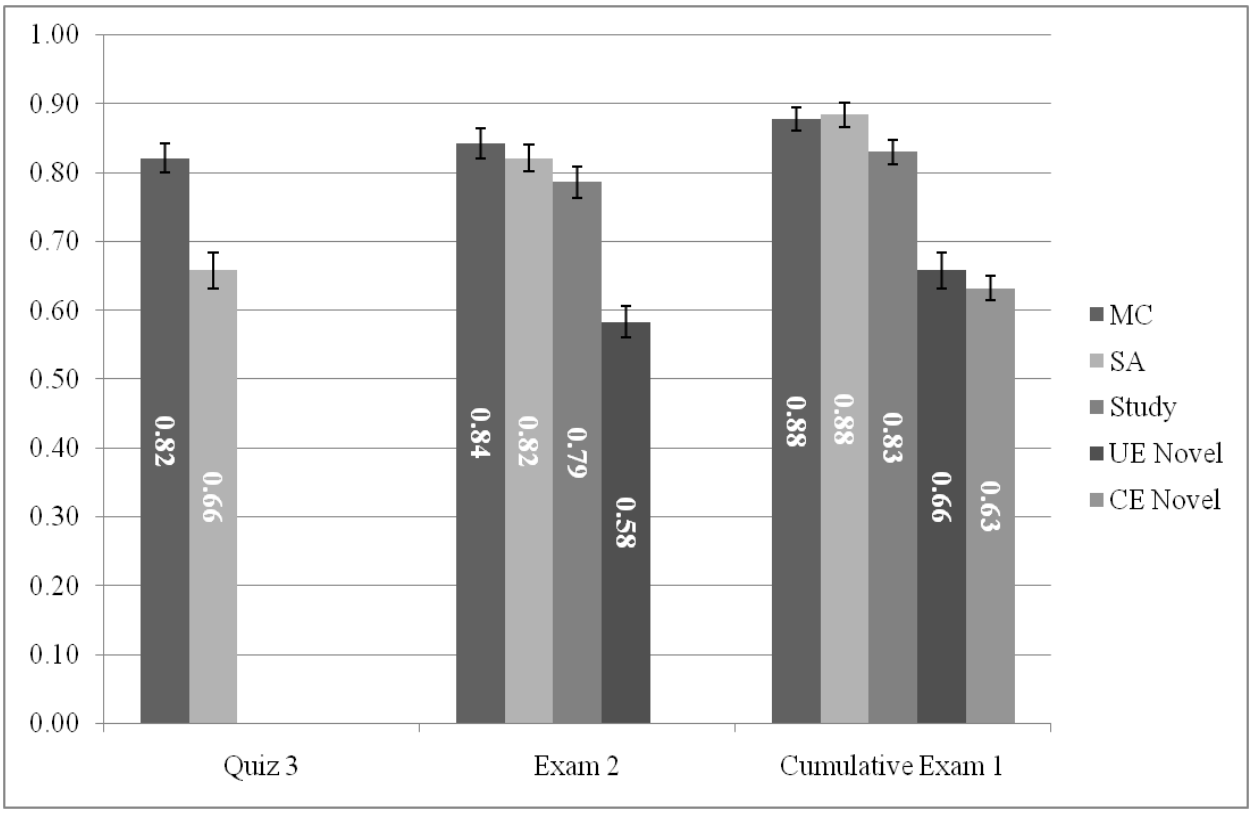

Figure 6. N = 112. Mean performance across assessments by original item type. Error bars represent standard error. MC = multiple-choice. SA = short-answer. UE Novel = unit exam novel. CE Novel = cumulative exam novel.

Repeated-measures ANOVA for proportion correct. We conducted a repeated-measures ANOVA to compare performance across assessments. There was a significant main effect of assessment time, F (10, 860) = 32.04, $p < .001$.

***Quiz performance*.** There was a significant difference between the scores on initial multiple-choice items and short-answer items. On average, participants scored 16.4% higher on the multiple-choice questions when compared to the short-answer questions. This finding supports similar findings for Chapters 1 and 2.

***Exam performance*.** There was no significant change in performance for multiple-choice items from Quiz 3 to Exam 2. There was, however, a significant increase in short-answer

performance. On average, participants increased their performance score by 16.4% from Quiz 3 to Exam 2. On Exam 2, there was no significant difference in the performance rates for questions that were previously assessed as multiple-choice or short-answer. In addition, there was no significant difference between the multiple-choice and short-answer conditions with the read-only study condition. All three conditions, multiple-choice, short-answer, and read-only, were significantly higher than novel questions appearing on Exam 2.

***Cumulative Exam 1 performance*.** There was no significant difference in performance for the multiple-choice, short-answer, and read-only conditions when comparing scores from Exam 2 to Cumulative Exam 1. There also was no significant difference for student performance for novel items from Exam 2 that were repeated Cumulative Exam 1. Although there was no significant difference between multiple-choice, short-answer, and read-only conditions for Cumulative Exam 1, performance in these conditions was significantly higher than performance for novel items repeated from Exam 2 and novel items appearing only on Cumulative Exam 1.

For the Memory chapter, there does appear to be some immediate advantage for multiple-choice quiz questions. However, students perform equally well on future questions regardless of how they were assessed initially. Study items produced similar levels of performance at later points in time. For the present portion of the study, there does not appear to be a differential effect or advantage of short-answer questions for future test performance.

Repeated-measures ANCOVA controlling for study time. To control for any effects of study time on student performance, we conducted a repeated-measures ANCOVA. We found a significant main effect for assessment time, F (10, 689) = 7.58, $p < .001$. There were no significant interactions of study time for Quiz 3, Exam 2, or Cumulative Exam 1 with assessment time (ps >.05).

Conditional analyses for item accuracy. We conducted a 2 (correct, incorrect) X 2 (multiple-choice, short-answer) within-subjects repeated-measures ANOVA to examine differential effects that may have resulted from getting an initial item correct or incorrect. The means and standard errors for Exam 2 and Cumulative Exam 1 by Quiz 3 accuracy are presented in Table 11.

Table 11

Mean Performance for Correct and Incorrect Items from Quiz 3

	Exam 2		Cumulative Exam 1	
Quiz Accuracy	MC	SA	MC	SA
Correct	.76 (.06)	.88 (.05)	.88 (.05)	.89 (.04)
Incorrect	.69 (.07)	.54(.07)	.73 (.07)	.63 (.07)

Note. N = 40. MC = Multiple-choice. SA = Short-answer. Standard error is included in parentheses.

For Exam 2, there was a significant main effect for initial accuracy, F (1, 39) = 18.38, $p < .001$. On average, participants performed 20.6% better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was no significant main effect for format, F (1, 39) = .16, *ns.* These effects were qualified by a significant interaction between initial accuracy and format, F (1, 39) = 5.81, $p = .02$. Post-hoc tests reveal that there was a significant difference for Exam 1 performance between multiple-choice and short-answer conditions for questions initially answered correctly, t (123) = .3.05, $p = .003$. There was, however, no significant difference for format among questions initially answered incorrectly, t (123) = 1.39, *ns.*

For Cumulative Exam 1, there was a significant main effect for initial accuracy, F (1, 39) = 11.83, p = .001. On average, participants scored 20.6% better on items that were initially answered correctly when compared to items that were initially answer incorrectly. There was no significant main effect for format, F (1, 39) = 1.00, *ns.* Also, there was no significant interaction between initial accuracy and format, F (1, 39) = 1.44, *ns.*

Conclusions from Quiz 3 content. The results from this chapter largely mirror those found in chapters 1 and 2. On the initial quiz, there was a significant difference between multiple-choice and short-answer performance. Overall, students performed better on multiple-choice questions. There was, however, no significant increase in multiple-choice performance from the quiz to the unit exam. Short-answer performance did significantly increase from the initial quiz to the unit exam. Performance in these conditions was not significantly different from performance in the study condition. Therefore, any prior exposure to the material led to similar performance rates. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

There was no significant increase performance for either multiple-choice, short-answer, read-only, or unit exam novel questions on the cumulative exam. These conditions were significantly better than novel questions appearing only on the cumulative exam. This finding, again, supports the notion that any pre-exposure to the item provides an advantage on subsequent assessments of the same item. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

Conditional analyses examining the effects of answering an initial item either correctly or incorrectly revealed a significant difference between the two conditions. Overall, participants performed better on items that were initially answered correctly when compared to items that

were initially answered incorrectly. Unlike the previous findings reported in earlier parts of this study, there was a significant interaction between multiple-choice and short-answer performance for items that were initially answered correctly. Participants had a higher accuracy rate for initially correct multiple-choice questions when compared to initially correct short-answer questions. There was no difference between multiple-choice and short-answer formats for items that were initially answered incorrectly.

The results from the present chapter are largely consistent with those of chapters 1 and 2. Although multiple-choice testing results in superior performance on the first assessment, there is no significant difference between any of the prior exposure conditions on subsequent assessments.

Performance for Items Repeated From Quiz 4

The means and standard errors for material from the Neuroscience chapter are presented in Figure 7.

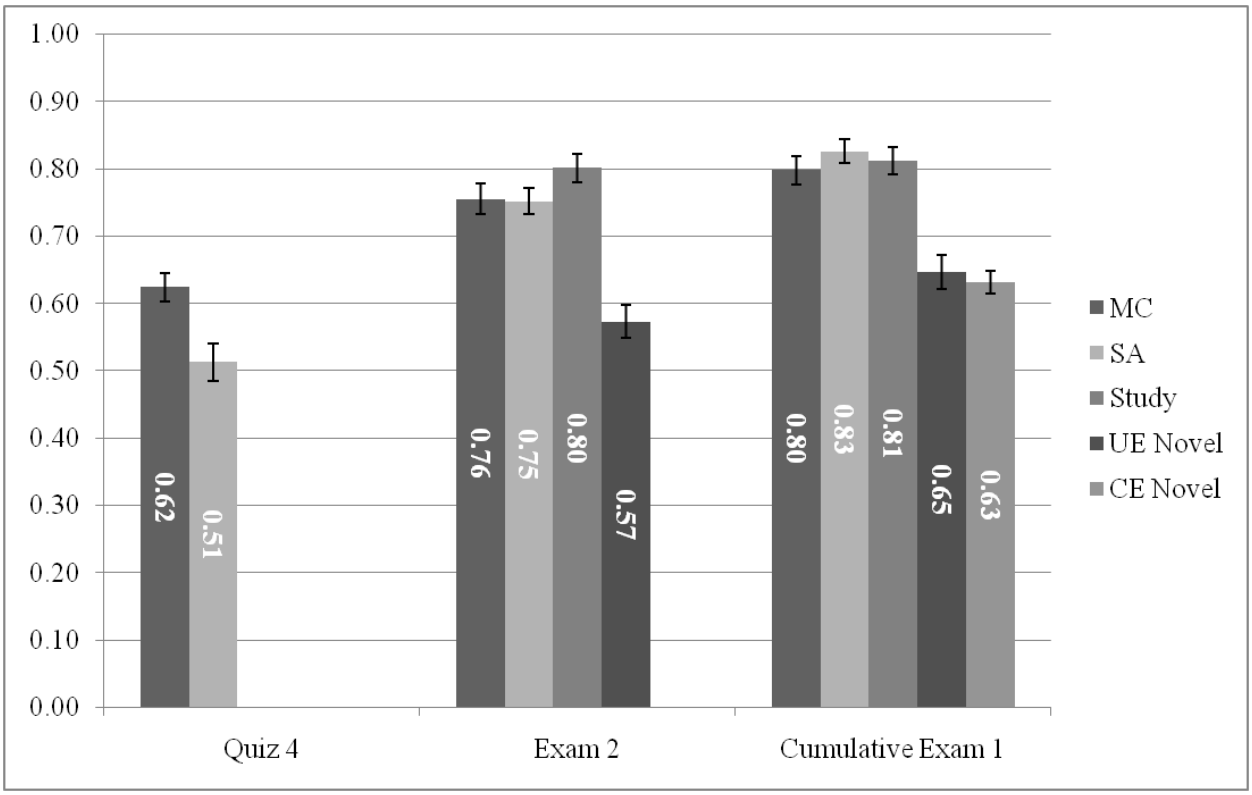

Figure 7. N = 117. Mean performance across assessments by original item type. Error bars represent standard error. MC = multiple-choice. SA = short-answer. UE Novel = unit exam novel. CE Novel = cumulative exam novel.

