DISCOVERY AND ADOPTION OF THE TESTING EFFECT: CHALLENGES OF ELICITING SELF-TESTING BEHAVIOR IN STUDENTS

by

Adam Patrick Young

B.A. Philosophy and Psychology, University of Massachusetts, 2011

M.A. Psychology – Research, University of Massachusetts, 2013

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado Boulder

In partial fulfillment of the requirements for the degree of

M.A. Cognitive Psychology

Department of Psychology and Neuroscience

2015

This thesis entitled:

The Discovery and Adoption of the Testing Effect:

Challenges of Eliciting Self-Testing Behavior in Students

Written by Adam P. Young

Has been approved for the Department of Psychology

Dr. Alice F. Healy_____

Dr. Matt Jones_____

Dr. Tim Curran_____

Date_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB Protocol #: 12-1465

Young, Adam Patrick

Discovery and Adoption of the Testing Effect:

Challenges of Eliciting Self-Testing Behavior in Students

Thesis directed by College Professor of Distinction Dr. Alice F. Healy

Engaging in self-testing is often a more effective strategy for retaining information than is restudying the target material. This testing benefit has been found in numerous studies and across a wide variety of manipulations and materials. Despite these common results, students remain largely unaware of the benefits of self-testing. We conducted three experiments using ecologically valid materials (statistics concepts and chemical elements) to determine whether student subjects could be made aware of the benefits of self-testing and consequently engage in voluntary self-testing. A within-subject testing benefit was successfully replicated in two experiments. However, little support was found for the hypothesis that subjects could be brought to exhibit voluntary, covert self-testing behavior. These results present a challenge to the field regarding effective means of eliciting appreciation and self-regulated use of the testing effect among students and thereby improve learning outcomes.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation of my committee members, Drs. Alice Healy, Matt Jones and Tim Curran for the challenges they have provided me; without challenge there is no growth, and without them there is no thesis. Thanks to my supportive family for making me the person I am today. Finally, thanks to my loving fiancée Monica, for believing in me.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

Introduction and Literature Review

Previous Research

In 1909 Edwina Abbott studied the effects of recall on subsequent memory and found that by engaging in the retrieval of previously studied information, subjects exhibited marked improvement in their memory for the target stimuli. Hundreds of studies have since been conducted on this phenomenon, which has come to be known as the *testing effect* (Roediger & Karpicke, 2006a). Behavioral (Rowland, 2014), neuropsychological (Keresztes, Kaiser, Kovacs, & Racsmány, 2015) and neurobiological (Rudy, 2014; Sara, 2000) research has converged on the view that these acts of *retrieval practice* constitute reconsolidation events that solidify the learner's memory of the retrieved information.

The beneficial effects of retrieval have recently been leveraged by the educational community seeking to enhance student learning outcomes. No longer administering tests simply as a means of assessment, the trend of using tests as learning opportunities has seen growing popularity with the emergence of educational technologies such as the iClicker in the classroom. This remote student response device is used to present questions to the class, requiring members of the audience to submit a response of their own to each question rather than listen passively as in traditional lectures. Previous research has demonstrated that the mnemonic benefits of using iClickers in the classroom (Shapiro & Gordon, 2012) are reliable and not reducible to various confounds such as heightened attention given to information presented in these contexts. Similarly, self-tests (acts of retrieval not formally assigned but instead the product of a learner's

self-regulated study activity) can also heighten the contribution of the study session to exam[1] preparation.

The testing effect has received much research attention in recent years. The most widely cited study on the testing effect was conducted by Roediger and Karpicke (2006b). In the primary experiment of their article, they asked a sample of college students to read two English prose passages ("The Sun" and "Sea Otters") for seven minutes each. Students were then asked either to restudy one of the passages or to take a recall test, writing down as much of the selected passage as possible. After either a five-min, two-day, or one-week retention interval, they were then asked to take a recall exam. Performance on this final exam indicated that retention for restudied material was superior to retention for tested material at the five-min retention interval by six percent, yet performance at both longer delays was superior by 14 percent. These results are not unique. A recent meta-analysis of 159 articles found an average testing benefit of .5 standard deviations in performance (Rowland, 2014) across a wide variety of manipulations. It should be noted, however, that there is some inconsistency regarding whether restudy benefits performance over short delays more so than does self-testing. This restudy advantage has been replicated (e.g., Kornell, Bjork, & Garcia, 2011; Tullis, Finley, & Benjamin, 2013), although enough studies have found the opposite result that the meta-analysis reports a testing advantage of .4 *sd* for delays categorized as shorter than one day (and .7 *sd* for delays greater than one day).

Despite the established research supporting the benefits of retrieval practice, studies have shown that those who stand to gain from the testing effect are largely unaware of its benefits. A study by Karpicke, Butler, and Roediger (2009) found that 84% of college students rated

[1] For clarity, I will refer to tests conducted for the sake of learning as *tests* and tests conducted for the sake of assessment (e.g., in a classroom or experiment) as *exams*.

restudying information as a highly important exam preparation strategy and 55% rated restudying as the most important strategy. Meanwhile, 11% of students rated self-testing as an important exam preparation strategy, with only 1% rating it as the most important strategy. Students in a different experiment also rated repeatedly studied information as more memorable than information initially studied and then tested upon, despite actually performing better to the extent that they were tested (Roediger & Karpicke, 2006a). The disparity between the findings from the research and the strategies employed by those who stand to gain from the testing effect is striking.

One explanation for the absence of appreciation of retrieval practice among students is a lack of experience comparing the effectiveness of studying and self-testing as exam preparation strategies. As a consequence, one potential method of eliciting awareness and engagement in retrieval practice among students is to allow them to directly experience the mnemonic benefits of retrieval practice by comparing their performance on tested and non-tested information. Following an experience in which performance clearly indicates the superiority of retrieval practice to restudy, one might expect a learner to choose to engage in retrieval practice in a subsequent exam preparation session. Meanwhile, learners not directly exposed to the benefits of retrieval practice would be unexpected to exhibit this realization and adoption of the testing benefit.

This rationale is justified in part by the success of a similar study in the context of the generation effect, conducted by deWinstanley and Bjork (2004). In this study. participants were asked to read several phrases with various words highlighted. These words were either incomplete (e.g., "*aff_ct_v_*") or complete (e.g., "*affective*") and distinguished from the rest of the phrases only by highlighting. After reading the phrases, participants completed a fill-in-the-

blank test, requiring retrieval of the previously highlighted words. They then read a second group of phrases similar in structure to the first group of phrases but novel in content, followed by a second fill in the blank test. Performance on the first test revealed a significant memory advantage for the incomplete words, which required generation, relative to the complete words, which required only reading; the generation effect had been observed. The striking finding was that performance on the second test showed a complete elimination of the generation advantage. The authors replicated their results, and concluded that due to exposure to both strategies (i.e., read and generate) and the apparent benefits of generation, participants changed their processing of to-be-read items to reduce the performance discrepancy on the second group of phrases. Without exposure to both conditions, participants did not change their processing of to-be-read items and continued to exhibit a generation benefit. The study by deWinstanley and Bjork (2004) demonstrates that experiencing multiple strategies with differing levels of effectiveness provides learners with the chance to discover the superior strategy and use it during later opportunities.

The Present Study

The findings by deWinstanley and Bjork (2004) have motivated the present investigation to determine whether the testing effect may be discovered and adopted in a manner similar to what was observed in the context of the generation effect. By allowing learners to experience the benefits of retrieval practice through direct, comparative experience, it is expected that these learners will come to appreciate the testing effect and adopt retrieval practice as a dominant strategy in subsequent exam preparation. The present work reports the results of a series of potential remedies for the divergence between the use of strategies that aid student learning outcomes and the strategies those students most frequently employ.

This investigation entails three experiments: A) a classroom quasi-experiment to determine whether direct instruction of the benefits of self-testing leads to improved learning outcomes, B) a laboratory experiment using statistics and research methods concepts with an emphasis on ecological validity, and C) a laboratory experiment using chemistry elements with an emphasis on simplicity of materials and internal validity.

Chapter II

Experiment 1: Classroom Instruction

The question may rightfully be asked why not simply tell students about the testing effect, that it leads to better learning outcomes, and determine whether these direct instructions lead to self-testing adoption? If this direct instruction method is effective at engendering use of the testing effect, discovery of the testing effect through comparative experience may be an unnecessary endeavor. To this end, a classroom quasi-experiment was conducted to test the effectiveness of this strategy of raising appreciation of the testing effect.

Method

Participants. Sixty-eight undergraduate students enrolled in a University of Colorado Boulder course on statistics and research methods in psychology provided consent and complete data. Students were enrolled in one of four weekly lab sections (distinct from the lectures). Two of these labs were chosen as experimental groups and two chosen as control groups by equating the labs as closely as possible in terms of class time. The groups exhibited no substantial differences with regards to academic year, major, or mathematical ability. The four labs were taught by two teaching assistants, each instructing one control group and one experimental group.

Materials. The instructions provided to students, explained in detail in the procedure below, were accompanied by PowerPoint slides provided in Figure 1**.**

Procedure. In the third week of the semester, the teaching assistant of each lab presented the PowerPoint presentation, which emphasized the importance of self-testing to the lab sections

composing the experimental group. The TA also informed the experimental group that it would be wise to use the practice tests that the professor makes available as a way of self-testing. In the control group's labs, the teaching assistant showed no comparable PowerPoint, but instead gave general tips for succeeding in class – reading the chapters, doing the homework, and also taking the practice tests. This last detail was included so all labs would be aware of the practice tests availability, but the control group did not receive special emphasis on these practice tests whereas this was the case for the experimental group.

In the week prior to each of the class's exams, the TAs reminded the experimental group of the importance of self-testing and recommended taking the practice test. The control group was made aware of the practice test's availability, but it was not explicitly recommended as a means of self-testing, nor as a preferable strategy for exam preparation. These reminders preceded all three class exams.

In the week following each exam, the professor used iClickers to survey the student audience during lecture by asking students to press a button on their iClicker corresponding to one of the options shown at the front of the classroom. Questions listed in Table 1 were asked on each occasion, and with the exception of the first question, the response alternatives were always "Yes", "No", and "Somewhat". The first question in Table 1, "While preparing for the last test, which best describes how you made use of the practice test?" was accompanied by the following response alternatives: A) "I went through and answered questions like I would on a real test," B) "I used it like a study guide to identify important concepts and skills for later study," C) "I tried to memorize the answers that were provided," and D) "I prepared in other ways, without using the practice tests." Each student's response was recorded and analyzed for differences between the experimental and control groups. Additionally, exam grades, final homework grades, and

final course grades were also analyzed for group differences. It was hypothesized that the experimental group would affirm use of the practice tests as real tests with greater frequency than would the control group, that the experimental group would self-report greater use of the practice tests in general, and that the experimental group would exhibit higher exam, homework, and course grades.

Results

To analyze responses of iClicker frequencies, a series of chi square tests of independence was conducted. Descriptive statistics of response rates are provided in Table 1. In all instances, the chi square test revealed non-significant differences in response rates among alternatives between groups, all $p > .05$. These consistently null results clearly indicate that the experimental manipulation was not effective in leading the experimental group to affirm more self-testing behavior than the control group. Although affirmative response rates were reasonably high for most questions, demand characteristics are not considered a viable criticism for the lack of differences between groups, as some questions (of which "yes" responses are normative and desired) elicited response rates where "No" was the modal response.

Exam, homework, and course grades were also analyzed for differences between groups, as it is possible that students who engaged in self-testing may have experienced learning benefits despite not responding that they used the self-testing strategy more than others. A series of *t*-tests was conducted on exam, homework, and course grades between groups. Descriptive statistics are provided in Table 2. In accord with the iClicker responses, all group differences on exam, homework, and course grades were non-significant, all $|t| < 1$ and $p > .05$. In combination with the null results in iClicker responses, these results clearly point to the ineffectiveness of the

classroom intervention in eliciting any evidence of increased self-testing usage, as well as lacking evidence for higher scores on any performance metric.

Discussion

Two salient interpretations exist for the results of this classroom intervention. The first is that students simply do not respond to being told how best to prepare for exams, and that efforts to instruct students on superior learning strategies are altogether unsuccessful. However, it is possible that different interventions may have exhibited greater success. For instance, the intervention used may have been too weak to elicit the desired changes in behavior. One possible improvement to the intervention would have been to make the instructional portion of the manipulation of a longer duration, or used reminders of greater detail and involvement than advice to use practice tests as a means of self-testing. An even more promising avenue for alternative interventions to increase self-testing appreciation would be to have students engage in self-testing during the lab as a means of establishing personal use of this strategy. These procedural details may be met with greater success than the brief and intermittent nature of the design used in the present experiment. Nonetheless, the failure to find any support for the view that the testing effect may be elicited from students by means of simple direct instruction constitutes firm footing for examining whether discovery of the testing effect through comparative experience can be successful in its place.

Chapter III

Experiment 2: Statistics Discovery

The success of deWinstanley and Bjork (2004) in increasing the learning outcomes of participants subjected to both generate and read-only trials motivates the present experiment into whether the testing effect may also be subject to this comparative, discovery-based learning phenomenon. Instead of using prose passages as in the generation effect experiments, the current experiment was conducted using statistics concepts in the hopes of drawing direct implications for science, technology, engineering and math (STEM) educational disciplines. It was hypothesized that by being forced to answer test questions on some trials and permitted to study without first providing a response on other trials, these subjects would discover that testing is associated with superior learning outcomes than is studying. Furthermore, this discovery would then alter the processing of subsequent study trials; students would engage in covert self-testing behavior in order to increase their processing of these otherwise less-helpful trials. As a consequence, we hypothesized that subjects exposed to this manipulation would exhibit a relatively weak testing benefit following their discovery experience.