Repeated-measures ANOVA for proportion correct. We conducted a repeated-measures ANOVA to compare performance across assessments. There was a significant main effect of assessment time, F (10, 870) = 28.89, $p < .001$.

***Quiz performance*.** There was no significant difference between multiple-choice and short-answer performance for Quiz 4. Student performance on multiple-choice questions was the lowest for this chapter when compared to all other chapters in the course. Student performance

on the short-answer questions was the second lowest for this chapter when compared to other chapters in the course.

***Exam performance*.** There was a significant increase in student performance for multiple-choice items from Quiz 4 to Exam 2. On average, student performance increased by 13.1%. Also, there was a significant increase in performance for short-answer items from Quiz 4 to Exam 2. Students increased their scores 24.2% on average. On Exam 2, there was no significant difference between initial multiple-choice and short-answer conditions, with the read-only condition. However, performance in all three conditions was significantly higher than performance for the novel questions appearing on Exam 2.

***Cumulative Exam performance*.** There was no significant increase in performance for either multiple-choice, short-answer, or read-only items repeated from Exam 2 to Cumulative Exam 1. Also, there was no difference in performance for novel items on Exam 2 that were repeated on the cumulative exam. There were, however, significant differences between performance on multiple-choice, short-answer, and read-only items when compared with novel items repeated from Exam 2 and novel items appearing only on the cumulative exam. On average, performance in the multiple-choice, short-answer, and read-only conditions ranged from 15.1-17.9% higher than Exam 2 novel items and ranged from 16.5-19.4% higher than novel items on the cumulative exam. There was no significant difference between performance on repeated novel items from Exam 2 and novel items on the cumulative exam.

Repeated-measures ANCOVA controlling for study time. To control for any effects of study time on student performance, we conducted a repeated-measures ANCOVA. We found a significant main effect for assessment time, F (10, 738) = 5.61, $p < .001$. There were no

significant interactions of study time for Quiz 4, Exam 2, or Cumulative Exam 1 with assessment time (*p*s >.05).

Conditional analyses for item accuracy. We conducted a 2 (correct, incorrect) X 2 (multiple-choice, short-answer) within-subjects repeated-measures ANOVA to examine differential effects that may have resulted from getting an initial item correct or incorrect. The means and standard errors for Exam 2 and Cumulative Exam 1 by Quiz 4 accuracy are presented in Table 12.

Table 12

Mean Performance for Correct and Incorrect Items from Quiz 4

	Exam 2		Cumulative Exam 1	
Quiz Accuracy	MC	SA	MC	SA
Correct	.89 (.03)	.88 (.03)	.91 (.03)	.90 (.03)
Incorrect	.38 (.05)	.44(.05)	.56 (.05)	.63 (.05)

Note. N = 80. MC = Multiple-choice. SA = Short-answer. Standard error is included in parentheses.

For Exam 2, there was a significant main effect for initial accuracy, F (1, 79) = 163.90, $p < .001$. On average, participants performed 47.5% better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was no significant main effect for format, F (1, 79) = .45, *ns.* There also was no significant interaction between initial accuracy and format, F (1, 79) = .41, *ns.*

For Cumulative Exam 1, there was a significant main effect for initial accuracy, F (1, 79) = 68.10, $p < .001$. On average, participants scored 31.2% better on items that were initially

answered correctly when compared to items that were initially answer incorrectly. There was no significant main effect for format, F (1, 79) = .88, *ns.* Also, there was no significant interaction between initial accuracy and format, F (1, 79) = .72, *ns.*

Conclusions from Quiz 4 content. Unlike the previous chapters, there was no difference between multiple-choice and short-answer performance on the initial quiz. This finding, coupled with this chapter having some of the lowest scores overall, suggests that this chapter was either: (a) more difficult in scope than the other chapters in the present study, or (b) that the assessment questions used for this chapter were substantially different from those used in other chapters. The former is a more plausible explanation because the material covered was neuroscience.

Students performed significantly better on the unit exam when compared to the initial quiz. This finding held for both repeated multiple-choice and short-answer questions. Performance for the read-only questions was also similar to that of multiple-choice and short-answer questions. Any prior exposure to the items (multiple-choice, short-answer, or read-only) resulted in superior performance when compared to novel questions appearing on the unit exam. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

There was no significant increase in performance for repeated multiple-choice, repeated short-answer, read-only, or novel questions from the unit exam on the cumulative exam. Again, any prior exposure to the questions resulted in superior performance when compared to novel questions appearing on the cumulative exam. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

The conditional analyses in this chapter largely mirror the results from chapters 1 and 2. Performance for questions that were initially answered correctly was higher on subsequent

assessments that performance for questions that were initially answered incorrectly. Unlike the results for chapter 3, there was no interaction between item format and initial accuracy.

In summary, there was no advantage of either multiple-choice or short-answer at either the initial quiz or any subsequent assessment. Any prior exposure, either through initial quizzing or studying, provided an advantage on later performance.

Performance for Items Repeated From Quiz 5

The means and standard errors for material from the Stress and Health chapter are presented in Figure 8.

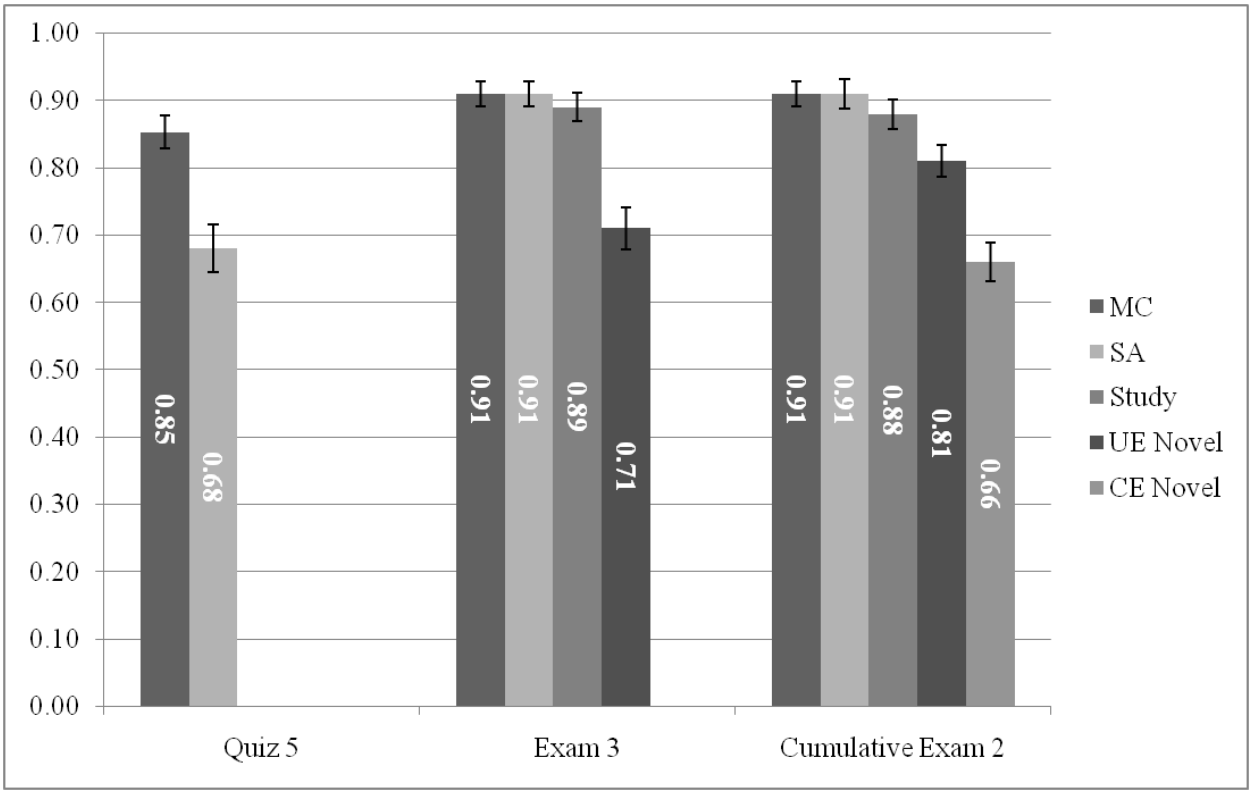

Figure 8. N = 77. Mean performance across assessments by original item type. Error bars represent standard error. MC = multiple-choice. SA = short-answer. UE Novel = unit exam novel. CE Novel = cumulative exam novel.

Repeated-measures ANOVA for proportion correct. We conducted a repeated-measures ANOVA to compare performance across assessments. There was a significant main effect of assessment time, F (10, 540) = 16.38, $p < .001$.

***Quiz performance*.** There was a significant difference between multiple-choice and short-answer performance for Quiz 5. On average, participants scored 17.3% higher on the multiple-choice when compared to short-answer. This finding mirrors findings from quizzes 1, 2, and 3.

***Exam performance*.** There was no significant difference in performance for multiple-choice items repeated from Quiz 5 to Exam 3. There was, however, a significant increase in performance for short-answer questions that were repeated from Quiz 5 to Exam 3. On average, participants increased their score by 22.9% from Quiz 5 to Exam 3. On Exam 3, there was no significant difference in performance among the initial multiple-choice, initial short-answer, and read-only conditions. Performance in these conditions was significantly higher than performance for novel questions on Exam 3.

***Cumulative Exam performance*.** There was no significant difference between the multiple-choice, short-answer, and read-only conditions when comparing performance from Exam 3 to Cumulative Exam 2. There also was no significant change in performance for items that appeared as novel items on Exam 3 and subsequently on Cumulative Exam 2. There was no significant difference in performance between the short-answer, read-only, and Exam 3 novel conditions for Cumulative Exam 2. The multiple-choice condition was significantly higher than the Exam 3 novel condition. Performance in these conditions (MC, SA, RO, and Exam 3 Novel) was significantly higher than for novel items appearing only on Cumulative Exam 2.

Repeated-measures ANCOVA controlling for study time. To control for any effects of study time on student performance, we conducted a repeated-measures ANCOVA. We found a

significant main effect for assessment time, F (10, 482) = 2.52, p = .015. There were no significant interactions of study time for Quiz 5, Exam 3, or Cumulative Exam 2 with assessment time (ps >.05).

Conditional analyses for item accuracy. We conducted a 2 (correct, incorrect) X 2 (multiple-choice, short-answer) within-subjects repeated-measures ANOVA to examine differential effects that may have resulted from getting an initial item correct or incorrect. The means and standard errors for Exam 3 and Cumulative Exam 2 by Quiz 5 accuracy are presented in Table 13.

Table 13

Mean Performance for Correct and Incorrect Items from Quiz 5

	Exam 3		Cumulative Exam 2	
Quiz Accuracy	MC	SA	MC	SA
Correct	.87 (.05)	.86 (.06)	.81 (.06)	.87 (.06)
Incorrect	.47 (.08)	.84(.06)	.69 (.08)	.79 (.07)

Note. N = 34. MC = Multiple-choice. SA = Short-answer. Standard error is included in parentheses.

For Exam 3, there was a significant main effect for initial accuracy, F (1, 33) = 20.27, p < .001. On average, participants performed 21.3% better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was also a significant main effect for format, F (1, 33) = 17.52, p < .001. On average, participants performed 18.4% better on multiple-choice when compared to short-answer items. These main effects were qualified by a significant interaction between initial accuracy and format, F (1, 33) = 14.23, p =

.001. There was no significant difference between multiple-choice and short-answer when they were answered correct initially. Students performed better on short-answer questions that were initially answered incorrectly when compared to multiple-choice questions that were initially answered incorrectly.

For Cumulative Exam 2, there was a marginally significant main effect for initial accuracy, F (1, 33) = 4.10, p = .05. On average, participants scored 9.6% better on items that were initially answered correctly when compared to items that were initially answer incorrectly. There was no significant main effect for format, F (1, 33) = 3.51, *ns.* Also, there was no significant interaction between initial accuracy and format, F (1, 33) = .30, *ns.*

Conclusions from Quiz 5 content. On the initial quiz, students performed better on multiple-choice items than short-answer items. This difference between formats dissipated on the unit exam: There was no difference between the multiple-choice, short-answer, and the read-only conditions. However, performance on these item types was better than performance on novel items. This finding is consistent with other chapters and suggests that any prior exposure to items, either through quizzing or studying, is beneficial on subsequent assessments. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

Also, there was no significant increase in performance for multiple-choice, short-answer, read-only, and unit exam novel items from the unit exam to the cumulative exam. Performance for these items was significantly higher than performance for novel items appearing on the cumulative exam. This finding, again, suggests that any prior exposure to the items has similar effects on performance across subsequent assessments. These effects remained even when

additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

Conditional analyses concerning performance on the unit exam revealed significant effects of both initial accuracy and format, as well as an interaction between these two variables. Overall, students scored better on items that were initially answered correctly when compared to items that were initially answered incorrect and on multiple-choice questions when compared to short-answer questions. However, there was no significant difference between formats for items that were initially answered correctly, but students performed better on initially incorrect short-answer questions than initially incorrect multiple-choice questions. For the conditional analyses concerning performance on the cumulative exam, only an effect of initial accuracy was influential. Students performed better on items that were initially answered correctly when compared to items that were initially answered incorrectly. This pattern, not the pattern from the unit exam, is similar to those found in previous chapters.

Results from the Stress and Health chapter are largely reflective of results found in previous chapters. There was an initial advantage for multiple-choice over short-answer, but this advantage was not maintained across subsequent assessments. There was, however, some benefit of prior exposure (MC, SA, or RO) compared to novel presentations (at either unit or cumulative exam).

Performance for Items Repeated From Quiz 6

The means and standard errors for material from the Social Psychology chapter are presented in Figure 9.

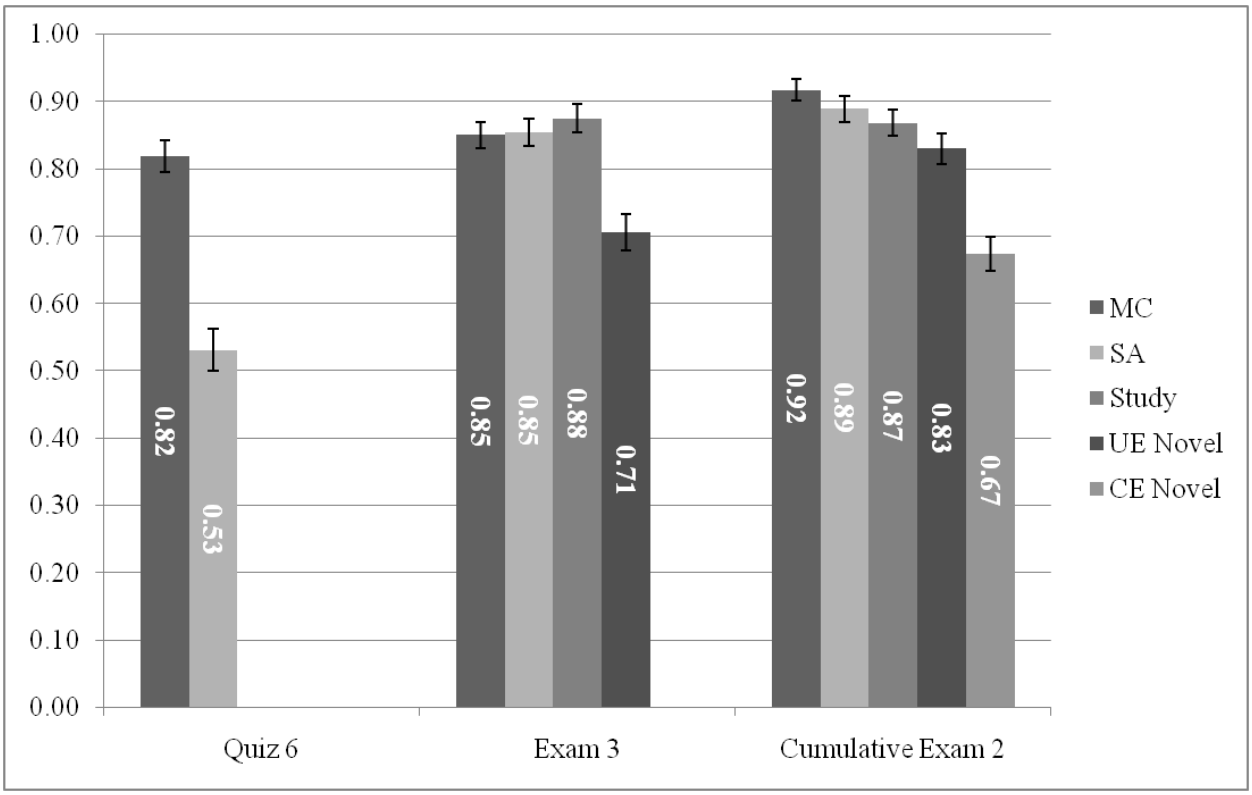

Figure 9. N = 96. Mean performance across assessments by original item type. Error bars represent standard error. MC = multiple-choice. SA = short-answer. UE Novel = unit exam novel. CE Novel = cumulative exam novel.

Repeated-measures ANOVA for proportion correct. We conducted a repeated-measures ANOVA to compare performance across assessments. There was a significant main effect of assessment time, F (10, 738) = 28.77, $p < .001$.

***Quiz performance*.** There was a significant difference between multiple-choice and short-answer performance for Quiz 6. On average, participants scored 28.8% better on the multiple-choice when compared with the short-answer questions. This finding is similar to findings from quizzes 1, 2, 3, and 5.

***Exam performance*.** There was no significant difference between performance on multiple-choice items from Quiz 6 to Exam 3. There was, however, a significant increase in short-answer performance. Participants, on average, scored 32.3% better on short-answer items repeated on Exam 3 when compared to performance on Quiz 6. Also, there were no significant differences in performance between repeated multiple-choice items, repeated short-answer items, and read-only item on Exam 3. However, these conditions were significantly higher than performance for novel items.

***Cumulative Exam performance*.** There was no significant difference in performance for multiple-choice, short-answer, and read-only from Exam 3 on Cumulative Exam 2. Also, there was no significant difference between performance for novel items on Exam 3 that were repeated on Cumulative Exam 2. Performance in these conditions (MC, SA, RO, novel from Exam 3) was significantly higher than performance on novel items appearing only on Cumulative Exam 2.

Repeated-measures ANCOVA controlling for study time. To control for any effects of study time on student performance, we conducted a repeated-measures ANCOVA. We found a significant main effect for assessment time, F (10, 657) = 4.27, $p < .001$. There were no significant interactions of study time for Quiz 6, Exam 3, or Cumulative Exam 2 with assessment time (ps >.05).

Conditional analyses for item accuracy. We conducted a 2 (correct, incorrect) X 2 (multiple-choice, short-answer) within-subjects repeated-measures ANOVA to examine differential effects that may have resulted from getting an initial item correct or incorrect. The means and standard errors for Exam 3 and Cumulative Exam 2 by Quiz 6 accuracy are presented in Table 14.

Table 14

Mean Performance for Correct and Incorrect Items from Quiz 6

	Exam 3		Cumulative Exam 2	
Quiz Accuracy	MC	SA	MC	SA
Correct	.85 (.05)	.70 (.06)	.69 (.07)	.64 (.07)
Incorrect	.44 (.07)	.70(.06)	.61 (.07)	.61 (.07)

Note. N = 42. MC = Multiple-choice. SA = Short-answer. Standard error is included in parentheses.

For Exam 3, there was a significant main effect for initial accuracy, F (1, 41) = 14.38, $p < .001$. On average, participants performed 20.2% better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was no significant main effect for format, F (1, 41) = 1.64, *ns*. These findings were qualified by a significant interaction between initial accuracy and format, F (1, 41) = 13.68, $p = .001$. Students performed better on initial multiple-choice items than initial short-answer items if they were answered correctly on Quiz 6. However, if the questions were answered incorrectly on Quiz 6 the pattern was reversed.