Method

Participants. One hundred six students at the University of Colorado Boulder were sampled from the introductory psychology subject pool. Informed consent was obtained from all students, and participation was rewarded with course credit. The sample was composed of 64 women and 42 men, who averaged 18.8 years of age (range 18-24). Ethnic background was varied, with 68 White participants, 14 Middle Eastern, 11 Hispanic, 10 Asian, 1 Black, 1 Native

American, and 1 Pacific Islander. Thirty-four participants were randomly assigned to the Study-Only group, 37 were assigned to the Test-Only group, and 35 participants were assigned to the Contrast group. Nineteen participants indicated prior experience with statistics in a higher education setting (college or advanced high school courses), and three participants reported prior experience with research methods and design. Results from a pilot study found no evidence that previous experience with these materials leads to meaningful differences in the data, so these participants were not excluded from the following analyses.

Materials. Subject matter consisted of concepts adapted from the textbook *Statistics for the Behavioral Sciences (8th ed.)* by Gravetter and Wallnau (2009), covering standard statistical terms and procedures as well as modules covering research design. Concepts were selected from a larger pool of concepts that were determined in a pilot experiment to allow sufficient learning with minimal presentation time, ruling out both floor and ceiling effects. Three new concepts were added to bring the total to 16 mostly independent concepts. All materials were presented on a computer screen using PsychoPy stimulus presentation software (Peirce, 2009). For each of the 16 concepts, four presentations were constructed: an information screen (Figure 2), an example screen (Figure 3), a practice problem screen (Figures 4 and 5) and an exam screen (Figure 6), for a total of 64 presentations. Figure 7 shows a screen used to generate metacognitive judgments of memory, to be described below.

Procedure. All participants were tested by the author in groups ranging in size from one to four, in sessions taking approximately 90 minutes to complete. Each participant was seated in a different room and randomly assigned to one of three between-subject conditions: Study-Only, Test-Only, and Contrast, to be described below. All participants began the experiment by providing consent, demographic information, and prior experience with statistics

and research methods. Participants were required to relinquish their mobile phones to the experimenter for the duration of the experiment to prevent cheating. The various stages of Experiment 1 are depicted in Figure 8.

Study phase #1. The experiment began by providing a blank piece of lined paper and pencil to the participant who was instructed to use them if they wanted to (e.g., for math calculations), but told that they were not required to do so. They then read instructions presented on the screen as the experimenter left the room. These instructions asked the participant to study the following statistics concepts presented one at a time, and that this presentation would proceed in a consistent pattern: information about a concept would be presented, followed by an example of the concept put into context. After the example, they would then be asked to rate the likelihood that they would remember the previous information at a later test "in approximately 15-20 minutes" on a 1-6 point scale (Figure 7). After this metacognitive judgment, a practice problem would be presented. After responding to this problem, the procedure would begin again with a new concept, until eight concepts (complete with information, example, metacognitive judgment, and practice problem) were presented.

Unknown to the participants, the type of practice problem he or she would view constituted the only experimental manipulation of the study. For all participants, the practice problems appeared as questions accompanied by four multiple-choice options. For members of the Study-Only group, the practice problem screen contained an instruction at the bottom of the screen reading "Hold down the left mouse button to view the correct answer." By providing this feature, the Study-Only participants were not required to test themselves using the practice problem but could instead view and study the answer. Self-testing was required for members of the Test-Only group as the answer-revealing feature and accompanying instruction were absent,

and progressing past the practice problem required the participants to press the letter (*a*, *b*, *c* or *d*) corresponding to their answer. For members of the Contrast group, practice problems alternated between trials presented in study format (i.e., the answer-revealing feature and instruction were present) and trials presented in test format (i.e., the feature was absent). The assignment of practice problems to study or test format was counterbalanced to ensure independence of content and format. Following the provision of a response from the Test-Only participants and test trials from the Contrast group, the correct answer was presented. For those instances in which a participant did not use the mouse button to reveal the correct answer on a study trial (10.4% of trials for the Study-Only group, and 31% of study trials for the Contrast group), a warning message followed the practice problem reminding them of its availability and recommending its use. In this manner the presence of feedback was largely controlled between groups.

Exam #1. After presentation of all eight concepts, the paper for writing notes or self-testing was taken away and the first half of a paper-based questionnaire was issued as an interpolated task, with participants taking approximately 10-15 min to complete the questionnaire. This length of delay was selected to minimize the possibility of restudied information remaining more memorable than tested information, as has been occasionally observed with immediate exams and 5-min delays, as noted earlier. The first half of the questionnaire contained items measuring self-regulated learning (Self-Regulated Learning: Self-Report Scale; Toering, Elferink-Gemser, Jonker, van Heuvelen, & Visscher, 2012), depth of processing (Study Process Questionnaire; Biggs, Kember, & Leung, 2001), and motivation (Work Preference Inventory; Amabile, Hill, Hennessey, & Tighe, 1994). Following the questionnaire, a new piece of blank paper upon which to do work was provided. All participants then received eight multiple-choice exam questions on the computer, presented in the order in

which their respective concepts were studied. All questions involving mathematical computation were deemed simple enough that math ability would not be likely to affect performance, helping to ensure that only memory for the procedure would be implicated by response accuracy.

Following the provision of each answer, feedback was presented as follows: "Earlier in this experiment you were asked to SOLVE (/ STUDY) a practice problem relevant to the previous topic. On the question a moment ago, you were CORRECT (/ INCORRECT)". This method of providing feedback was used in order to make the connection between studying or solving (i.e., engaging in retrieval practice) and their differential rates of success salient to the participants in the Contrast condition. As neither the Study-Only nor the Test-Only participants engaged in both strategies during the practice problems, they would be expected to gain no insight into the benefits of retrieval practice from the exam feedback. The conclusion of this exam marked the end of the first phase of the experiment.

Study phase #2. The second phase of the experiment was in most respects identical to the first. Eight new concepts were presented with information, an example, a prospective memory judgment, and a practice problem. The key distinction between study phases 1 and 2 is that in the second phase, all participants, regardless of initial group, were treated as if they were members of the Contrast group. In other words, half of all practice problems afforded study by using the answer-revealing feature, and half did not, requiring the production of a response to that question. The purpose of this procedural change was to elicit measures of performance on both tested and studied concepts for all participants in the second exam.

Exam #2. The second exam was undertaken following completion of the presentation of the eight new concepts and the second half of the questionnaire, which contained items measuring trait metacognition (Metacognitive Awareness Inventory; Schraw & Dennison,

1994), personality (Big Five Inventory; John & Srivastava, 1999), and goal orientation (Button, Mathieu, & Zajac, 1996). The delay imposed by the second half of the questionnaire was roughly equivalent in length to the first.

It was hypothesized that the testing effect, operationally defined as the retention advantage for tested concepts relative to studied concepts, would be significantly greater among the Study-Only and Test-Only groups than among the Contrast group on Exam 2 performance. This hypothesis originates from the assumption that the Contrast group, once aware of the benefits of engaging in retrieval practice, would test themselves on practice problems even when not required to do so, as is the case in the study trials accompanied with the answer-revealing feature. The veracity of this assumption and the support for this hypothesis are evaluated below.

Results

Prior to any statistical analyses, the data were investigated for outliers. All participants exhibiting a median stimulus (i.e., information and example screen) presentation time of less than 9 s were excluded, resulting in 13 exclusions. This procedure was used to ensure that only participants who actually read the lengthy materials would be used in the following analyses. Three additional subjects were excluded for consistently failing to use the answer-revealing feature that characterized study trials, leaving a final sample size of 90 participants (29 in Study-Only, 32 in Test-Only, 29 in Contrast).

The present study sought to determine whether direct experience of the mnemonic benefits of testing relative to restudying would elicit an appreciation of self-testing. One major prerequisite for participants to discover and adopt the testing effect is that the testing benefit must be present; participants who do not experience greater memory for tested items relative to

studied items would have no basis for appreciating the testing effect. To determine whether a testing benefit was present for the Contrast group, a paired *t*-test was conducted on the Contrast group's within-subject variable of question type. This analysis revealed a marginally significant testing benefit, t (28) = 1.90, p = .07, which is significant under a one-tailed hypothesis test, p = .035. Descriptive statistics are provided in Table 3. The use of a one-tailed test is justified by the directional prediction that testing performance would be superior to study performance, as is intrinsic to the testing effect. It can be concluded that the Contrast group as a whole did experience a performance advantage for tested items. A between-subjects testing effect was also investigated by comparing performance between Study-Only and Test-Only subjects' performance on Exam 1, but revealed no difference between the groups' performance, F (2, 86) < 1.

To test the primary hypothesis that the testing effect would be discovered and adopted by the Contrast group, a 3 (group: study-only, test-only, contrast) by 2 (practice type: test, study) ANOVA was conducted with repeated measures on practice type, using proportion correct at Exam 2 as the dependent variable. An interaction was predicted, such that the difference in proportion correct between studied and tested concepts would be substantially smaller among the Contrast group relative to the other groups. Descriptive statistics for this analysis are reported in Table 4. Contrary to the hypothesis, the ANOVA revealed no main effect of group, F (2, 86) < 1, no main effect of practice question type, F (1, 87) < 1, and no interaction, F (2, 90) <1.

It was also possible that the hypothesized interaction would be observed not in accuracy but in response time. To test this idea, an identical 3 x 2 ANOVA was conducted, this time using response time (for correct answers only) as the dependent variable. Descriptive statistics for this analysis are reported in Table 5. Once again, the ANOVA revealed no main effect of group, F (2,

86) < 1, no main effect of practice question type, F (1, 87) < 1, and no interaction, F (2, 86) = 1.71, MSE = 154.50.

Despite the failure of these analyses to yield the predicted results, it is possible that the hypothesized reduction in the testing effect by the Contrast group would be observed as a within-subject comparison between Contrast group performance on the two question types between Exam 1 and Exam 2. If participants covertly adopted the testing effect during study trials in the second study session, it would be expected that any advantage for tested items might decrease on the second test. To evaluate this prediction, a 2 (question type: test, study) x 2 (exam number: 1, 2) ANOVA with repeated measures on both variables was conducted for proportion correct. As shown in Figure 9, the testing effect was decreased on the second exam, but this interaction was not significant, F (1, 27) = 1.20, $p > .05$.

These results indicate that although a testing benefit was present on the first test for the Contrast group, this benefit was either not discovered or not put to use on the second test. However, it is reasonable to suspect that the hypothesized effect was not observed because not all participants in the Contrast group experienced a testing benefit on the first test. It stands to reason that Contrast group members who did not exhibit a testing benefit on the first test would have no reason to adopt a retrieval practice strategy during subsequent study trials. The same 2 x 2 repeated measures ANOVA was conducted on the Contrast group after excluding participants who did not exhibit higher Exam 1 scores for tested items than for studied items. This time, the hypothesized interaction between question type and exam number was significant, F (1, 10) = 14.03, $p < .001$. This interaction can be observed in Figure 10.

By restricting the analysis to Contrast group participants who experienced a testing benefit on Exam 1, the hypothesized reduction of the testing benefit on the second exam was

observed. It is possible, however, that this test effect reduction was not due to the discovery and adoption of the testing effect, but instead constitutes an example of regression to the mean. To attempt to address this competing interpretation, a similar analysis was conducted on participants who performed better on studied items on Exam 1. If a parallel study benefit is reduced from Exam 1 to Exam 2, this would suggest that the previous test effect reduction is not specific to the act of testing. Such an interpretation would either imply a parallel study-effect adoption (which is not possible, given that one can't choose to study on a test trial - as the answer is missing - in a comparable way to the choice of self-testing on study trials), or regression to the mean. As can be seen in Figure 10, these participants exhibited a decreased study benefit on Exam 2 relative to Exam 1. This result was not significant, F (1, 5) < 1, MSE = .117, p > .05, perhaps due to the small sample of applicable participants (n = 6). The finding that participants showed a decreased benefit of their superior strategy from Exam 1 to Exam 2, whether that strategy was testing or studying, implies that regression to the mean is a likely cause for this pattern of results.

A final set of analyses was conducted to explore the accuracy of student metacognitive judgments. Student metacognitive accuracy can be evaluated in several ways. In the first, each students' metacognitive judgments, averaged across items, was correlated with student accuracy on the two exams similarly averaged, resulting in an overall significant, positive correlation between judgments of future memory and eventual performance, r (89) = .48, p < .001, which is depicted in Figure 12. The correlations were larger for judgments and performance on Exam 1 (r = .46) than on Exam 2 (r = .35), perhaps due to the inclusion on Exam 2 of trial types not encountered on Exam 1 among the Study-Only and Test-Only groups. This analysis reveals a reliable trend whereby student confidence is positively associated with student exam success. However, when metacognitive judgments are matched to the exam questions they corresponded

with when prompted rather than averaged across all items, subjects' correlations between their judgments and their exam performance was greatly diminished. In general, the correlation fell to r (88) = .19, p < .001. Again the correlations were larger on Exam 1, r (75) = .19, p < .001 than on Exam 2, r (82) = .15, p < .001.