For Cumulative Exam 2, there was no significant main effect for initial accuracy, F (1, 41) = 2.35, *ns.* Also, there was no significant main effect for format, F (1, 41) = .49, *ns* nor an interaction between initial accuracy and format, F (1, 41) = .36, *ns.*

Conclusions from Quiz 6 content. Similar to findings from chapters 1, 2, 3, and 5, there was an advantage of multiple-choice quizzing when compared to short-answer quizzing. This advantage disappeared on the unit exam because performance for both conditions, as well as the read-only condition, was not significantly different. However, performance for repeated items from these groups was better than performance on novel questions appearing on the unit exam.

These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

There was no significant increase in performance for the multiple-choice, short-answer, read-only, or unit exam novel items when they were repeated from the unit exam to the cumulative exam. Performance for these repeated items were higher than performance for novel questions appearing on the cumulative exam. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

For the unit exam, students performed better on repeated presentations of the same item when the initial presentation was answered correctly. Conditional analyses also revealed a significant interaction between initial accuracy and format: Performance on multiple-choice questions that were initially correct was higher than performance on short-answer questions that were initially correct. Interestingly, this pattern was reversed for items that were answered incorrectly. For the cumulative exam, subsequent performance was better for items that were initially answered correctly when compared to items that were initially answered incorrectly. There were no differential effects by format.

The findings from the Social Psychology chapter are consistent with findings from other chapters. Although there appears to be some initial advantage of multiple-choice quiz questions when compared to short-answer quiz questions, this difference disappears at both the unit and cumulative exam. Some prior exposure to the item does appear to benefit performance when compared to information with which there was no previous exposure or assessment.

Performance for Items Repeated From Quiz 7

The means and standard errors for material from the Psychological Disorders chapter are presented in Figure 10.

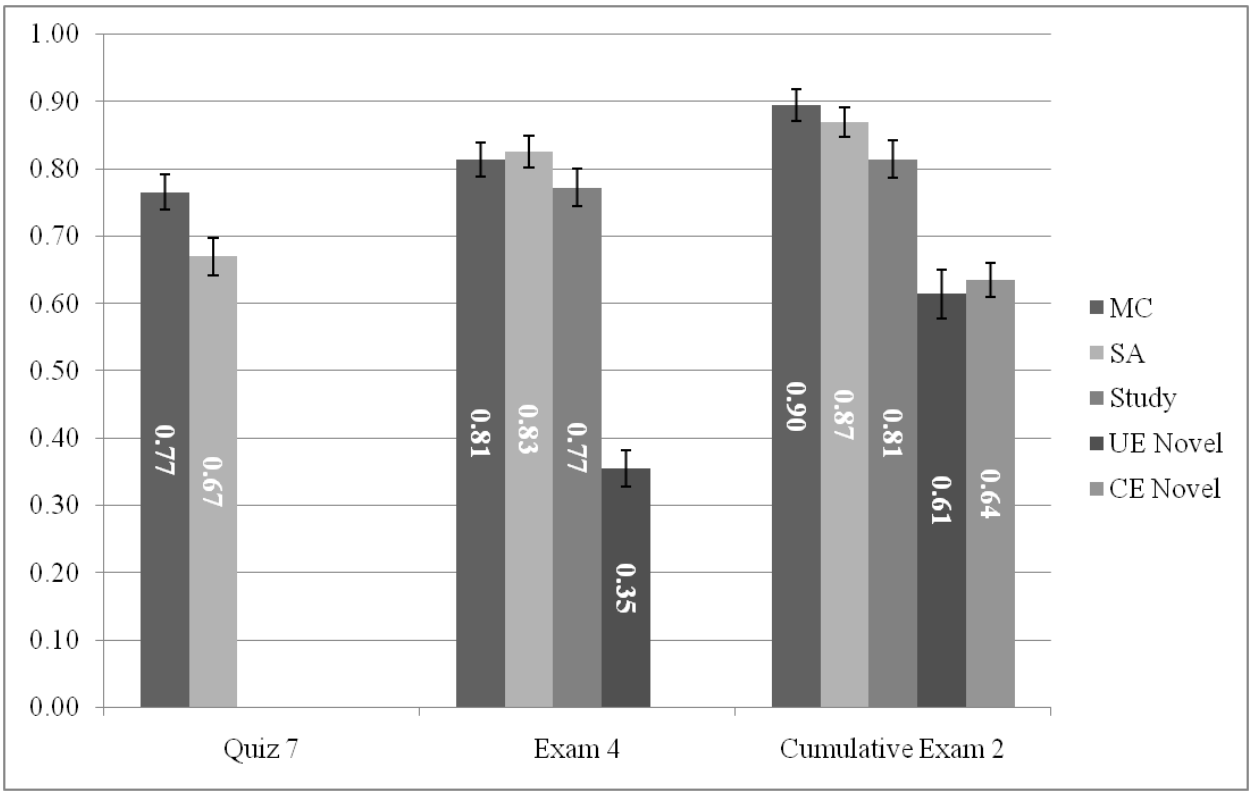

Figure 10. N = 95. Mean performance across assessments by original item type. Error bars represent standard error. MC = multiple-choice. SA = short-answer. UE Novel = unit exam novel. CE Novel = cumulative exam novel.

Repeated-measures ANOVA for proportion correct. We conducted a repeated-measures ANOVA to compare performance across assessments. There was a significant main effect of assessment time, F (10, 675) = 37.33, $p < .001$.

***Quiz performance*.** There was no significant difference between multiple-choice and short-answer performance on Quiz 7. This finding, unlike the finding for the neuroscience chapter, is not the result of lower overall performance on this chapter compared to the others.

***Exam performance*.** There was not a significant difference in performance for repeated multiple-choice between Quiz 7 and Exam 4. There was, however, a significant increase in performance for repeated short-answer items from Quiz 7 to Exam 4. On average, participants increased their score on these items by 15.4%. On Exam 4, there were no significant differences in performance for items that were previously administered as multiple-choice, short-answer, or read-only. Performance in all three conditions was significantly higher than performance on novel items appearing on Exam 4. This finding should be interpreted with caution because performance on Exam 4 novel items was uncharacteristically low when compared with novel items across all other chapters.

***Cumulative Exam performance*.** There was no significant difference in performance for multiple-choice, short-answer, and read-only items when compared from Exam 4 to Cumulative Exam 2. There was, however, a significant increase in performance for novel items from Exam 4 that were repeated on Cumulative Exam 2. On average, student performance increased 26%. Again, this finding should be interpreted with caution because of the uncharacteristically low performance for novel items appearing first on Exam 4. There were no significant difference in performance among multiple-choice, short-answer, and read-only items on Cumulative Exam 2. These three conditions were significantly higher than both the novel questions repeated from Exam 4 and the novel questions appearing only on Cumulative Exam 2. The novel questions from Exam 4 and the novel questions appearing only on Cumulative Exam 2 were not significantly different from each other.

Repeated-measures ANCOVA controlling for study time. To control for any effects of study time on student performance, we conducted a repeated-measures ANCOVA. We found a significant main effect for assessment time, F (10, 606) = 10.01, $p < .001$. There were no

significant interactions of study time for Quiz 7, Exam 4, or Cumulative Exam 2 with assessment time (*p*s >.05).

Conditional analyses for item accuracy. We conducted a 2 (correct, incorrect) X 2 (multiple-choice, short-answer) within-subjects repeated-measures ANOVA to examine differential effects that may have resulted from getting an initial item correct or incorrect. The means and standard errors for Exam 4 and Cumulative Exam 2 by Quiz 7 accuracy are presented in Table 15.

Table 15

Mean Performance for Correct and Incorrect Items from Quiz 7

	Exam 4		Cumulative Exam 2	
Quiz Accuracy	MC	SA	MC	SA
Correct	.81 (.05)	.83 (.05)	.74 (.06)	.75 (.06)
Incorrect	.58 (.07)	.67 (.07)	.74 (.06)	.67 (.07)

Note. N = 44. MC = Multiple-choice. SA = Short-answer. Standard error is included in parentheses.

For Exam 4, there was a significant main effect for initial accuracy, F (1, 43) = 11.33, p = .002. On average, participants performed 19.3% better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was no significant main effect for format, F (1, 43) = 1.36, *ns*. There was no significant interaction between initial accuracy and format, F (1, 43) = .47, *ns.*

For Cumulative Exam 2, there was no significant main effect for initial accuracy, F (1, 42) = .89, *ns.* Also, there was no significant main effect for format, F (1, 43) = .45, *ns,* nor an interaction between initial accuracy and format, F (1, 43) = 1.60, *ns.*

Conclusions from Quiz 7 content. Similar to the results for chapter 4, there was no significant difference between multiple-choice and short-answer conditions on the initial quiz. This finding, however, was not coupled with lower rates of performance for this chapter when compared to the other chapters in the study. There was no increase in performance for repeated multiple-choice items from the quiz to the unit exam, but there was a significant increase for repeated short-answer items. There was, however, no significant difference in performance between these two conditions, as well as the read-only condition, on the unit exam. Performance on multiple-choice, short-answer, and read-only questions were higher than performance on novel items appearing on the unit exam. Performance on the novel items was uncharacteristically low, however. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

There was no significant increase in performance in Cumulative Exam 1 for either the multiple-choice, short-answer, or read-only items when they were repeated from the unit exam to the cumulative exam. And, although there was a significant increase in performance for the novel questions repeated from the unit exam, performance for these questions was still lower than for the multiple-choice, short-answer, and read-only questions. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

For conditional analyses for both unit and cumulative exam performance, students performed better on subsequent presentation of items when they had initially answered the item

correctly. There was no effect of format or an interaction between these two variables. This finding is consistent with many of the chapters discussed previously.

The findings from this chapter are generally consistent with findings from other chapters. Similar to Chapter 4, there were no differences in quiz performance for item type. There were no differences across any of repeated conditions on either the unit or the cumulative exam. Participants scored significantly higher on repeated items when compared to novel items at both the unit and cumulative exam.

Performance for Items Repeated From Quiz 8

The means and standard errors for material from the Treatment of Psychological Disorders chapter are presented in Figure 11.