It has been argued that Goodman-Kruskal gamma is a more appropriate correlation technique for analyzing metacognitive accuracy (Nelson, 1984) because metacognitive judgments may not be assumed to adhere to an interval scale of measurement. The previous analyses were redone using gamma instead of Pearson's *r*. The overall metacognitive accuracy of students, averaging across items, was gamma (88) = .38, p <.001, down from .48 found with Pearson's *r,* but still substantially positive, as gamma varies from -1 to +1 and may be interpreted similar to the Pearson correlation coefficient. Again, metacognitive accuracy was higher on Exam 1, gamma (75) = .37, p <.001, than on Exam 2, gamma (82) = .28, p <.01. When analyzed by item, again the correlations were reduced. Across both exams, subjects exhibited positive relationships between judgments and future performance, gamma = .28, higher than the .19 coefficient found using Pearson's *r*. Yet again, Exam 1 correlations were higher (gamma = .28) than on Exam 2 (gamma = .20). In summary, students exhibited substantial metacognitive accuracy when making judgments about their future memory of the target material in the present experiment.

Discussion

Participants exposed to the benefits of testing exhibited a decreased testing effect on the following exam. This result is in keeping with the primary hypothesis, as it would be expected that only participants who were exposed to the benefits of testing would discover and adopt the

strategy for later use. Whether this result was due to the discovery and adoption of the testing effect is not certain, as regression to the mean is a viable competing interpretation.

It is important to consider why test trials did not outperform study trials for a large number of Contrast group participants. It is likely that the experimental manipulation was not strong enough to cause all participants to experience the benefits of testing relative to studying. Although deWinstanley and Bjork (2004) successfully employed a similar procedure to elicit awareness of the related generation effect, it is perhaps the case that the differences between these experiments negatively impacted the ability to elicit the testing benefit. For instance, in the aforementioned study, the experimental manipulation consisted of varying whether phrases containing one critical word were to be read or to be generated. In contrast, the present experiment, in the attempt to use ecologically valid statistics materials, introduced many times as much information (i.e., information, example, and practice problem screens, each containing several phrases per critical concept). Furthermore, the exams used by deWinstanley and Bjork to assess memory for the studied or generated words were fill-in-the-blank problems and used identical phrases as those previously viewed by participants during the training session, with only the critical to-be-remembered word removed. In contrast, the present study used entirely new exam questions to assess memory for the concepts, and exam performance required critical thinking to apply the studied or tested information to the context of the exam question. One might conclude that participants in the present study were being tested for conceptual understanding whereas participants in the deWinstanley and Bjork experiment were being tested for episodic memory. The extra processing that these tests required might have been too much to ask from some participants who otherwise might have exhibited greater performance on simpler self-tested materials with more direct exam questions.

Although participants who exhibited the testing effect also exhibited a significant reduction in that effect on the second phase of the experiment, it is possible that this reduction would have been even more substantial if the effectiveness of the feedback provided was improved. As there were four study trials and four test trials, perhaps it was overly difficult for participants to compare performance between these trial types. Feedback was presented immediately after each exam question, likely requiring the participant to remember past feedback in order to contrast rates of success. An explicit performance summary might have better permitted direct comparison of performance between trial types. These three divergent elements -- greater complexity of presented materials, use of novel exam questions, and suboptimal feedback -- are perhaps sufficient to explain why some participants did not become aware of the testing effect during the first phase of the present experiment.

As a consequence of the variety of procedural details that may have hindered the ability to elicit the testing effect among more participants, a follow-up study was justified to better examine the possibility of learning and using the testing effect through direct comparative experience.

Chapter IV

Experiment 3: Chemistry Discovery

Experiment 3 was conducted to rectify the procedural challenges observed in Experiment 2 while attempting to replicate its primary conclusions; that a within-subject testing effect is reliably observed using STEM materials, yet despite the availability of direct experience attesting to the superiority of testing over restudying, students do not adopt self-testing as a strategy to enhance their subsequent learning. Whereas Experiment 2 emphasized ecological validity in the selection of materials and procedural details, Experiment 3 was conducted to maximize internal validity through simplistic materials and the use of more direct methods of assessing covert self-testing. It was hypothesized that the testing effect would be replicated in Experiment 3, and that indicators of covert self-testing would be observed among the Contrast group following exposure to both testing and restudying strategies.

Method

Participants. One hundred thirty five participants (68 female, 67 male) were sampled from the University of Colorado Boulder introductory psychology subject pool (mean age = 19.24) and randomly assigned to one of three between-subjects groups (45 subjects per group). Participants were moderately diverse in ethnicity, with 89 White, 18 Asian, 13 Hispanic, 6 Middle Eastern, 5 Black, and 4 other ethnicities observed. Fifty-two participants reported prior experience with college chemistry courses, and 13 were chemistry majors. Only one participant correctly identified more than two of the 32 relevant materials on the pre-test and was excluded from all analyses, leaving a final sample size of 134.

Materials. Stimuli consisted of simple images depicting real chemical element symbols paired with the appropriate atomic numbers. These stimuli were selected in order to retain ecological validity while heightening experimental control. No elements under atomic number 32 were included, as many of these elements are commonly known even to non-chemists. A rough attempt was made to control for frequency of element letters and atomic numbers; no letter was used more than four times for either the initial or final letter, no initial number was used more than five times, and no final number was used more than four times. Examples of how stimuli appeared in this experiment are depicted below, showing items presented for study (Figure 13) and for tests as well as exams (Figure 14). All materials were presented on a computer screen in PsychoPy stimulus presentation software. The full list of elements used is shown in Figure 15.

Procedure. All participants were tested by the author in groups ranging in size from one to four, in sessions taking approximately 60 min to complete. Each participant was seated in a different room and randomly assigned to one of three between-subjects strategy conditions: Study-Only, Test-Only, or Contrast, as well as to one of three between-subjects delay conditions: short, medium, and long, leading to the formulation of nine between-subjects conditions in a 3x3 design. The various stages of Experiment 2 are depicted in Figure 16.

Pretest. Participants began the experiment by providing consent, demographic information and prior chemistry experience such as college major and number of chemistry courses previously taken. Participants were required to relinquish their mobile phones to the experimenter for the duration of the experiment to prevent cheating. Each participant was then provided with a blank periodic table, showing only the atomic numbers and their boxes in the familiar layout (Figure 15). Participants were asked to input the chemical symbols where they

belonged, and to work on the task for a minimum of 1 min, and afterwards to inform the experimenter when no further elements could be provided.

Study A. To begin the first study phase, participants were seated at a computer and left to read and follow the presented instructions. These instructions were identical for all conditions, and read:

> *Today you will be asked to memorize the atomic elements. Each element has both an atomic symbol (such as He for Helium), and an atomic number associated with it (such as 8 for Oxygen). Although there are 118 different elements, today you will be asked to learn only 32 of these. Your goal in this experiment is to memorize the atomic symbol and atomic number for these 32 elements.*
>
> *When the experiment begins, you will be shown 16 elements, one by one. On the top of the screen you will see an atomic symbol, and at the bottom of the screen, you will see an atomic number. Each element will be presented for 10 seconds, and then the next element will appear automatically. You should try to memorize the atomic symbol and atomic number during this study session.*
>
> *Later in the experiment, you will be given a memory test for these elements, and you will be shown how many correct answers you received. Consider yourself challenged to do as well as possible!*

Participants were then given the opportunity to ask the experimenter any questions, and pressed a button to begin. All participants proceeded to study the first 16 elements (List A) for 10 s each. No writing materials were provided, preventing off-loading of the material.

Survey 1. Following this initial study phase, a delay period was imposed by requiring participants to respond to 27 questionnaire items randomly selected from the Motivated Strategies for Learning Questionnaire (MSLQ; Duncan & McKeachie, 2005; Pintrich, Smith, Garcia & McKeachie, 1993) assessing intrinsic and extrinsic goal orientation, task value, control beliefs about learning, self-efficacy for learning and performance, test anxiety, rehearsal, elaboration, metacognition, and effort. The delay took an average of 3 min 47 s to complete. After these items were responded to on the computer, the strategy manipulation commenced (Figure 17).

Strategy manipulation. Participants in the Study-Only group repeated the same experience as in the first study phase, viewing the same 16 elements of List A for 10 s each, albeit in a newly randomized order. Participants in the Test-Only group experienced List A as a series of test questions (Figure 14) with immediate feedback (identical in presentation to the respective Study-Only trial; see Figure 13). Instructions informed these participants that they would have 5 s to type a response, and afterwards would be shown the correct response for an additional 5 s. Participants in the Contrast group were told that they would alternate between the aforementioned strategies. The Contrast group was counterbalanced such that one sub-group studied the even-numbered items (i.e., their order within the presentation list, not their atomic numbers) and tested on the odd-numbered items, and another sub-group did the opposite. No results differed between these sub-groups, so they will not be distinguished going forward.

Delay manipulation. Following the strategy manipulation phase, a variable delay period was imposed prior to taking the List A exam. The purpose of this delay was to maximize the possibility of participants experiencing a testing benefit, as previous research has indicated that the testing effect is sometimes not observed at very short durations and is most apparent

following substantial delays (Rowland, 2014). By manipulating the delay between item exposure and the exam for those items, it becomes possible to evaluate the influence of delay on the magnitude of the testing effect. Three between-subject delay conditions were used, differing in tasks used to fill the delay. In the short delay condition, participants studied List B items in the same manner that List A items were experienced during the initial study phase. Participants in this delay condition took an average of 3 min 20 s to complete the task. In the medium delay condition, participants completed a pencil-and-paper letter detection packet and then studied list B items, taking an average of 9 min 13 s to complete both tasks. Participants in the long delay condition completed the letter detection task, then took a multiple-choice exam on various Psychology 101 concepts (that did not contain any explicit references to learning or memory), and then studied the List B items. This long delay condition took an average of 17 min 30 s to complete. It is worth noting that the study phase for List B always occurred immediately prior to the List A exam, regardless of delay condition, so proximity of the interfering list to Exam A was controlled, whereas proximity of the interfering list to the previous exposure of List A was not.

Exam A. Following this delay period, the List A exam was conducted. Exam A was similar in implementation to the Test-Only group's experience during the strategy manipulation phase; all 16 items were presented in a newly randomized order, and participants had to type in a response before progressing to the next screen. A timeout feature was used and set to 20 s in order to derive more response time measures than would be possible with a far shorter timeout period (it was set to 5 s during the strategy manipulation in order to equate stimulus presentation time between strategy conditions). Unlike the strategy manipulation phase, no immediate feedback was provided in order to protect against criticism that exam accuracy was affected by a reduction in valid answer options as the exam period progressed. Importantly, because items

were randomized on Exam A, there was no obvious manner for the Contrast group to discern from which sub-list (restudied or tested) a given exam item originated.

Following the final exam item, a feedback screen was presented that detailed and summarized the performance of the participant on the exam. Two example feedback screens are depicted in Figure 18. After the feedback screen, which terminated after 30 s, participants responded to an open-response reflection prompt: "Which strategies did you find most effective for learning the elements?" These two screens were used in the hopes of making explicit the relative effectiveness of the test and study strategies to the Contrast group; neither of the other groups were expected to learn about strategy effectiveness from these screens but instead were expected to interpret them in light of their overall exam performance.

Choice phase. After Exam A concluded with the feedback screen and reflection prompt, all participants experienced the remainder of the experiment identically regardless of condition. The choice phase instructed participants that List B items would be presented again, this time without the accompanying answer. By pressing the space bar at any time (before trial completion at 10 sec) the correct answer would be revealed, hence this phase constituted a choice between self-testing with the primary prompt or studying the correct answer after pressing the space bar. Importantly, the duration between stimulus onset and space bar press was assumed to function as an index of covert self-testing behavior.

Survey 2. Following the choice phase, participants completed the second half of the MSLQ survey, which took on average 2 min 53 s to complete.

Exam B. Subjects then took the exam for List B in identical fashion to that employed for List A, albeit without a reflection prompt at the end. It is worth reiterating that, due

to the presence of the choice phase during List B learning, there was no overt distinction between studied items and tested items on List B.

Post-test. After Exam B, another blank periodic table was administered as a post-test. Participants then took the letter-detection task and multiple-choice psychology exam if they had not already done so and if materials were available, then were debriefed and thanked for participation.

Results

It was hypothesized that the testing effect would be replicated in Experiment 2. To evaluate this hypothesis, Exam A scores were first compared between Study-Only and Test-Only groups (a between-subjects testing effect) and then between the contrast group's scores on restudied and tested items (a within-subject testing effect). Prior to these analyses, two subjects were removed as they exhibited mean reaction times of under 1 s, a time deemed insufficient to have responded with any intent of exhibiting accuracy. The between-subjects analysis was conducted using a 3 (strategy condition: Study-Only, Test-Only, Contrast) x 3 (delay condition: short, medium, long) ANOVA on Exam A accuracy. The main effect of strategy condition was significant, F (2, 123) = 4.14, $p < .05$ (Figure 19). A post-hoc analysis using Tukey's HSD revealed that the Study-Only group significantly outperformed the Test-Only group; no other significant differences were observed. The main effect of delay condition was not significant, F (2, 123) = 1.58, $p = .21$, nor was the interaction between the two factors, F (4,123) < 1. These results indicate that the Study-Only group (M = 6.45, s = 3.62) significantly outperformed the Test-Only group (M = 4.24, s = 3.62), and the magnitude of this performance advantage did not vary meaningfully as delay increased. As half of the Contrast group's trials consisted of restudy

opportunities and the other half consisted of testing opportunities, it was predicted that this group would perform at a level between the two other groups. Figure 19 does not support this prediction, as the Contrast group exhibited the lowest score. Most importantly, and in direct opposition to the hypothesis of a testing advantage, the Study-Only group learned or retained the List A information better than did the Test-Only group across all delay conditions.