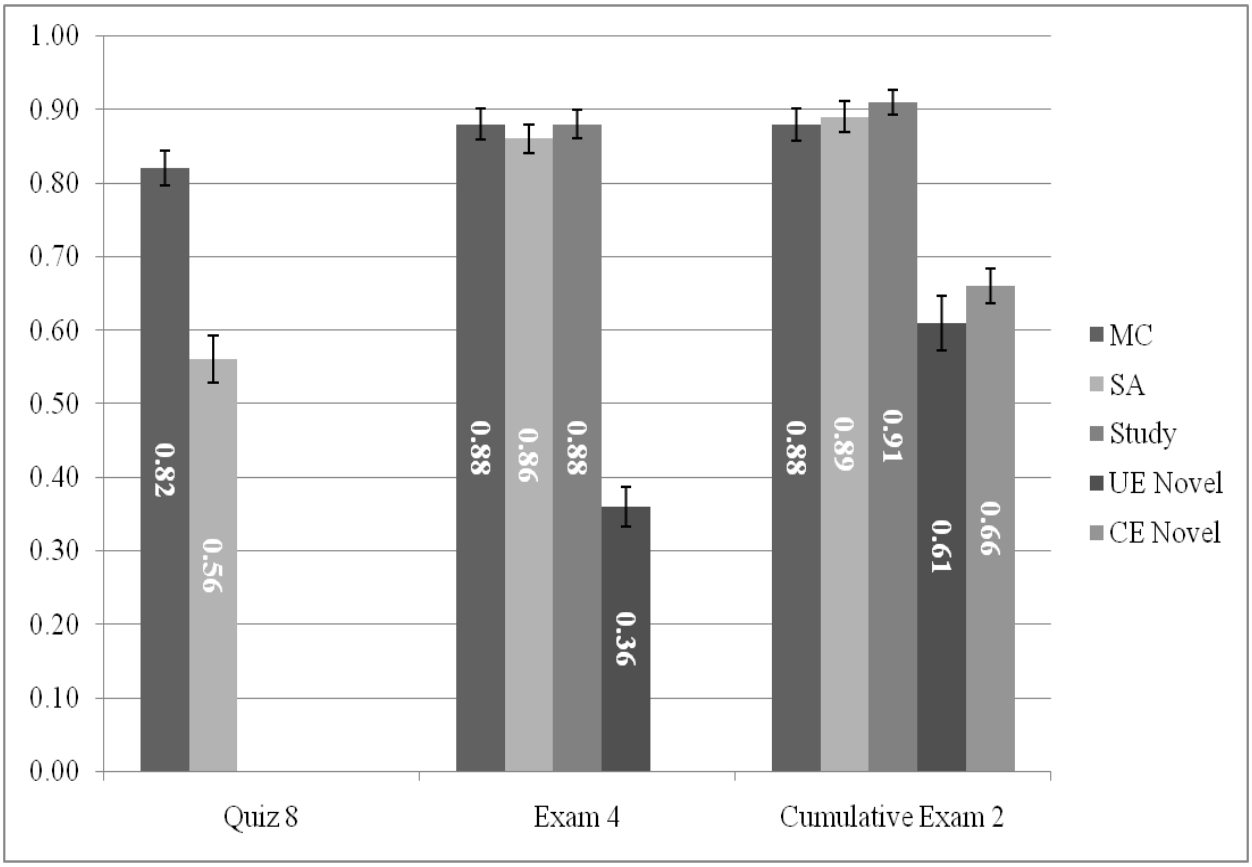

Figure 11. N = 94. Mean performance across assessments by original item type. Error bars represent standard error. MC = multiple-choice. SA = short-answer. UE Novel = unit exam novel. CE Novel = cumulative exam novel.

Repeated-measures ANOVA for proportion correct. We conducted a repeated-measures ANOVA to compare performance across assessments. There was a significant main effect of assessment time, F (10, 625) = 57.57, $p < .001$.

***Quiz performance*.** There was a significant difference between multiple-choice and short-answer performance for Quiz 8. On average, participants scored 26.1% better on the multiple-choice items when compared to the short-answer items. This finding is consistent with findings from quizzes 1, 2, 3, 5, and 6.

***Exam performance*.** There was no significant difference in performance for multiple-choice items from Quiz 8 to Exam 4. There was, however, a significant increase in performance for short-answer items repeated from Quiz 8. On average, participants scored 29.8% better on Exam 4 when compared with Quiz 8. Also, there was no significant difference in performance for repeated multiple-choice, repeated short-answer and read-only items on Exam 4. Performance in these conditions was significantly higher than performance for novel questions appearing on Exam 4. Similar to data from the previous chapter, performance on the novel questions were uncharacteristically low.

***Cumulative Exam performance*.** There was no significant difference in performance for repeated multiple-choice, repeated short-answer, or read-only items from Exam 4 to Cumulative Exam 2. There was, however, a significant increase in performance for novel items that appeared first on Exam 4 and were subsequently repeated on Cumulative Exam 2. On Cumulative Exam 2, there was no significant difference in performance among repeated multiple-choice items,

repeated short-answer items, and repeated read-only items. These three groups were significantly higher than both novel questions repeated from Exam 4 and novel items appearing only on Cumulative Exam 2. There was no significant difference between performance on novel questions repeated from Exam 4 and novel items appearing only on the cumulative exam.

Repeated-measures ANCOVA controlling for study time. To control for any effects of study time on student performance, we conducted a repeated-measures ANCOVA. We found a significant main effect for assessment time, F (10, 518) = 16.05, $p < .001$. There were no significant interactions of study time for Quiz 8, Exam 4, or Cumulative Exam 2 with assessment time (*p*s >.05).

Conditional analyses for item accuracy. We conducted a 2 (correct, incorrect) X 2 (multiple-choice, short-answer) within-subjects repeated-measures ANOVA to examine differential effects that may have resulted from getting an initial item correct or incorrect. The means and standard errors for Exam 4 and Cumulative Exam 2 by Quiz 8 accuracy are presented in Table 16.

Table 16

Mean Performance for Correct and Incorrect Items from Quiz 8

	Exam 4		Cumulative Exam 2	
Quiz Accuracy	MC	SA	MC	SA
Correct	.83 (.05)	.81 (.05)	.72 (.06)	.72 (.06)
Incorrect	.55 (.07)	.66(.06)	.67 (.06)	.71 (.06)

Note. N = 53. MC = Multiple-choice. SA = Short-answer. Standard error is included in parentheses.

For Exam 4, there was a significant main effect for initial accuracy, F (1, 52) = 17.89, $p < .001$. On average, participants performed 21.7% better on items that were initially answered correctly when compared to items that were initially answered incorrectly. There was no significant main effect for format, F (1, 52) = 1.81, *ns*. There was no significant interaction between initial accuracy and format, F (1, 43) = 2.82, *ns.*

For Cumulative Exam 2, there was no significant main effect for initial accuracy, F (1, 52) = .46, *ns.* Also, there was no significant main effect for format, F (1, 52) = .23, *ns*, nor an interaction between initial accuracy and format, F (1, 52) = .23, *ns.*

Conclusions from Quiz 8 content. Consistent with findings from chapters 1, 2, 3, 5 and 6, performance on multiple-choice quiz items was better than short-answer quiz items. Although there was no increase in performance for multiple-choice items from the quiz to the unit exam, there was an increase in short-answer performance. On the unit exam, however, there was no difference between performance on repeated multiple-choice items, repeated short-answer items, or read-only items. Performance on items from those groups was better than performance on novel items appearing on the unit exam. Similar to Chapter 7, performance for novel items on the unit exam were uncharacteristically low. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

On the cumulative exam, there was no increase in performance for initial multiple-choice, initial short-answer, or read-only items that were repeated from the unit exam to the cumulative exam. Although performance for repeated novel items from the unit exam increased, performance on these items was lower than for any other repeated items. Performance on novel questions appearing only on the unit exam was lower that for the repeated multiple-choice, short-

answer, or read-only questions. These effects remained even when additional analyses controlled for study times across the quiz, unit exam, and the cumulative exam.

Similar to findings from conditional analyses from Chapter 7, students performed better on items that were initially answered correctly when compared to items that were initially answered incorrectly. This finding held for both the unit and cumulative exam. There was no effect of format or an interaction between the two variables.

For the Treatment of Disorders chapter, the results presented here are consistent with findings in earlier chapters. There was an initial advantage of multiple-choice over short-answer, but this advantage held only for the initial quiz. There were differential effects of item format (including studying) at both the unit and cumulative exam. Performance on items that had some prior exposure, either through testing or studying, was higher than performance on items that had never been encountered.

Multiple-regression Models for Individual Difference Variables

We calculated two new variables (termed *testing change scores*) by subtracting quiz performance from cumulative exam performance for each format type. Therefore, we created a multiple-choice testing change score and a short-answer testing change score, which we used as the dependent variables in two regression models. Although each model used the dependent variable for either multiple-choice or short-answer, both models followed the same ordering for regression analyses. First, gender, aptitude score, and estimated course average were entered into the first block. Grade point average was not included because a large percentage of students failed to provide estimates for their current semester GPA. In addition, first-year or entering freshman reported their high school GPA, while returning college students reported their college GPA. Both high school and college GPA were not considered for inclusion in this block because

of the disparate nature these indicators. The second block was a stepwise inclusion of the 10 LASSI subscales. The second block was exploratory in nature and we had no a priori conclusions about potential ordering for the LASSI subscales.

The results from the regression analyses are broken down by original chapter.

Multiple-regressions for Quiz 1 content. Neither the regression for the multiple-choice change score nor the short-answer change score was significant, F (1, 59) =.36, *ns*, and F (3, 60) = 1.03, *ns*, respectively.

Multiple-regressions for Quiz 2 content. Neither the regression for the multiple-choice change score nor the short-answer change score was significant, F (3, 117) = 1.96, *ns*, and F (3, 117) = .85, *ns*, respectively.

Multiple-regressions for Quiz 3 content. The first regression model, in which the multiple-choice change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was not significant, F (3, 104) = 1.73, *ns*.

The second regression model, in which the short-answer change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, approached significance, F (4, 104) = 2.36, p = .059. The only significant predictor that emerged in this model was the Time Management subscale from the LASSI, t (104) = -2.42, p = .018. The short-answer test change score was expected to decrease .01 units for every point increase in the Time Management subscale.

Multiple-regressions for Quiz 4 content. The first regression model, in which the multiple-choice change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was not significant, F (3, 108) = .24, *ns*.

The second regression model, in which the short-answer change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was approaching significance, F (3, 108) = 2.68, p = .051. This model accounted for 7.1% of the variance in the short-answer test change score. The only significant predictor that emerged in this model was estimated course average, t (108) = -2.02, p = .046. The short-answer test change score was expected to decrease by .79 units when the estimated course average increases by a single point.

Multiple-regressions for Quiz 5 Content. The first regression model, in which the multiple-choice change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was not significant, F (3, 85) = 1.21, *ns*.

The second regression model, in which the short-answer change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was approaching significance, F (3, 85) = 7.33, $p < .001$. This model accounted for 21.1% of the variance in the short-answer test change score. The only significant predictor that emerged in this model was estimated course average, t (85) = -4.23, $p < .001$. The short-answer test change score was expected to decrease by 2.07 units for every single point increase in estimated course average.

Multiple-regressions for Quiz 6 content. The first regression model, in which the multiple-choice change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was significant, F (4, 94) = 10.44, $p < .001$. This model accounted for 31.7% of the variance in the multiple-choice test change score. Average course performance and the Test Strategies subscale of the LASSI both emerged as significant predictors of the multiple-choice change score, t (94) = -.322, p = .002 and t (94) = -.3.07, p = .003, respectively. The short-answer test change score was expected to decrease by 1.08 units when the estimated course performance increases by a single point (when the other predictor was held constant). The short-

answer test change score also was expected to decrease by .02 units when scores on the Test Strategies subscale of the LASSI increases by a single point (when estimated course performance was held constant).

The second regression model, in which the short-answer change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was also significant, F (3,91) = 6.01, p = .001. This model accounted for 16.5% of the variance in the short-answer test change score. The only significant predictor that emerged in this model was estimated course average, t (94) = -3.03, p = .003. The short-answer test change score was expected to decrease by 1.45 units for every single point increase in estimated course average.

Multiple-regressions for Quiz 7 content. The first regression model, in which the multiple-choice change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was significant, F (3, 99) = 4.65, p = .004. This model accounted for 12.7% of the variance in the multiple-choice test change score. The only significant predictor that emerged in this model was estimated course average, t (99) = -2.31, p = .02. Thus, the multiple-choice test change score was expected to decrease by .89 units when estimated course average increases by one point.

The second regression model, in which the short-answer change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was also significant, F (3,96) = 9.37, $p < .001$. This model accounted for 22.7% of the variance in the short-answer test change score. The only significant predictor that emerged in this model was estimated course average, t (99) = -4.36, $p < .001$. The short-answer test change score was expected to decrease by 1.84 units when estimated course average increases by a single point.

Multiple-regressions for Quiz 8 content. The first regression model, in which the multiple-choice change score was regressed on gender, aptitude score, estimated course average, and 10 LASSI subscales, was not significant, F (3, 98) = 2.04, *ns*.

The second regression model, in which the short-answer change score was regressed on gender, aptitude score, estimated grade point average, and 10 LASSI subscales, was significant, F (4, 98) = 3.17, p = .017. This model accounted for 11.9% of the variance in the short-answer test change score. The only significant predictor that emerged in this model was the Selecting Main Ideas subscale from the LASSI, t (98) = -3.13, p = .002. The short-answer test change score was expected to decrease by .021 units for every increase of a subscale point.

Conclusions from multiple-regression models. This study was the first to examine learning and study skills for any possible effects on the testing effect. We conducted regression models for both multiple-choice performance and short-answer performance for each chapter. Although no predictors were constant across all analyses, there were two major findings.

First, the most common predictor of the testing effect (as measured by a testing change score for each format) was estimated course performance. In contrast to previous findings (e.g., Marsh et al., 2009), we found that the magnitude of the testing effect decreases as a function of an increase in estimated course average: Testing effect diminished as a student's overall performance increases. Estimated course performance was the sole predictor in 80% percent of the significant short-answer regression models and accounted for 7.1-22.7% of the variance in these models. Estimated course performance, also, was a significant predictor in both of the significant multiple-choice models. This predictor combined with the Test Strategies subscale of the LASSI accounted for 31.7% of the variance for one chapter and estimated course performance, alone, accounted for 12.7% of variance in another chapter.

Second, contrary to our expectations, there was no consistent and reliable effect for learning and study skills in the regression models. Three different subscale scores emerged as significant predictors for several of the regression models: the Time Management, Selecting Main Ideas, and Test Strategies subscales. A closer examination of the effects of learning and study skills on the testing effect is needed before any firm conclusions can be drawn about the interaction of learning and study skills and repeated testing.

General Discussion

The purpose of the present study was two-fold: First, we investigated the occurrence of the testing effect in an introductory psychology course. This study, along with numerous others (for a review, see Roediger & Karpicke, 2006a), attempted to bridge experimental research in cognitive psychology with pedagogical research on effective strategies to promote retention in the college classroom. Second, this study examined several factors—including item format, academic achievement, aptitude, study time, learning skills, and study strategies—that may influence the prevalence and magnitude of the testing effect. To our knowledge, this study was the first to examine the effect of learning skills and study strategies, and one of only several to examine the effect of academic achievement and aptitude on the testing effect. We discuss the significance of our findings, along with their connection to previous research on the testing effect, are presented below. We also discuss the limitations of our work and potential directions for future research.

The Testing Effect

We predicted that unit exam performance would be higher for items that were quizzed when compared to novel items on the unit exam: Performance for questions that were assessed two times (on the initial quiz and the unit exam) would be higher than for questions that were assessed only once (on the unit exam). This general finding held across six of the eight chapters. In these instances, there was no difference between items that had been initially assessed on a quiz and the read-only condition. However, each instance resulted in superior performance when

compared to novel items on the unit exam. This finding suggests that prior exposure, either through quizzing or repeated study, was beneficial on the unit exam. Most likely, there was no difference between quizzing and studying on the unit exam because students were engaged in both activities in preparation for the upcoming exam. Therefore, students took the quizzes as part of the course, but also studied course material independently. The absence of significant differences between the read-only and quiz conditions may be the result of our applied experimental design. That any prior exposure leads to significantly better performance than items to which students were not previously exposed, suggests students may benefit from repeated presentations of items across the semester. This repeated exposure may assist students in selecting key concepts or ideas that are central to the course, or give students insight about the expectations of the professor.

In addition, the spacing of assessments across the semester may also influence the retention of information. In a meta-analysis concerning the impact of distributed practice on learning, Cepeda, Pashler, Vul, Wixted, and Rohrer (2006) concluded that lags between learning sessions promote better long-term retention. Although the optimal delay between learning sessions has not yet been defined, Cepeda et al. suggested that longer lags (from 1 day to several months) are often better than shorter lags (less than 1 day). In the present study, the temporal spacing of quizzes and exams (and the cumulative nature of these assessments) served as distributed practice across the semester. Other researchers have shown that spacing is also important when delivering feedback about performance (Smith & Kimball, 2010). The present study, however, did not directly exam the impact of feedback.

Second, we hypothesized that performance for items repeated across an initial quiz and a unit exam would be better than for novel items on the cumulative exam. It was predicted that

performance for questions that were assessed three times (on the initial quiz, the unit exam, and the cumulative exam) would be higher than for questions that were assessed only once (on the cumulative exam). This general finding held for seven of the eight chapters. Across the chapters, students fared better when items were initially assessed on a quiz and then repeated on both the unit and cumulative exam than when items were only assessed on the cumulative exam. However, there was no significant change in performance from unit exam to the cumulative exam for these repeated items, which suggests that performance remained largely the same after the initial quiz.

Finally, we predicted that performance for items that were first administered on the unit exam and subsequently repeated on the cumulative exam would be higher than for items that were first administered on the cumulative exam: Performance for questions that were assessed two times (on the unit exam and the cumulative exam) would be higher than for questions that were assessed only once (on the cumulative exam). This general finding held for five of the eight chapters. For the remainder, there was no significant difference between the conditions. Therefore, our data only partially supported this prediction.

Generally, we supported our three predictions for an enhancement of performance due to repeated testing. Our finding that performance is better on repeated items (specifically, from the quiz to the unit exam or exam performance when compared to novel items) partially supports other researchers. McDaniel et al. (2007) also found an increase in exam performance as a function of classroom quizzing. Although there were some methodological differences between the present study and McDaniel et al., they also concluded that students performed better on items that were repeated from initial quizzes to unit and cumulative exams when compared to novel items appearing on the unit or cumulative exam. Our conclusions were also similar to that

of Butler and Roediger (2008) who found that an initial test led to superior performance when compared to a re-study condition or a control condition where participants did not engage in either intervening test or study trials. We found an advantage for taking intervening tests over certain items when compared to taking only a final test over some items (our novel conditions).

However, our results contrast previous findings in one important way: We found that prior exposure, not necessarily prior testing, resulted in enhanced performance on later assessments. Therefore, studying items (through study items or summary statements) was as equally effective as taking an initial quiz over the material. This contrasts Butler and Roediger (2007) who found an advantage for short-answer tests, but no relative advantage of multiple-choice over a study condition. We address our finding regarding item format in the next section, however, we did find that there was no significant difference in performance for items that were studied when compared to items that were initially quizzed as either short-answer or multiple-choice.

There are two plausible explanations for the absence of a significant difference between study and initial test conditions: First, there was probably considerable overlap in our test and study conditions. We believe that because chapter content is often interrelated, students who studied the course material (independent of our study conditions) had exposure to items from both our study and test conditions. This unrestricted access to chapter material covered in both the study and test conditions may have eliminated any possible effects of testing. Second, we agree with Butler and Roediger (2007) who argued that the study condition used (in both their experiment and ours) was quite artificial. Presenting to-be-tested information, specifically the exact or similar questions, to students probably does not accurately reflect current practices in higher education.

Item Format

We predicted that multiple-choice performance would be significantly higher than short-answer performance on quizzes, but that this pattern would reverse for unit and cumulative exams. For six of the eight chapters, performance for multiple-choice quiz questions was higher than that for short-answer quiz questions. The initial advantage for multiple-choice over short-answer has been found in several studies examining the testing effect (e.g., McDaniel et al., 2007; Wheeler et al., 2003). This advantage reflects the increased difficulty of recall questions compared to recognition questions. Interestingly, McDaniel, et al. (2007) and Wheeler et al. (2003) found a reversal in performance for subsequent assessments. Although students performed better on initial multiple-choice tests, student performance was significantly better on repeated short-answer items when compared to repeated multiple-choice items. Based on these findings, and those Bjork (1975), we expected that short-answer quiz performance would be lower, but later performance on these items would be enhanced (relative to multiple-choice) because of the effortful retrieval that takes place for these items. However, we did not find a significant advantage of short-answer over multiple-choice on either the unit or cumulative exams for any of the eight chapters.

Also, there was neither an advantage of multiple-choice nor short-answer quizzing over engaging in study (through read-only items) for to-be-tested information on unit or cumulative exams. This finding is consistent with the findings discussed in the previous section: Prior exposure, not necessarily prior testing, was key for enhanced performance relative to items for which there was no prior exposure (either through study or testing). Therefore, we failed to support our hypotheses for an enhancement of performance for short-answer items.

Our findings are in contrast to earlier researchers who have found an advantage for short-answer testing (e.g., Butler and Roediger, 2007; Kang et al., 2007; McDaniel et al., 2008). In a comprehensive set of experiments examining test format and test feedback, Kang et al. found that short-answer testing promoted better long-term retention than either multiple-choice testing or restudying. Although Butler and Roediger (2007) and McDaniel et al. (2008) found no significant difference between initial multiple-choice testing and restudying, Kang et al. (2007) found that initial multiple-choice testing benefitted retention better and restudying. We, however, found no significant advantage of either initial item format when compared to studying. As we discussed in the previous section, we believe that overlap between our experimental investigation of repeated testing and student preparation for assessments (independent of our experiment) may be partially responsible for our failure to find an effect of format.