A parallel 3 x 3 ANOVA was conducted using the same set of predictors but using reaction time (to correct items only) as the dependent variable. Results of this analysis indicate no main effect of strategy condition, F (2, 114) = 1.52, $p > .05$, no main effect of delay condition, F (2, 114) = 1.05, $p > .05$, but a marginally significant interaction, F (4, 114) = 2.13, $p = .08$. Closer scrutiny of this interaction was warranted, and a simultaneous multiple regression was used with contrast coded predictors to average across the effects of short and medium delay conditions compared to the long delay condition. Furthermore, the Contrast group was omitted from this 1-df interaction test,[2] which revealed a significant interaction F (1, 114) = 8.23, $p = .005$. The Study-Only group (short: $M = 4.93$ sec; medium: $M = 5.14$ sec) and Test-Only group (short: $M = 4.38$ sec; medium: $M = 4.80$ sec) group responded with roughly equivalent reaction time collapsing across short and medium delay conditions, but at the long delay condition, the Study-Only group ($M = 4.37$ sec) was significantly faster than the Test-Only group ($M = 6.24$ sec). The Contrast group ($M = 5.53$ sec) exhibited reaction times between these two groups. This interaction is shown in Figure 20.

---

[2] The contrast coded predictors of this analysis were constructed as follows:
Group: Study-Only (1/2), Test-Only (-1/2), Contrast (0);
Delay: Short (-1/3), Medium (-1/3), Long (2/3)
Complementary contrast codes were also entered into the regression to ensure orthogonality, although these predictors were not significantly predictive and are not reported herein.

A parallel analysis was conducted using delay as a continuous variable (preserving subject-to-subject variability in the duration of the delay tasks) rather than as a discrete group variable (i.e., short, medium and long), and the same interaction was found, F (1, 117) = 8.30, p = .005, ensuring the previous interaction was not an artifact of discretizing the delay variable. This interaction can be seen in Figure 21.

The within-subject analysis of the testing effect was conducted using a repeated measures t-test on item type (previously restudied or previously tested during the manipulation phase). The results of the t-test indicated the presence of a marginally significant effect of item type, t (44) = 1.78, p = .08, significant under a one-tailed test. The mean restudied item score for the Contrast group was 2.47 (31%; s = 2.05) whereas the mean tested item score was 2.89 (36%; s = 2.19; see Figure 22).

These analyses were supplemented by a 3 (delay condition: short, medium, long) x 2 (item type: restudied, tested) mixed-model ANOVA with repeated measures on item type. This analysis was conducted to determine whether the effect of item type interacts with delay condition. The results of this analysis yielded no significant interaction, F (2, 42) = 1.1, p > .05, nor was the effect of item type significant, t (42) = 1.46, p = .15. These null results should be interpreted in the context of the delay conditions; because both the medium and long delay conditions contained self-paced distractor tasks (letter detection and/or multiple choice Psychology 101 quiz), there was substantial variability in the delay length between subjects in the same delay conditions (see Figure 22). To account for this variability, the previous analysis was run using delay time (in s) as a continuous predictor rather than as a discrete grouping variable. This analysis would be expected to be more statistically powerful than the discrete analysis.

To conduct such an analysis, a linear regression was used that was nearly identical to the discrete analysis except a continuous delay time variable was used in place of the discrete, contrast-coded delay predictors in the aforementioned ANOVA (the continuous nature of the predictor variable required regression to be adopted over ANOVA). Additionally, the continuous delay predictor was centered (i.e., each subject's score was subtracted from the mean delay time) so that the analysis of item type would strictly evaluate whether the difference in performance by item type was significant when delay time was at its mean (606 sec). This analysis revealed a marginal effect of item type, t (43) = 1.8, p = .08, and no interaction with delay, t (43) = 1.31, p > .05. A series of similar targeted analyses were then conducted to examine whether performance differed by item type when delay time was held at the mean for each of the delay conditions (short = 200 sec, medium = 551 sec, long = 1,067 sec). To conduct each analysis, a predictor variable was created by centering the delay variable around the average delay duration for each level of the discrete delay conditions. The variable *Delay1* was created by subtracting 200 seconds from all subjects' delay score (the actual mean duration of time that subjects in the short delay condition experienced the delay phase). This transformation provided the ability to test whether the within-subject effect of interest (item type) was significant when the delay was 200. Centering predictor variables around theoretically motivated values is a valid and useful approach to determining the conditions under which an effect does or does not hold (Judd, McClelland, & Ryan, 2009). Similar centering transformations were conducted to create the variables *Delay2* and *Delay3* to allow tests of significance of item type at the average duration of medium and long delay conditions, respectively. This series of linear regressions revealed a significant within-subjects testing effect at the short delay, t (43) = 2.18, p < .05, a marginally significant effect at the medium delay, t (43) = 1.97, p = .055, but no effect of item type at the

long delay, t (43) = .1, p >.05. These findings suggest that the marginally significant within-subjects testing effect found in the original *t*-test (ignoring the delay predictor) is averaging over substantial item type differences at short delays with minimal item type differences at longer delays. The descriptive statistics bear this point out, as seen in Table 6. Short and medium delays were accompanied by tested item advantages in accuracy, whereas long delays saw this advantage disappear – a finding in direct contradiction to previous research showing more substantial test effects at longer delays. The aforementioned tests suggest that a within-subjects testing effect was observed at short delays.

It may be surprising that a testing benefit was observed within-subjects, yet a restudy benefit was observed between-subjects. The cause of such a surprising divergence in results is a topic of focus in the discussion to follow. To note, a parallel series of analyses was conducted using item type and delay (both discretely and continuously) as predictors of reaction time to correct items, but analyses revealed non-significant results, all $p > .05$. This Chemistry Experiment was conducted to determine not only that the testing effect could be replicated in the present paradigm, but that participants in the Contrast group would become aware of this benefit and consequently engage in covert self-testing to improve future performance. This hypothesis was tested by computing an average choice reaction time (the latency between stimulus presentation and press of the space bar to reveal the answer during the choice phase; see Figure 16) for use as an operational definition of covert self-testing. This choice RT was used as the dependent variable in a 3 (strategy condition: Study-Only, Test-Only, Contrast) x 3 (delay condition: short, medium long) ANOVA. This analysis revealed no main effect of either strategy, F (2, 4) = 1.35, $p > .05$, or delay condition ($F < 1$), and no interaction ($F < 1$). This analysis

indicates that the Contrast group did not exhibit greater indicators of covert self-testing than did the other groups (see Figures 24 and 25).

Subjects in the Contrast group who exhibited a testing benefit on Exam A would be expected to appreciate the testing benefit more than contrast subjects who did not experience the testing benefit on Exam A. As a result, it is reasonable to speculate that a between-subject analysis restricted to Contrast group participants would reveal that those who did enjoy a testing benefit would exhibit longer choice RT than subjects who performed equivalently across item type or even show a study advantage. A *t*-test was used on the choice RT measure, such that Contrast subjects with the test benefit present were compared with those with the test benefit absent. This analysis revealed a trend in the predicted direction, $t\ (43) = 1.60$, $p = .12$ (see Figure 26). With only 19 subjects in the test benefit present condition, this null result may be due in part to deficient sample size. However, when conducting a parallel analysis using testing effect as a continuous variable (i.e., the magnitude of the testing effect per subject) instead of as a grouping variable (i.e., its presence or absence), the analysis is decidedly non-significant, $|t|\ (43) < 1$. It is not certain whether these conflicting results indicate a lack of any true effect, or whether the presence or absence of a testing benefit is more predictive of future self-testing behavior than is the extent of the testing benefit, as might be the case if one's strategy adoption occurs due to the presence of a superior strategy without regard to the magnitude of that superiority.

It is important to note that if choice RT is really an index of covert self-testing, one would expect a positive relationship to exist between choice RT and Exam B performance. This relationship was found across participants in all three groups, as shown in Figure 25. A strong, positive correlation was observed between choice RT and Exam B accuracy, $r\ (132) = .33$, $p < .001$ (blue line). Two subjects exhibited scores far removed from the general trend (see Figure

27). With these subjects excluded, the correlation rose to r (130) = .35 (green line). The correlation between choice RT and Exam B performance may be interpreted as evidence that subjects entered the experiment with individual differences in proclivity to self-test, and this behavior was positively associated with Exam B performance. Importantly, choice RT was also significantly correlated with Exam A performance, r (132) = .37, $p < .001$, including subjects identified as Exam B outliers, and r (130) = .39 without those subjects. This finding implicates a stable individual difference among subjects in exam preparation, because if the former correlation were due only to a temporary decision to self-test, this activity in the choice phase could not affect performance on Exam A. In other words, because choice RT cannot affect exam performance that preceded it, the observation that choice RT was significantly correlated with Exam A performance as well as Exam B performance strongly suggests the importance of stable individual differences in self-testing (and related learning strategies) that are not captured by scales in the Motivated Strategies for Learning Questionnaire (see Table 10).

Discussion

Experiment 3 (Chemistry) improved upon Experiment 2 (Statistics) in several ways. First, Experiment 3 used a stronger manipulation, with trial types distinguished by stimulus presentation (paired associate present or absent, pending a response) instead of by presence of an on-screen instruction to reveal the answer when desired. Second, Experiment 3 used more tightly controlled stimuli that more strictly implicated memory, whereas Experiment 2 – due to the complex, lengthy stimuli – may have been assessing not only memory but additional cognitive phenomena such as critical thinking, conceptual understanding, and procedural skills. Third, Experiment 3 presented feedback in a summative and explicit manner, allowing an easier

comparison of performance between item types (when applicable), ideally facilitating the induction of the benefits of one strategy over another. Most importantly, Experiment 2's reliance upon observing a decreased testing effect on Exam 2 made it difficult to rule out the regression to the mean interpretation, whereas this interpretive obstacle was avoided in Experiment 3. As a consequence of these procedural improvements, the results of Experiment 3 are largely more convincing than those gained from Experiment 2.

The observation of a within-subject testing effect permitted the analysis of testing benefit awareness by comparing participants who did exhibit the testing benefit to those who did not on a measure of covert self-testing, choice reaction time. By abstaining from revealing the correct answer for a longer period of time, it can be assumed that such subjects are engaging in self-testing more so than subjects who choose to see the answer after shorter durations. Although the difference in choice RT was not significant, the results trended in the predicted direction and the null results may be due to the small sample of 45 subjects in the Contrast condition, which was subdivided into those exhibiting a testing benefit and those who did not. With a sufficiently large sample, the trend observed may in fact reveal itself significant and indicate an adoption of the testing effect among those who experienced a testing advantage. However, the magnitude of the testing benefit did not predict choice reaction time, which may indicate that the null result is a faithful representation of reality, or that the magnitude of a testing effect does not engender greater self-testing behavior so long as the testing effect is observed at all.

Nevertheless, it must be questioned why subjects in the Study-Only and Test-Only groups did not exhibit shorter choice RT than did subjects in the Contrast condition, as the latter was the only group expected to appreciate the benefits of self-testing due to the availability of feedback on Exam A that compared the effectiveness of the two strategies. It is possible that

demand characteristics caused all subjects to spend considerable time self-testing, or at least waiting before pressing the button to advance to the next screen. It is also possible that the Contrast group on average did not appreciate the testing effect enough to change their well-ingrained pace of exam preparation.

Results from Experiment 3 thus provide mixed evidence for the awareness and adoption of the testing effect. Although subjects who experienced a testing benefit waited longer to reveal an answer than did subjects who did not experience a testing benefit, this difference was non-significant, and combined with the lack of a between-groups effect of choice RT, it is perhaps prudent to conclude that the elicitation of testing effect awareness was not successful in Experiment 3. Subjecting more participants to the Contrast group manipulation and determining whether the trend towards test effect adoption solidifies as statistical power increases with sample size would likely be a worthwhile endeavor.

Chapter V

General Discussion

Summary

The present study was conducted with the primary aim of determining whether students can come to appreciate the beneficial effects of testing as a means of enhancing learning outcomes, and to put this appreciation to use by engaging in this strategy at subsequent opportunities. Such an investigation has clear practical implications, given that a majority of students contend that restudying, not self-testing, is the most effective way to learn and retain information, even though the psychological and educational research on this matter is definitive that testing is more conducive to long-term learning than is restudying (e.g., Rowland, 2014). Methods for lessening the disparity between how students prepare for exams and how they should be doing so are surprisingly unaddressed in existing research. By subjecting students to situations that either directly inform them of self-testing's benefits or by directly exposing them to those benefits in relation to those received by restudying, it was hoped that the present experiment would constitute a strong step towards bringing student practices more in line with research-based recommendations. Instead, the current set of experiments found little evidence that students can be brought to engage in self-testing as a self-directed learning strategy. The mitigation of suboptimal student strategy use and voluntary embracement of the testing effect remains a fundamental challenge to the domain of applied and educational psychology.

As a whole, the present investigation succeeded in demonstrating that STEM materials, whether ecologically valid and demanding conceptual understanding of statistics (Experiment 2) or simplistic and emphasizing associative memory of chemical elements (Experiment 3), can be used to elicit within-subject testing effects; that is, experiencing both test-required items and

study-required items led to consistent advantages for the tested items, statistically significant under one-tailed hypothesis tests. Surprisingly, between-subjects testing effects were not observed; in fact, accuracy was superior for the Study-Only groups. This restudy advantage has been found elsewhere (e.g., Roediger & Karpicke, 2006a), and is often explained with reference to the role of delay and decay; restudied items are more available immediately after exposure, but these representations decay at a faster rate than do representations strengthened by tested experiences (Delaney, Verkoeijen & Spirgel, 2010). It is interesting, however, that the immediate restudy advantage was only found in the between-subjects comparisons, especially considering that a recent meta-analysis has supported the claim that between-subjects testing effects are stronger than within-subjects testing effects (Rowland, 2014). One explanation may concern the role of displaced rehearsal (Slamecka & Katsaiti, 1987); Contrast group subjects may have prioritized the encoding of items being tested at the expense of items being restudied, while participants in the Study-Only and Test-Only groups, lacking exposure to both trial types, had no such difference in emphasis. If this interpretation is correct, it would become important to understand whether learners believe tested information is inherently more important than information only restudied, or whether importance does not differ but cognitive processes are naturally biased towards encoding tested information. Stated differently, are within-subjects testing effects due to the belief that tested items deserve more attention, or due to the workings of cognitive mechanisms that enhances the memorability of tested information? The former interpretation has been offered as an explanation for the superior generation effects obtained in mixed-list designs relative to generation effects obtained in pure-list designs (Slamecka & Katsaiti, 1987), whereas the latter is arguably a more theoretically interesting interpretation.

Several other unexpected yet noteworthy results were also obtained in the present investigation. First, students were surprisingly accurate in predicting their later performance on statistics concepts, despite many previous studies demonstrating that students are often poor at metacognitive judgments of their own learning (Dunlosky & Lipko, 2007), especially when metacognitive judgments are made immediately after initial encoding (Nelson & Dunlosky, 1991). Second, it was also found that the extent to which subjects wait to reveal a chemical's atomic number (an index of covert self-testing) positively predicts performance on those items in a final exam. This finding suggests that self-testing is conducive to superior performance, although the third variable problem cannot be ruled out. It could be the case that good students, however defined, tend to take their time during the experimental trials and are skilled learners, leading to high performance. It should be noted, however, that this hypothetical third variable was not identified using the individual differences surveys, which found no correlation with performance nearly as strong as choice latency. Lastly, interesting interactions emerged between performance on the first chemistry exam (as defined by reaction time to correct items) and the amount of delay they were forced to endure prior to that exam. This result may highlight the importance of interference from competing lists in the relative effectiveness of testing and restudying on performance; if testing and studying are differentially affected by the close proximity of competing lists, this result may serve to underscore the merits of competing theories of the testing effect. This last finding was quite unexpected and may offer a valuable vantage point into the mechanisms responsible for the testing effect should follow-up studies confirm these results.

Theoretical Implications

Many theories have been advanced which attempt to explain the basis of the testing effect; that is, how and why does testing enhance memory? The present investigation may provide support for a subset of these theories, and so each will be described in brief. The Transfer Appropriate Processing theory claims that the testing effect is due to the compatibility of processes that occur during learning via testing and during later retrieval; by virtue of the overlap in processing at learning and at performance, the process of retrieval itself is strengthened (Roediger & Karpicke, 2006a). The Elaborative Encoding theory states that retrieval enhances memory by forging semantic mediators between the original cue and its target, and these mediators function as additional cues in the later search for the target memory (Rawson, Vaughn, & Carpenter, 2015; Carpender & DeLosh, 2006). The Retrieval Effort/ Desirable Difficulties view claims that because retrieval is inherently more difficult than is restudying, and difficulty of initial learning engenders a more durable representation, testing is effective to the extent that it is difficult and involves effort by the learner (Bjork & Bjork, 1992). Another theory, which can be referred to as the Reliable Error Signal theory, claims that retrieval is more effective than restudying because testing affords more opportunities for explicit feedback than does restudying, and this differential feedback availability underlies the testing effect's advantages (Mozer, Howe & Pashler, 2004). Finally, the Test-Potentiated Learning view (Arnold & McDermott, 2013) states that in addition to the direct benefits of testing, there are indirect benefits as well, as restudying benefits learning after a testing experience. It is outside the scope of this paper to compare and contrast each relevant theory, but it should be noted how the results of the present investigation can be reconciled with these explanations of the testing effect.

One theory stands out in its ability to explain why some students are reluctant to appreciate and utilize the testing effect: the Desirable Difficulties view and its parent theory, the New Theory of Disuse (Bjork & Bjork, 1992). According to this view, restudying may boost *retrieval strength*, defined by its temporary and decaying nature, whereas testing may differentially boost *storage strength*, defined by its permanence. If students typically make judgments about the effectiveness of competing strategies based on their immediate effects on memory, then such learners may be selecting strategies that differentially boost retrieval strength over storage strength, thus handicapping their performance over the long term. This interpretation explains why the testing effect grows as retention interval increases, because retrieval strength diminishes as time elapses yet storage strength does not; consequently, if testing is a greater boost to storage strength than is restudying, the difference in performance is most apparent after a considerably delay. If students base their strategy use decisions on a moment-to-moment basis caused by implicit feelings-of-knowing primarily derived from retrieval strength, they may favor restudying over self-testing to an extent impervious to any interventions designed to make salient the benefits of self-testing.

The Desirable Difficulties view also finds support in the surprising interaction between group and delay on correct reaction time on chemistry Exam A. At long delays (but not at medium or short delays), the Study-Only group exhibited faster reaction times to correct responses than did the Test-Only group, in direct opposition to the common finding that the testing effect increases with retention interval. This result can also be interpreted with reference to storage and retrieval strength. At short delays, the Study-Only and Test-Only groups performed comparably. Under the premise that there was a greater boost to immediate memory among the Study-Only group due to the benefits of retrieval strength from restudying, one would

expect superior performance by the Study-Only group at short delays. However, it is important to emphasize that the delay manipulation was not idle time but rather included a highly interfering experience, the encoding of List B, which occurred immediately before the List A exam (at the conclusion of the delay period). It has been proposed that testing drives context change, which helps to differentiate between temporally proximal experiences (Jang & Huber, 2008). By testing immediately prior to encoding List B, the Test-Only group's List A representations are interfered with less than are those among the Study-Only group, for whom the two lists are difficult to distinguish due to their contextual overlap. The two groups thus perform equally, but due to two different mechanisms: greater retrieval strength among the Study-Only group, and greater contextual differentiation among the Test-Only group.

The difference in performance at the long delay can be explained by the same phenomena. Whereas retrieval strength continually decays among the Study-Only group, the harmful effects of encoding List B are mitigated by the passage of time and changing context occurring between encoding Lists A and B (which, in the long delay condition, were separated by a letter-detection task and a multiple-choice psychology exam). The Study-Only group's loss of retrieval strength over time is thus offset by contextual differentiation due to the intervening tasks between lists. These combined factors have the effect of maintaining relatively high performance (i.e., fast reaction time) among the Study-Only group at the long delay. The Test-Only group, however, saw its performance degrade (i.e., reaction time slow down) in the long delay condition. This degradation may have occurred because, as mentioned in the discussion of the short delay, the act of testing drives contextual change. The benefit of list differentiation attributable to contextual change among the Test-Only group is not aided by the passage of additional time nor the intervening tasks; at the short delay the Test-Only group was unique in

gaining benefits from contextual differentiation, whereas at the long delay the Study-Only group gained this benefit as well, and only additional decay distinguishes the Test-Only group at the long versus the short delay. It is presumably the case that at much longer delays the Test-Only group's memories would be stronger than the Study-Only group's as a result of having greater storage strength. Figure 27 depicts these four scenarios.

In sum, the present work has found much support for the Desirable Difficulties theory and New Theory of Disuse, both in their quality of explaining the failure to adopt the testing effect, and in their ability to account for the surprising interaction between group and delay condition on reaction time data in the chemistry experiment. The latter explanation is entirely post-hoc and would certainly gain from further experimentation to determine whether this interpretation stands up to closer scrutiny. Nonetheless, the combined explanatory value of these theories is useful for interpreting the most noteworthy results reported herein.

Limitations and Future Directions

Weaknesses of the present study include the brevity of the experiments relative to the tremendous amount of time students have had with well-ingrained exam preparation strategies, primarily restudying. Longer training with testing strategies and repeated exposure to the benefits of self-testing may be necessary to elicit awareness of the testing effect, and would constitute a valuable follow-up study to determine whether the results reported herein change as a function of intervention duration. A longer training regimen would also have the benefit of assessing the effects of competing strategies at longer retention intervals, which previous research has shown are far more likely to elicit between-subjects testing effects (Roediger & Karpicke, 2006a; Rowland, 2014). Furthermore, the present investigation attempted to elicit

adoption of the testing effect by direct instruction and by comparative experience in distinct studies. At no point were both methods attempted in the same sample of students. Combining instruction and direct experience is effective in formal education of facts and skills, perhaps this combinatory approach would be effective in eliciting adoption of the testing effect. Finally, future studies will need to take special care to motivate subjects. Unfortunately, subjects often have only a modest level of intrinsic motivation to succeed on the task they are asked to perform in the lab, and without extrinsic motivators (e.g., monetary reward for success), getting students to engage in a more difficult self-testing strategy as opposed to an easier restudy strategy may be overly optimistic. Future studies should find methods of motivating subjects, or selecting subjects already exhibiting motivation to learn and remember. One such method of selecting motivated subjects could be to sample students enrolled in a professional certificate course who must achieve a passing grade in order to apply for a desired career. Such subjects would have a clear motivation to embrace strategies that lead to superior learning and retention of information. Alternatively, testing may be incorporated into an enjoyable experience, such as a high quality interactive video game, which may reduce subjects' apathy. In sum, a longer training manipulation should be used, longer retention intervals should be assessed, a combination of direct instruction and comparative experience should be required, and subjects' motivation should be maximized in order to stand the best chance of eliciting awareness and adoption of the testing effect.

Conclusion

The mnemonic benefits of retrieval practice have been established for over a century. Unfortunately, there exists a wide and glaring chasm between learning practices that individuals

put to use, and those they should adopt if long-term learning is their goal. Well over a hundred scientific studies have been conducted on the testing effect, yet one of the most basic but important questions has not been adequately answered: How can educators raise the frequency of student adoption of this effortful yet effective learning strategy? The experiments reported in the present work were a first attempt to answer this important question. Instructing students that testing themselves leads to better learning outcomes was not successful, nor was arranging students to experience the benefits of self-testing firsthand. The logical manipulations employed herein were largely ineffective at addressing this divergence between ideal strategy use and actual strategy employment, despite the studies' basis in a common-sense intervention (direct instruction) and in empirically-justified methods of discovering learning (comparative experience). Nevertheless, the current work represents an important step in the process of uncovering effective methods to elicit adoption of the testing effect among a population of students who have much to gain from its use. It is hoped that by appreciating the difficulty of eliciting self-motivated retrieval practice, future researchers may design interventions that go beyond those used in the present work, and in doing so, improve academic learning outcomes by fueling the adaptive adoption of the testing effect.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs*, *11*, 159–177. Doi:10.1037/h0093018.

Amabile, T. M., Hill, K. G., Hennessey, B. A., & Tighe, E. M. (1994). The Work Preference Inventory: Assessing intrinsic and extrinsic motivational orientations. *Journal of Personality and Social Psychology*, *66*, 950-967. Doi :10.1037/0022-3514.66.5.950.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 940-945. Doi: 10.1037/a0029199.

Biggs, J., Kember, D., & Leung, D. P. (2001). The revised two-factor Study Process Questionnaire: R-R-SPQ-2F-2F. *British Journal of Educational Psychology*, *71*, 133-149. Doi :10.1348/000709901158433.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *Essays in honor of William K. Estes, vol. 1: From learning theory to connectionist theory; vol. 2: From learning processes to cognitive processes.* (pp. 35-67) Lawrence Erlbaum Associates, Hillsdale, NJ.

Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation. *Organizational Behavior and Human Decision Processes*, *67*, 26-48. Doi:10.1006/obhd.1996.0063.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268-276. Doi: 10.3758/BF03193405.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times recursive review of the literature. *Psychology of Learning and Motivation, 53*, p. 63-147. Doi: 10.1016/S0079-7421(10)53003-2.

de Winstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*, 945-955. Doi: 10.3758/BF03196872.

Duncan, T. G., & McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational Psychologist, 40*, 117-128. Doi: 10.1207/s15326985ep4002_6.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*, 228-232. Doi: 10.1111/j.1467-8721.2007.00509.x.

Gravetter, F. J. & Wallnau, L. B. (2009). Statistics for the Behavioral Sciences, 8th ed. Wadsworth, Cengage Learning. Belmont, CA.

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 112-127. Doi:10.1037/0278-7393.34.1.112.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. O. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). NewYork: Guilford Press.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). Data Analysis: A Model Comparison Approach, 2nd ed. Routledge. New York, NY.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*, 471-479. Doi: 10.1080/09658210802647009.