In addition to our analyses of proportion correct for each item format, we conducted conditional analyses examining unit and cumulative exam performance for quiz items that were either answered correctly or incorrectly. McDaniel et al. (2007) concluded the testing effect was present for missed short-answer items when using similar conditional analyses. For the majority of the chapters, we found a significant effect of initial accuracy. Participants' performance on repeated items was higher if the participant initially answered the item correctly than when compared to answering it incorrectly initially. Our findings, unlike those of McDaniel et al. (2007), did not find a significant effect of format or an advantage for short-answer questions.

Study Time

When study time was controlled across analyses, our results on the testing effect and item format remained unchanged. Participants performed significantly better on items that were repeated from initial quizzes to exams relative to novel information on the exams. There was no

differential effect for items that were quizzed or study items, and no differential effects of format (multiple-choice and short-answer) when compared with restudying. There are two plausible explanations for this finding: First, student self-report may not have been an adequate method for measuring study duration. Estimates of study time were retrospective, because students completed the estimates after completing each assessment. These estimates may have been affected by any number of factors, including an inability to accurately calculate hours spent studying or the desire to provide socially appropriate responses. In future studies, we may be able to obtain a more accurate estimate of study time by having students log study hours throughout the semester. Second, study time may have had similar effects across assessments and formats because of the interrelated nature of chapter content. Therefore, studying may have benefited the student in similar ways regardless of the anticipated assessment (quiz or exam) or the format of the items on that assessment.

Aptitude

We hypothesized that students with higher aptitude scores would benefit more from repeated testing than students with lower aptitude scores. To test our prediction, we performed a series of regressions for the change in proportion correct from the initial quiz to the cumulative exam. Aptitude score did not emerge as a significant predictor across any of these analyses. Therefore, we failed to support this hypothesis.

To our knowledge, our study was only the second to examine the relationship between aptitude and the testing effect. The first study, conducted by Marsh et al. (2009), found that higher aptitude students showed a greater effect of testing when compared to lower aptitude students. Although our results contrast those of Marsh et al. (2009), there is one important difference between the studies which make direct comparisons difficult. Marsh et al. used an

estimated aptitude score to predict gains in repeated presentations of SAT II questions. However, our study used actual aptitude scores to predict gains in repeated presentations of material from an introductory psychology course. Therefore, aptitude performance may relate more directly to aptitude-type testing than performance on Introductory Psychology assessments.

Academic achievement

We predicted students who had higher levels of academic achievement would show greater benefit of testing than students who had lower levels of academic achievement. To test our prediction, we performed a series of regressions for the change in proportion correct from the initial quiz to the cumulative exam. In contrast to our predictions, we found that, in general, higher academic performance (as measured by estimated course average) was associated with a decrease in the proportion correct from the initial quiz to the cumulative exam. Therefore, lower performing students may actually benefit more from repeated testing than higher performing students.

Our finding contradicts the most recent finding by Marsh et al. (2009) who found the opposite effect. In their study, participants who scored higher overall (on the final criterion test) showed greater benefit of repeated testing when compared to participants who scored lower overall. Again, attempts to directly compare our study with that of Marsh et al. (2009) are compounded by the differences between both studies, including differences in the content of the assessments. Our results, however, are more similar to those of Graham (1999), who examined the impact of unannounced and announced quizzes on exam performance. Although there was no significant effect for announced quizzes, Graham concluded that unannounced quizzes had a significant benefit for lower performing students, not higher performing ones.

Our findings that lower performing students benefitted more from repeated testing when compared to higher performing students may be a product of overall performance. One plausible explanation for this finding is that higher performing students may be constricted by a ceiling effect, while lower performing students have a larger margin for improvement. Additional studies are needed to fully understand the impact of academic achievement. More valid measures of academic achievement or repeated assessment of academic achievement may provide better indices for future comparisons.

Learning and Study Skills

Our study was the first to examine the effect of learning and study skills on the testing effect. We predicted that students who had developed better learning and study skills would benefit more from repeated testing than individuals with less developed learning and study skills. To test our prediction, we performed a series of regressions for the change in proportion correct from the initial quiz to the cumulative exam with learning and study skills as one of several predictors. For the majority of analyses, learning and study skills (as measured by LASSI subscales) did not emerge as significant predictors of change in the proportion correct. For the analyses where a LASSI subscale emerged as a significant predictor for change in proportion correct, there was no consistent pattern. Three different scales emerged significant in three separate regression analyses. These results, taken together, suggest that learning and study skills may not play an influential role in the testing effect. However, additional studies on learning and study skills variables are needed before a definitive conclusion can be reached.

Limitations of the Present Study

The present study represents an innovative first step in translating cognitive research into pedagogical practice. Due to the applied nature of this work, there are several limitations that

deserve some attention. First, we had a considerable amount of missing data across our participants. Although we believe this did not affect our general conclusions, it did limit our ability to perform some analyses and led us to segment our data in a way we had not anticipated. Second, inherent in our applied work was a lack of experimental control, which is readily apparent when attempting to compare quiz conditions to study conditions. We could not control study and test time across conditions because it would have altered the nature of the course and potentially hindered educational progress. Finally, some of our measures, including self-report for study time and estimated grade point average, may not have accurately captured those dimensions we intended to measure. Previous research (e.g., Glenberg, et al., 1987) found that students were not proficient in providing accurate judgments related of their own academic performance. We believe that students may also be inaccurate with retrospective self-report of study time, as well. In future studies, more accurate measures, such as official transcripts for grade point average or having students keep logs for study hours, may provide better estimates for these constructs.

Conclusions

Our study is one of the most recent investigations of the testing effect in higher education. Although many of our results contradicted previous findings (e.g., Roediger & Karpicke, 2006a), we believe that this work is an important step bridging cognitive psychology and educational practice. Our work shows some of the clear pedagogical benefits for frequent quizzing and testing. Students performed better on items to which they were exposed to them several times throughout the semester than items that are merely presented a single time.

In the opening paragraph of this paper, we quoted Roediger and Karpicke (2006a) who called for an in-depth research on the effects of repeated testing and its application in education.

We believe our study represents an important first step in this process. It is our hope that educators can implement research-based strategies, such as repeated and frequent testing, to promote long-term retention of learning.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger III, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861-876.

American College Test. (2009). *Concordance between ACT composite score and sum of SAT critical reading and mathematics scores*. Retrieved July 24, 2009 from ACT ACT-SAT Concordance. Web site: http://www.act.org/aap/concordance/index.html

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85,* 89-99.

Bjork, R.A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R.L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp.123–144). Hillsdale, NJ: Erlbaum.

Butler, A. C., & Roediger, III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514-527.

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

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

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354-380.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438-481.

Dustin, D. S. (1971). Some effects of exam frequency. *Psychological Record, 21,* 409-414.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119-136.

Graham, R. B. (1999). Unannounced quizzes raise test scores selectively for mid-range students. *Teaching of Psychology, 26,* 271-273.

Grover, C. A., Becker, A. H., & Davis, S. F. (1989). Chapters and units: Frequent versus infrequent testing revisited. *Teaching of Psychology, 16,* 192-194.

Jones, H. E. (1923). Experimental studies of college teaching. *Archives of Psychology, 10,* 1-70.

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

Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology, 25,* 427-436.

Kling, N., McCorkle, D., Miller, C., & Reardon, J. (2005). The impact of testing frequency on student performance in a marketing course. *Journal of Education for Business, 81,* 67-72.

Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2008). Test-enhanced learning in medical education. *Medical Education, 42,* 959-966.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*, 210-212.

Marsh, E. J., Agarwal, P. K., & Roediger III, H. L. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied, 15*, 1-11.

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., et al. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 34*, 51-57.

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

McDaniel, M. A., Roediger III, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200-206.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74*, 18-22.

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect of true-false examination questions. *Journal of Educational Psychology, 17*, 52-56.

Richland, L. E., Linn, M. C., & Bjork, R. A. (2007). Cognition and instruction: Bridging laboratory and classroom settings. In F. T. Durso (Ed.), *Handbook of applied cognition (2nd ed.).* New Jersey: John Wiley & Sons.

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

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

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 233-239.

Seidel, L. F., Benassi, V. A., & Lewis, J. B. (2008). The testing effect: Teaching to enhance learning in health administration education. *Journal of Health Administration Education,* 63-72.

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 80-95.

Sones, A. M., & Stroud, J. B. (1940). Review, with special reference to temporal position. *Journal of Educational Psychology, 31*, 665-676.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*, 641-656.

Spunzar, K. K., McDermott, K. B., & Roediger, III, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition, 35*, 1007-1013.

Sumowski, J. F., Chiaravalloti, N., & DeLuca, J. (2010). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. *Neuropsychology, 24*, 267-272.

Svinicki, M., & McKeachie, W. J. (2010). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers (13th ed.).* Wadsworth Publishing: Belmont, CA.

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true-false examinations. *Journal of Educational Research, 83*, 119-124.

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Psychology, 86,* 357-362.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175-184.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571-580.

Appendix A

The naturally occurring schedule for assessments in this study.

Week	Scheduled Assessment Type	Course Content
2	Reading Quiz	Research Methods
4	Reading Quiz	Psychological Theories of Learning
5	Unit Exam	Research Methods and Learning
6	Reading Quiz	Memory
8	Reading Quiz	Neuroscience and Behavior
9	Unit Exam	Memory and Neuroscience and Behavior
	Cumulative Midterm Exam	Research Methods, Learning, Memory, Neuroscience and Behavior
11	Reading Quiz	Stress and Health
12	Reading Quiz	Social Psychology
	Unit Exam	Stress and Health and Social Psychology
14	Reading Quiz	Psychological Disorders
15	Reading Quiz	Treatment of Psychological Disorders
	Unit Exam	Psychological Disorders and Treatment of Psychological Disorders
16	Cumulative Final Exam	Stress and Health, Social Psychology, Psychological Disorders, Treatment

Appendix B

Demographic Questionnaire

1. What is your gender?
    a. Female
    b. Male

2. What is your class standing?
    a. Freshman (If selected, please answer questions 3 and 4.)
    b. Sophomore (If selected, please skip to question 5.)
    c. Junior (If selected, please skip to question 5.)
    d. Senior (If selected, please skip to question 5.)
    e. Other (If selected, please skip to question 5.)

3. Please choose the statement that best describes you:
    a. This is my first semester in college
    b. I am a freshman, but this is not my first semester in college.