Keresztes, A., Kaiser, D., Kovács, G., & Racsmány, M. (2014). Testing promotes long-term learning via stabilizing activation patterns in a large network of brain areas. *Cerebral Cortex, 24*, 3025-3035. Doi: 10.1093/cercor/bht158.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*, 85-97. Doi: 10.1016/j.jml.2011.04.002.

Mozer, M. C., Howe, M,. & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. *Proceedings of the Twenty Sixth Annual Conference of the Cognitive Science Society* (pp. 975-980). Hillsdale, NJ: Erlbaum Associates.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109-133. Doi: http://10.1037/0033-2909.95.1.109

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*, 267-270.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*, 1-8. Doi: 10.3389/neuro.11.010.2008.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement, 53*, 801-813. Doi: 10.1177/0013164493053003024.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition, 43*, 619-633. Doi: 10.3758/s13421-014-0477-z.

Roediger, H. L., I.,II, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210. Doi: 10.1111/j.1745-6916.2006.00012.x.

Roediger H. L., I.,II, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255. Doi:10.1111/j.1467-9280.2006.01693.x.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432-1463. Doi:10.1037/a0037559.

Rudy, J. W. (2014). *The Neurobiology of Learning and Memory*, 2nd ed. Sinauer Associates. Sunderland, MA.

Sara, S. J. (2000). Retrieval and reconsolidation; Toward a neurobiology of learning and remembering, *Learning & Memory, 7,* 73-84. Doi: 10.1101/lm.7.2.73

Schraw, G., & Dennison, R. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*, 460-475. Doi :10.1006/ceps.1994.1033.

Shapiro, A. M., & Gordon, L. T. (2012). A controlled study of clicker-assisted memory enhancement in college classrooms. *Applied Cognitive Psychology*, *26*, 635-643. Doi: 10.1002/acp.2843.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26*, 589-607. Doi: 10.1016/0749-596X(87)90104-5.

Toering, T., Elferink-Gemser, M. T., Jonker, L., van Heuvelen, M. J. G. & Visscher, C. (2012). Measuring self-regulation in a learning context: Reliability and validity of the self-regulation of learning self-report scale (SRL:SRS). *International Journal of Sport and Exercise Psychology,* 1-15. Doi: 10.1080/1612197X.2012.645132.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41*, 429-442. Doi: 10.3758/s13421-012-0274-5.

Appendix: Tables and Figures

Table 1. iClicker questions used in the classroom intervention.

	% "Yes" Within Group for each of Three Assessments	
Question	*Exp.*	*Control*
While preparing for the last test, which best describes how you made use of the practice test?*	.89, .82, .73	.70, .85, .54
For the last test, did you study the notes and lecture slides?	.87, .84, .79	.81, .90, .63
For the last test, did you make up your own practice test and test yourself?	.10, .12, .19	.15, .07, .06
For the last test, did your look over your past HW and lab assignments?	.53, .48, .68	.38, .38, .67
For the last test, did you rework problems from HW and lab assignments?	.32, .44, .41	.25, .39, .43
For the last test, did you work through problems in the book that were not required for homework?	.10, .10 ^	.15, .12

** For this question, "Yes" corresponds to "I went through and answered questions like I would on a real test"*

^ Due to technical error, this question was accompanied by two instead of three data points

Table 2. Descriptive statistics of classroom performance metrics.

	Mean Grade per Group (sd)	
	Exp.	*Control*
Homework Total	90.17 (9.56)	91.43 (8.31)
Test 1	79.04 (9.61)	79.15 (9.92)
Test 2	70.03 (17.39)	69.42 (19.63)
Test 3	80.03 (17.95)	82.03 (16.32)
Final Exam	69.89 (15.91)	67.57 (16.91)
Final Course Grade	79.91 (11.35)	79.46 (12.23)

Table 3. Descriptive statistics for Statistics Experiment Exam 1 proportion correct.

	Studied	Tested	Mean (sd)
Study-Only	.76 (.15)	NA	.76 (.13)
Test-Only	NA	.79 (.16)	.75 (.14)
Contrast	.71 (.26)	.82 (.22)	.75 (.11)
Mean (sd)	.73 (.21)	.80 (.19)	.75 (.12)

Table 4. Descriptive statistics for Statistics Experiment Exam 2 proportion correct.

	Studied	Tested	Mean (sd)
Study-Only	.79 (.19)	.72 (.25)	.75 (.16)
Test-Only	.73 (.25)	.69 (.23)	.71 (.20)
Contrast	.73 (.20)	.76 (.21)	.75 (.11)
Mean (sd)	.75 (.22)	.72 (.23)	.74 (.16)

Table 5. Means and standard deviations for Exam 2 response time for correct responses.

	Studied	Tested	Mean
Study-Only	22.17 (6.90)	26.18 (10.66)	23.40 (6.02)
Test-Only	25.24 (12.63)	24.26 (9.67)	24.60 (8.14)
Contrast	23.02 (10.75)	21.63 (7.49)	22.25 (7.40)
Mean	23.52 (10.39)	24.03 (9.45)	23.44 (7.24)

Note that condition means are weighted by number of respective trials answered correctly

Table 6. Descriptive statistics for Exam A proportion correct among the Contrast group, by item type, at various delays (within-subject analysis).

	Short (M = 200 s)	Medium (M = 551 s)	Long (M = 1067 s)
Tested Items	.333	.358	.392
Restudied Items	.258	.267	.400
Difference	.075	.091	-.008

Table 7. Descriptive statistics for Exam A proportion correct among Study-Only and Test-Only groups, at various delays (between-subjects analysis).

	Short (*M* = 200 s)	Medium (*M* = 551 s)	Long (*M* = 1067 s)
Study-Only	.43 (.26)	.35 (.20)	.42 (.22)
Test-Only	.25 (.20)	.22 (.22)	.32 (.26)
fContrast	.29 (.21)	.31 (.22)	.40 (.30)

Table 8. Descriptive statistics for Exam A correct item reaction time among the Contrast group, by item type, at various delays (within-subject analysis).

	Short (*M* = 200 s)	Medium (*M* = 551 s)	Long (*M* = 1067 s)
Tested Items	4.85 (1.45)	5.66 (1.98)	5.72 (2.26)
Restudied Items	2.00 (4.79)	5.12 (1.70)	5.16 (1.72)
Difference	2.85	.54	.56

Table 9. Descriptive statistics for Exam A correct item reaction time among Study-Only and Test-Only groups, at various delays (between-subjects analysis).

	Short (*M* = 200 s)	Medium (*M* = 551 s)	Long (*M* = 1067 s)
Study-Only	4.93 (1.31)	5.14 (2.04)	4.37 (0.64)
Test-Only	4.38 (1.21)	4.80 (1.69)	5.79 (2.47)
Contrast	5.17 (2.08)	5.66 (1.68)	5.54 (2.00)

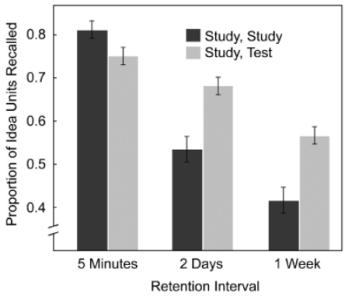
Table 10. Correlations between Choice RT, MSLQ scales Exam A & B Performance.

	Scale Reliability (α)	Exam A Accuracy (*r*)	Exam B Accuracy (*r*)
Choice RT	NA	.37*	.33*
Intrinsic Goal Orient.	.65	-.10	-.04
Extrinsic Goal Orient.	.61	.10	-.08
Beliefs about Learn.	.65	-.04	-.04
Effort Management	.75	.11	.08
Self-Efficacy	.90	.03	.07
Rehearsal Strategy	.56	.00	-.07
Elaboration Strategy	.68	.02	.07
Metacognition	.78	-.09	-.02
Test Anxiety	.66	-.02	-.07

** denotes significance at $p < .001$*

A Test Preparation Strategy for Statistics: Quiz Yourself!

- The most common test preparation strategy used by college students: Re-read notes and/or textbook.
  - *This is not the best way to prepare for your tests*
- Research in cognitive and educational psychology has shown that **you remember information better if you quiz yourself on it** than if you re-study it, but most people are unaware of this fact.
- Practice makes perfect; re-reading information does not.
  - When you quiz yourself, you are practicing the act of remembering that information. It becomes easier to recall that information later.
  - When you re-study, you are attempting to memorize the information you already know. This does not strengthen your memories as much as self-quizzing.



- Roediger, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17* (3).

- This graph comes from a classic paper (cited over 390 times in research articles), and the results it shows have been replicated many times.
- Participants in this experiment studied information, and then either studied it again later, or took a practice test on it.
- When they took the real test 2 days or 1 week later, **information that was quizzed earlier was remembered far better than information only restudied.**

Moral of the Story: Quiz Yourself for Better Test Scores

Figure 1. Three PowerPoint slides used in Classroom Quasi-Experiment instruction

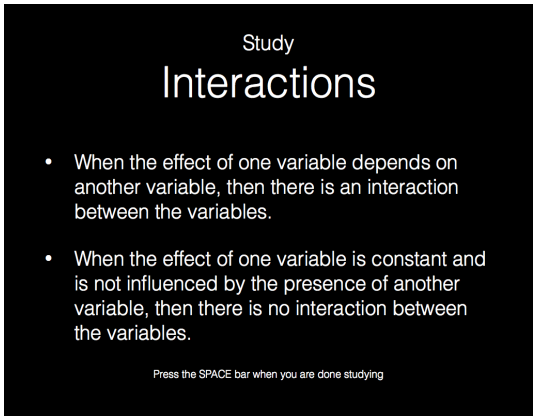


Figure 2. An information screen.

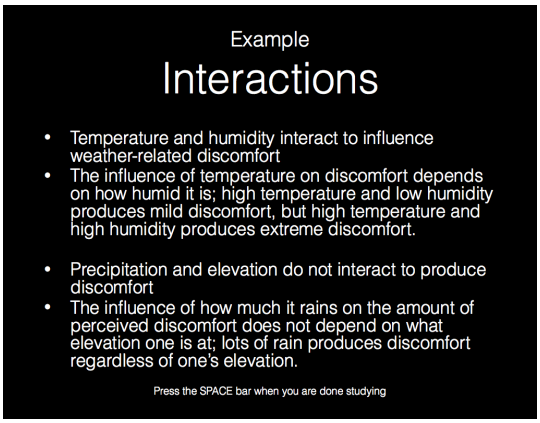
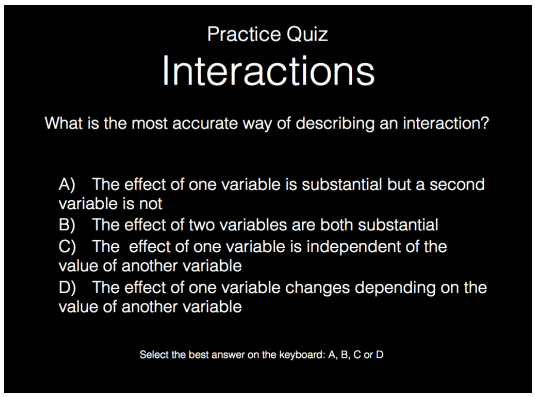
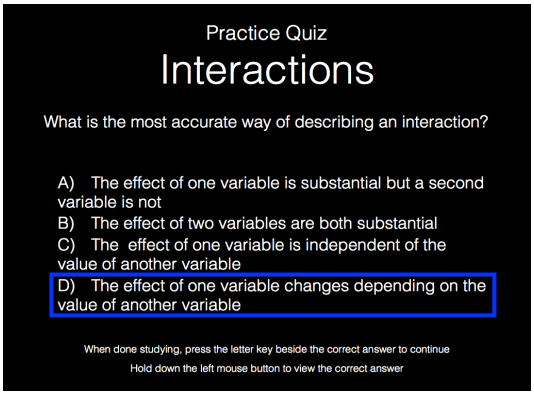


Figure 3. An example screen.





Figures 4 and 5. Practice problem screens seen on Test (3) and Study (4) trials.

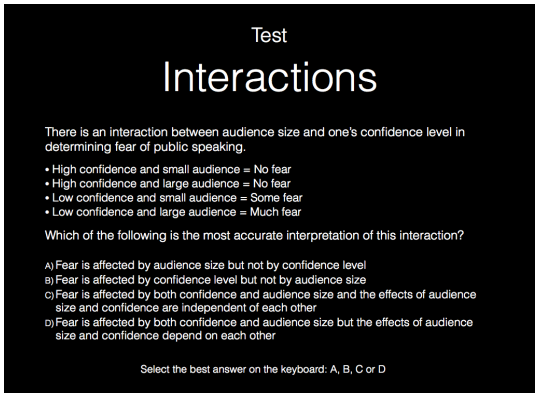


Figure 6. An exam screen.

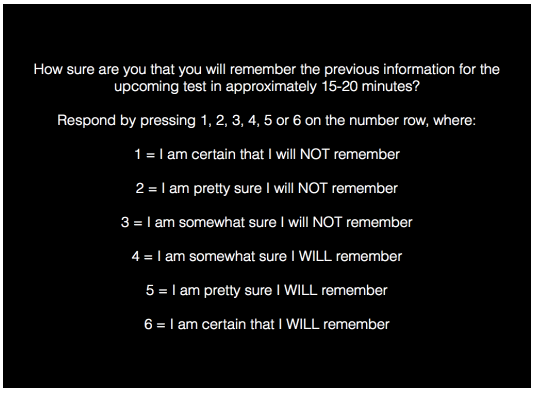


Figure 7. A metacognitive judgment screen.

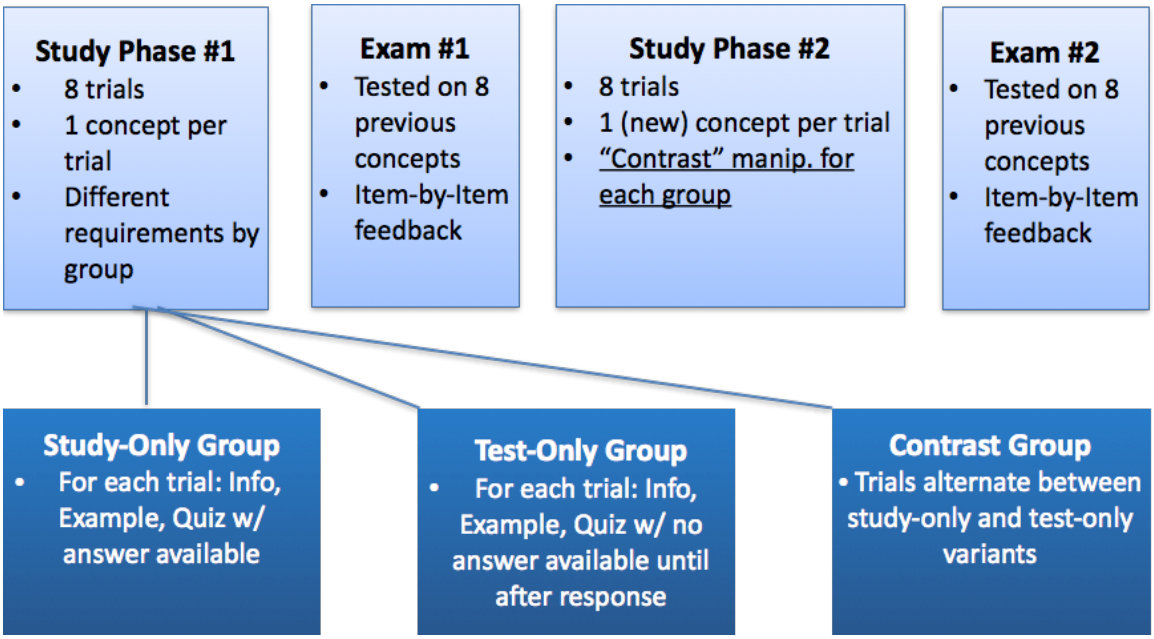


Figure 8. The general procedure for the Statistics Experiment.

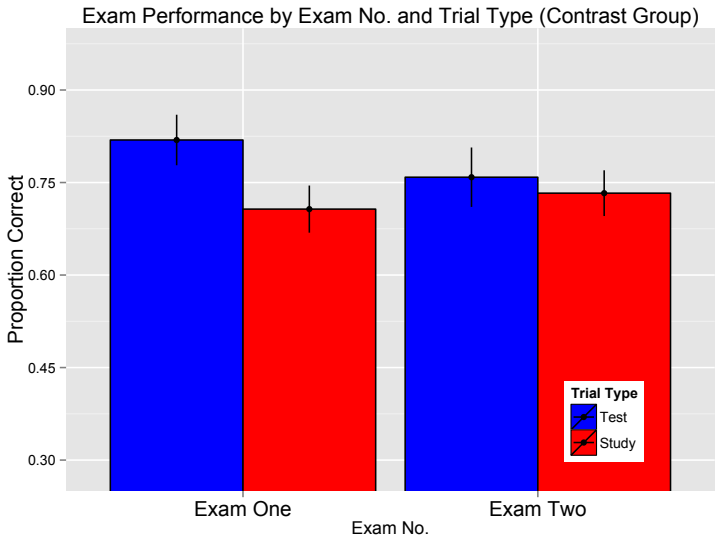


Figure 9. The effect of trial type for all Contrast subjects on both exams. Here and elsewhere, error bars represent standard errors.

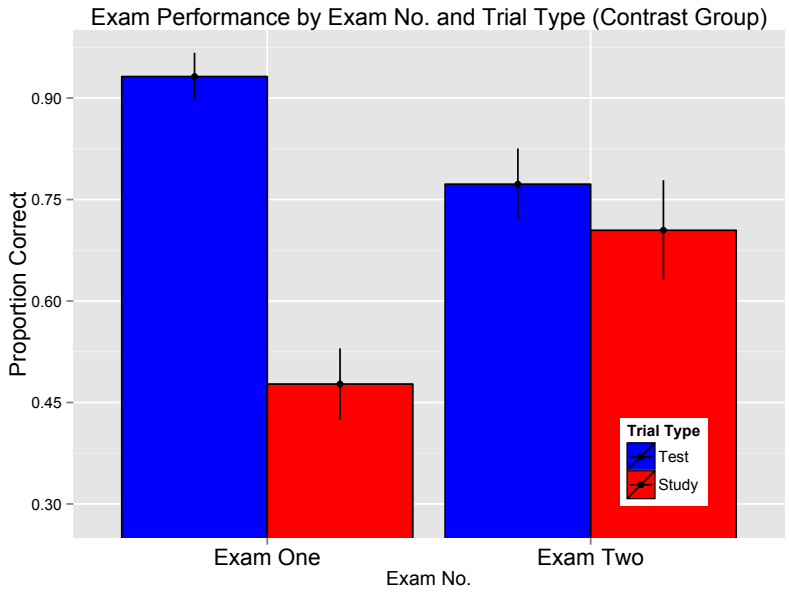


Figure 10. The effect of trial type for Contrast subjects with test benefit on Exam 1.

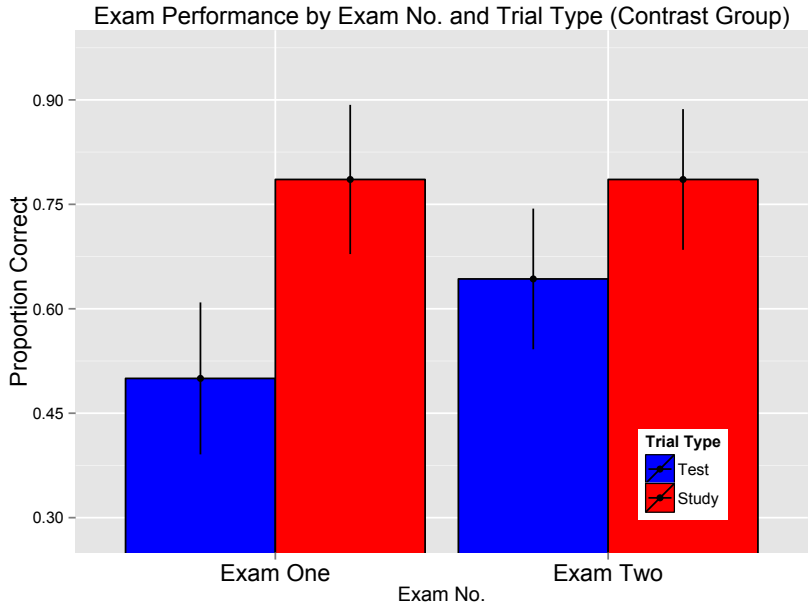


Figure 11. The effect of trial type for Contrast subjects with study benefit on Exam 1.

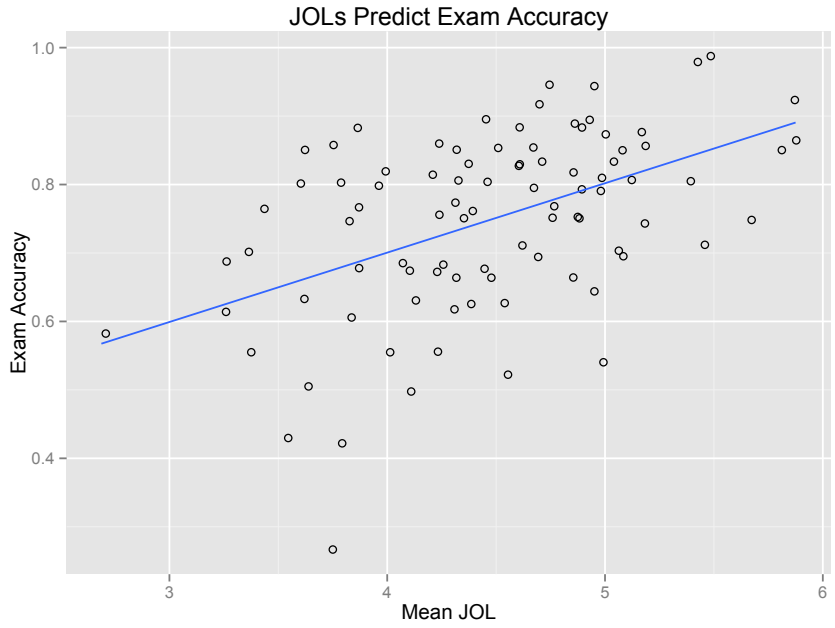


Figure 12. Relationship between confidence in future memory and performance in Statistics Experiment.

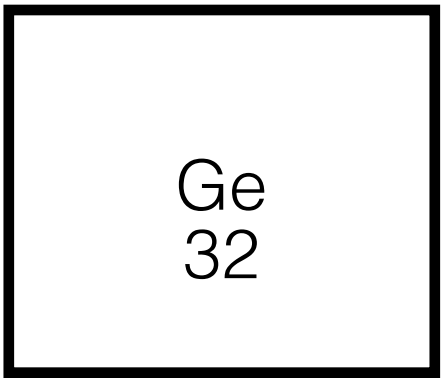


Figure 13. Elemental symbol study slide.

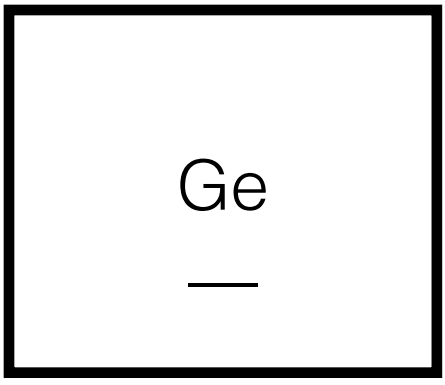


Figure 14. Elemental symbol test/exam slide.

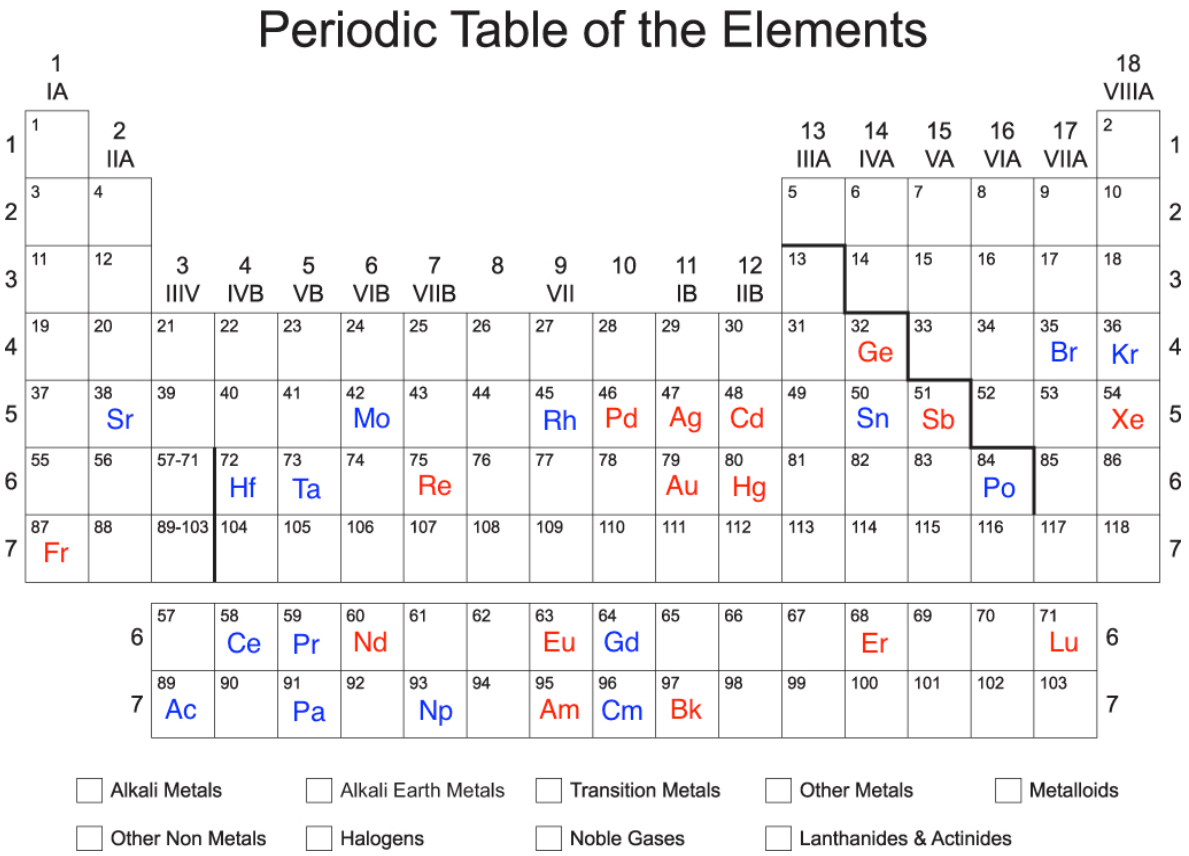


Figure 15. The periodic table used at pre-test and post-test (provided blank). Red symbols indicate elements used in List A list and blue symbols indicate elements used in List B.