4. What was your estimated high school GPA (grade point average) when applying to college? Remember, your grade point average ranges on a scale from 0.0 to 4.0.
    _____

5. What is your estimated college GPA (grade point average)? _____

6. What was your ACT or SAT scores when applying to college (Please provide your highest score for either the ACT, SAT, or both.)
    a. ACT _____
    b. SAT _____

7. Overall, how would you rate your academic ability?
    a. Very strong
    b. Good
    c. Okay
    d. Fair
    e. Poor

Appendix C

Learning and Study Strategies Inventory

The Learning and Study Strategies Inventory (LASSI) contains 80 statements related to how you learn and study. Please read each statement and select a response according to the following key:

Sample Question:

99. I would like to learn more about myself by taking inventories like this.

Not typical of me - does not necessarily mean that the statement would never describe you, but it would be true of you only in rare circumstances.

Not very typical of me - means that the statement generally would not be true of you.

Somewhat typical of me - means tahat the statement would be true of you about half the time.

Fairly typical of me - means that the statement would generally be true of you.

Very much typical of me - does not necessarily mean that the statement would always describe you, but that it would be true of you almost all the time.

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I concentrate fully when studying.		○	○	○	○	○
I am unable to summarize what I have just heard in a lecture or read in a textbook.		○	○	○	○	○
I try to find relationships between what I am learning and what I already know.		○	○	○	○	○
I find it hard to stick to a study schedule.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
In taking tests, writing papers, etc., I find I have misunderstood what was wanted and lose points because of it.		○	○	○	○	○
I am able to study subjects I do not find interesting.		○	○	○	○	○
When I decide to study, I set aside a specific length of time and stick to it.		○	○	○	○	○
Because I don't listen carefully, I don't understand some course material.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I try to identify potential test questions when reviewing my class material.		○	○	○	○	○
During class discussions, I have trouble figuring out what is important enough to put in my notes.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
To help me remember new principles we are learning in class, I practice applying them.		○	○	○	○	○
My underlining is helpful when I review text material.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
When it comes to studying, procrastination is a problem for me.		○	○	○	○	○
I set high standards for myself in school.		○	○	○	○	○
When I am studying a topic, I try to make everything fit together logically.		○	○	○	○	○
I find it difficult to maintain my concentration while doing my course work.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I only study the subjects I like.		○	○	○	○	○
When preparing for an exam, I create questions that I think might be included.		○	○	○	○	○
When I take a test, I realize I have studied the wrong material.		○	○	○	○	○
If there is a web site for my textbook, I use the information provided there to help me learn the material.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I have difficulty identifying the important points in my reading.		○	○	○	○	○
When work is difficult, I either give up or study on the easy parts.		○	○	○	○	○
To help me learn the material presented in my classes, I relate it to my own general knowledge.		○	○	○	○	○
There are so many details in my textbooks that it is difficult for me to find the main ideas.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I review my notes before the next class.		○	○	○	○	○
I have difficulty adapting my studying to different types of courses.		○	○	○	○	○
I translate what I am studying into my own words.		○	○	○	○	○
I put off studying more than I should.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I get discouraged because of low grades.		○	○	○	○	○
Even if I am having difficulty in a course, I can motivate myself to complete the work.		○	○	○	○	○
I spread out my study times so I do not have to "cram" for a test.		○	○	○	○	○
My mind wanders a lot when I study.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I stop periodically while reading and mentally go over or review what was said.		○	○	○	○	○
I go to the college learning center for help when I am having difficulty learning the material in a course.		○	○	○	○	○
I feel very panicky when I take an important test.		○	○	○	○	○
I have a positive attitude about attending my classes.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I test myself to see if I understand what I am studying.		○	○	○	○	○
When I study for a test, I have trouble figuring out just what to do to learn the material.		○	○	○	○	○
Even if I do not like an assignment, I am able to get myself to work on it.		○	○	○	○	○
When they are available, I attend review sessions for my classes.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I would rather not be in school.		○	○	○	○	○
I set goals for the grades I want to get in my classes.		○	○	○	○	○
When I am taking a test, worrying about doing poorly interferes with my concentration.		○	○	○	○	○
I try to see how what I am studying would apply to my everyday life.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I have trouble understanding exactly what a test question is asking.		○	○	○	○	○
I worry that I will flunk out of school.		○	○	○	○	○
To help make sure I understand the material, I review my notes before the next class.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I do not care about getting a general education, I just want to get a good job.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I find it hard to pay attention during lectures.		○	○	○	○	○
I try to relate what I am studying to my own experiences.		○	○	○	○	○
I dislike most of the work in my classes.		○	○	○	○	○
I review my answers during essay tests to make sure I have made and supported my main points.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
When studying, I seem to get lost in the details and miss the important information.		○	○	○	○	○
I use special study helps, such as italics and headings, that are in my textbook.		○	○	○	○	○
I am very easily distracted from my studies.		○	○	○	○	○
Even when I don't like a course, I work hard to get a good grade.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
It is hard for me to decide what is important to underline in a text.		○	○	○	○	○
To help me learn the material, I complete at least some of the practice problems in my textbooks.		○	○	○	○	○
I do not have enough time to study because I spend too much time with my friends.		○	○	○	○	○
To check my understanding of the material in a course, I make up possible test questions and try to answer them.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
Even when I am well prepared for a test, I feel very anxious.		○	○	○	○	○
I set aside more time to study the subjects that are difficult for me.		○	○	○	○	○
I do poorly on tests because I find it hard to plan my work within a short period of time.		○	○	○	○	○
During a demonstration in class, I can identify the important information I need to remember.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
I am up-to-date in my class assignments.		○	○	○	○	○
When I am having trouble with my coursework, I do not go to the instructor for help.		○	○	○	○	○
I end up "cramming" for every test.		○	○	○	○	○
When I listen to class lectures, I am able to pick out the important information.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
When I am studying, worrying about doing poorly in a course interferes with my concentration.		○	○	○	○	○
I do not care if I finish college as long as I have a good time.		○	○	○	○	○
I try to find a study partner or study group for each of my classes.		○	○	○	○	○
Courses in certain subjects, such as math, science, or a foreign language, make me anxious.		○	○	○	○	○

		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
When completing a problem-solving task, it is difficult for me to pick out the important information.		○	○	○	○	○
After a class, I review my notes to help me understand the information that was presented.		○	○	○	○	○
If I get distracted during class, I am able to refocus my attention.		○	○	○	○	○
In my opinion, what is taught in my courses is not worth learning.		○	○	○	○	○
		Not at all typical	Not very typical	Somewhat typical	Fairly typical	Very much typical
If I am having trouble studying, I ask another study or the instructor for help.		○	○	○	○	○
I get so nervous and confused when taking an examination that I fail to answer questions to the best of my ability.		○	○	○	○	○
I find that during lectures I think of other things and don't really listen to what is being said.		○	○	○	○	○
Even when study materials are dull and uninteresting, I manage to keep working until I finish.		○	○	○	○	○

Appendix D

Sample Multiple Choice Items

1. When every member of a population has an equal chance of being included in a sample, what sampling process is being used?
    a. Reliability sampling
    b. Random assignment
    c. Random sampling
    d. Convenience sampling
2. A person has suffered damage to the region of the brain responsible for regulating body temperature, hunger, thirst, and sexual behavior. Which region of the brain was injured?
    a. Thalamus
    b. Cerebral cortex
    c. Hypothalamus
    d. Hippocampus

Sample Short-Answer Items

1. When every member of a population has an equal chance of being included in a sample, what sampling process is being used?_____
2. A person has suffered damage to the region of the brain responsible for regulating body temperature, hunger, thirst, and sexual behavior. Which region of the brain was injured?

    _____

Sample Study Items

1. When every member of a population has an equal chance of being included in a sample, a random sampling process is being used.

2. A person has suffered damage to the region of the brain responsible for regulating body temperature, hunger, thirst, and sexual behavior. The region of the brain that was injured was the hypothalamus.

Appendix E

Summary Table of Significant Effects

This table is provided as a summary for the significant results for proportion correct for each item type. Each block that contains text describes a significant effect. Non-significant effects were omitted from this table. Gray cells are not applicable to a given assessment.

	Performance				
	MC	SA	Study	UE Novel	CE Novel
Quiz 1	Higher than SA	Lower than MC			
Unit Exam 1		Increased from Q1			
Cumulative Exam 1		Higher than CE novel		Higher than CE novel	Lower than SA and UE novel

Quiz 2	Higher than SA	Lower than MC			
Unit Exam 1		Increased from Q2			
Cumulative Exam 1	Higher than CE novel	Higher than CE novel	Higher than CE novel	Higher than CE novel	Lower than MC, SA, Study, UE novel

	Performance				
	MC	SA	Study	UE Novel	CE Novel
Quiz 3	Higher than SA	Lower than MC			
Unit Exam 2	Higher than UE novel	Increased from Q3; Higher than UE novel	Higher than UE novel	Lower than MC, SA, Study	
Cumulative Exam 1	Higher than UE novel, CE novel	Higher than UE novel, CE novel	Higher than UE novel, CE novel		Lower than MC, SA, Study
Quiz 4					
Unit Exam 2	Increased from Q4; Higher than UE novel	Increased from Q4; Higher than UE novel	Higher than UE novel	Lower than MC, SA, Study	
Cumulative Exam 1	Higher than UE novel, CE novel	Higher than UE novel, CE novel	Higher than UE novel, CE novel	Lower than MC, SA, Study	Lower than MC, SA, Study
Quiz 5	Higher than SA	Lower than MC			
Unit Exam 3	Higher than UE novel	Increased from Q5; Higher than UE novel	Higher than UE novel	Lower than MC, SA, Study	
Cumulative Exam 2	Higher than UE novel, CE novel	Higher than CE novel	Higher than CE novel	Lower than MC; Higher than CE novel	Lower than MC, SA, Study, UE novel

	Performance				
	MC	SA	Study	UE Novel	CE Novel
Quiz 6	Higher than SA	Lower than MC			
Unit Exam 3	Higher than UE novel	Increased from Q6; Higher than UE novel	Higher than UE novel	Lower than MC, SA, Study	
Cumulative Exam 2	Higher than CE novel	Higher than CE novel	Higher than CE novel	Higher than CE novel	Lower than MC, SA, Study, UE novel
Quiz 7					
Unit Exam 4	Higher than UE novel	Increased from Q7; Higher than UE novel	Higher than UE novel	Lower than MC, SA, Study	
Cumulative Exam 2	Higher than UE novel, CE novel	Higher than UE novel, CE novel	Higher than UE novel, CE novel	Increased from UE4; Lower than MC, SA, Study	Lower than MC, SA, Study
Quiz 8	Higher than SA	Lower than MC			
Unit Exam 4	Higher than UE novel	Increased from Q8; Higher than UE novel	Higher than UE novel	Lower than MC, SA, Study	
Cumulative Exam 2	Higher than UE novel, CE novel	Higher than UE novel, CE novel	Higher than UE novel, CE novel	Increased from UE4; Lower than MC, SA, Study	Lower than MC, SA, Study

Note. Q = quiz. UE = unit exam. CE = cumulative exam. MC = multiple-choice. SA = short-answer. Study = study items.