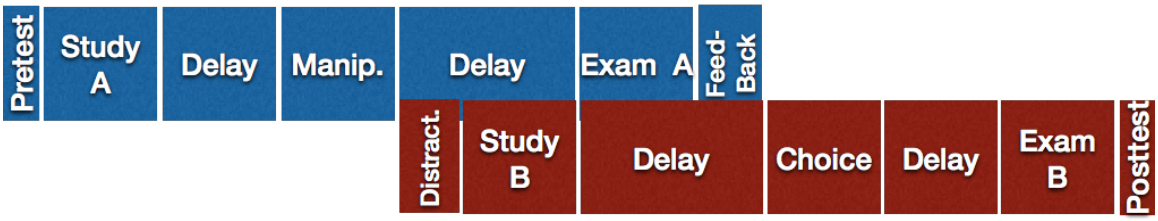


Figure 16. The general procedure for the Chemistry Experiment. Time progress from left to right. Overlap indicates that delays for one list's memory are used to learn or assess the other list.

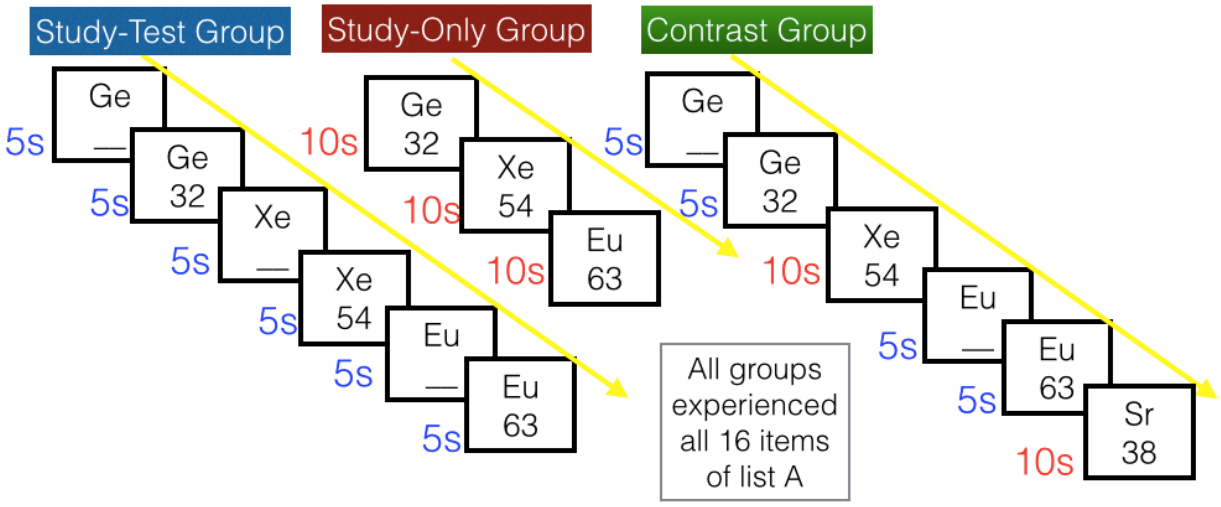


Figure 17. The strategy manipulation phase of the Chemistry Experiment.

Elements	Your Score	Elements	Your Score
Cd	CORRECT	Lu	X
Au	X	Er	CORRECT
Ge	CORRECT	Hg	X
Eu	X	Sb	CORRECT
Re	X	Ag	CORRECT
Am	CORRECT	Bk	X
Fr	CORRECT	Nd	X
Xe	X	Pd	CORRECT
		Overall Score:	8 /16

Restudied Elements	Your Score	Tested Elements	Your Score
Cd	CORRECT	Lu	X
Au	X	Er	CORRECT
Ge	CORRECT	Hg	X
Eu	X	Sb	CORRECT
Re	X	Ag	CORRECT
Am	CORRECT	Bk	X
Fr	CORRECT	Nd	X
Xe	X	Pd	CORRECT
Restudied Score	4 / 8	**Tested Score**	4/8

Figure 18. Post-Exam A feedback screen for the Study-Only group and Test-Only group (top) and Contrast group (bottom). Note the column headers and summary scores.

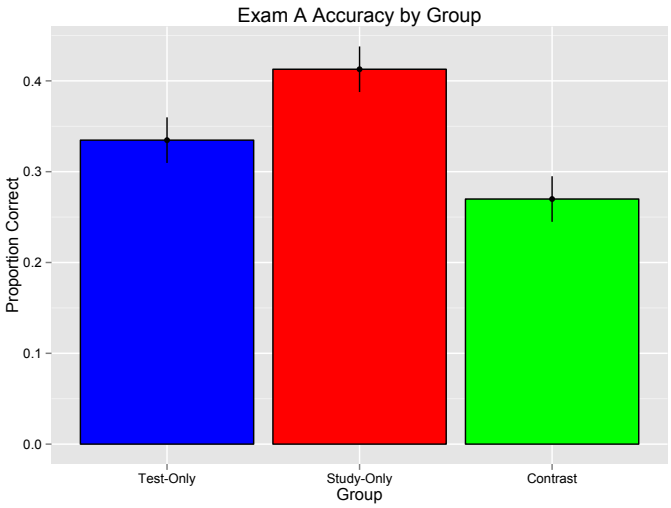


Figure 19. Exam A performance between groups (proportion correct).

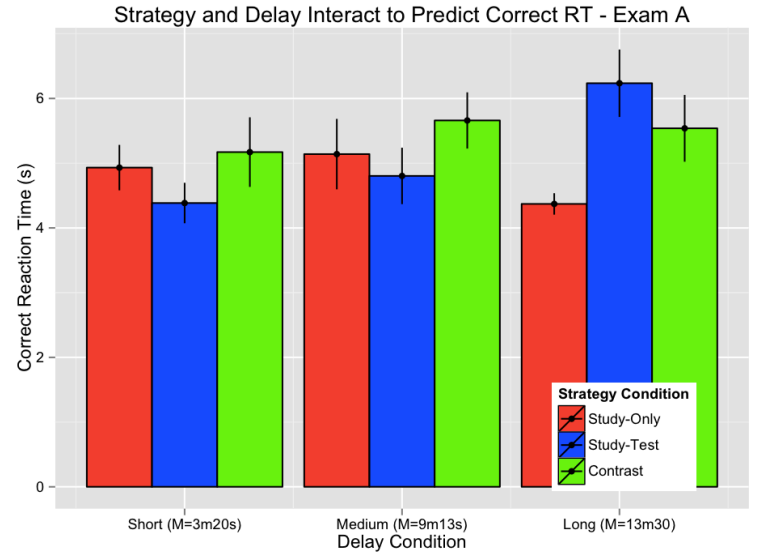


Figure 20. Exam A performance between groups (reaction time; discrete delay).

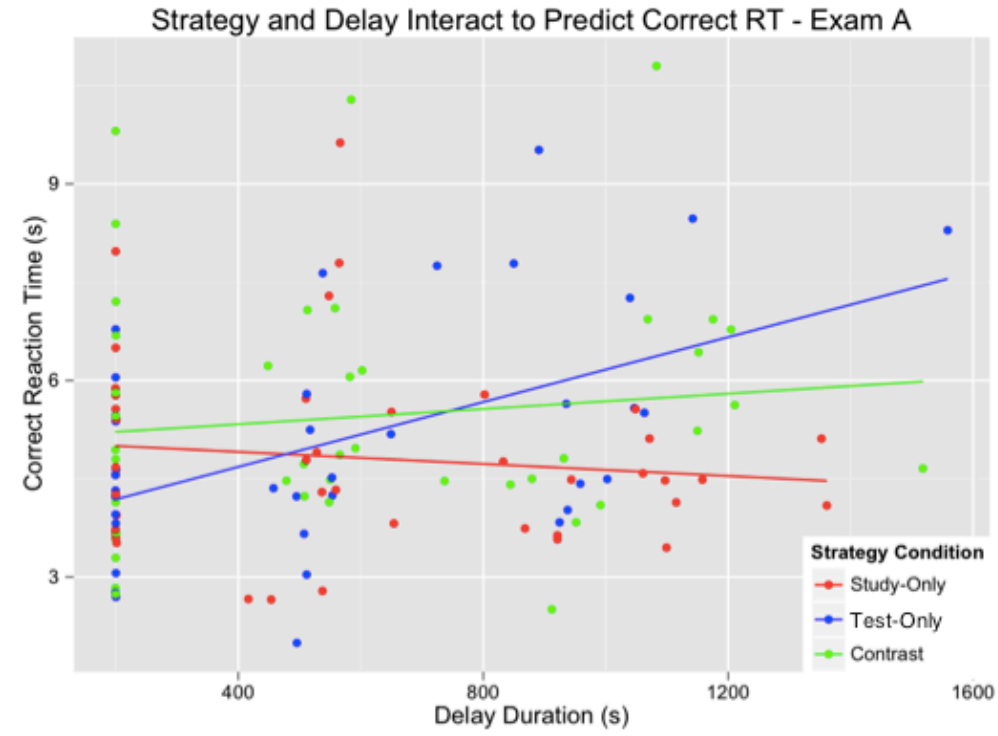


Figure 21. Exam A performance between groups (reaction time; continuous delay).

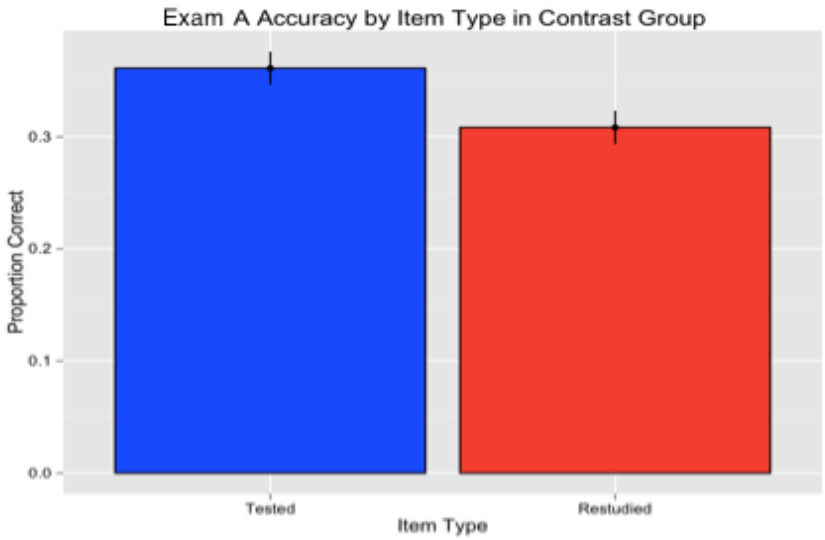


Figure 22. The effect of trial type among the Contrast group, Exam A

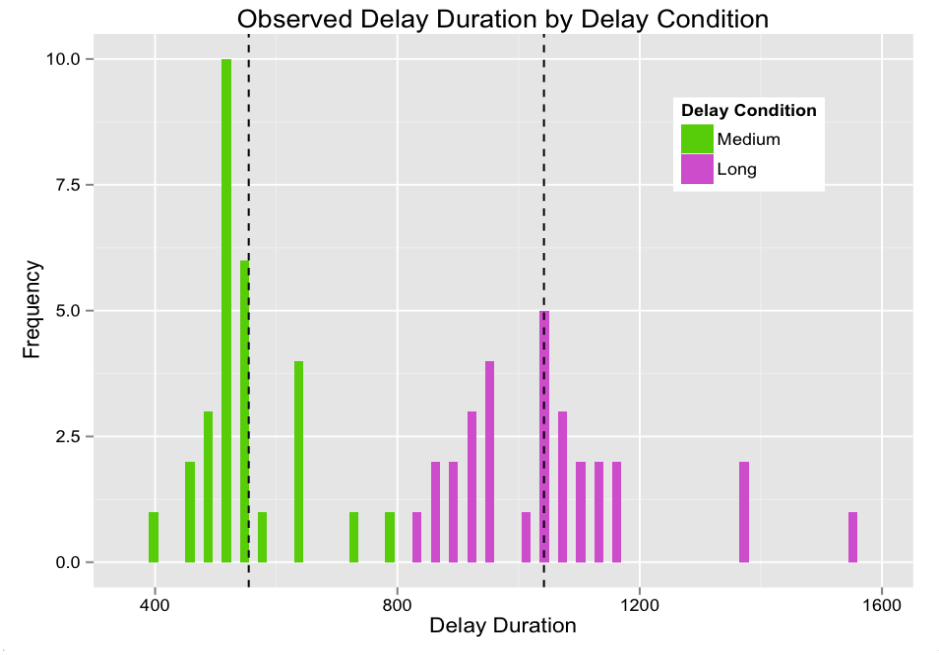


Figure 23. Variation of delay within conditions (note that Short delay was constant at 200 s, not shown).

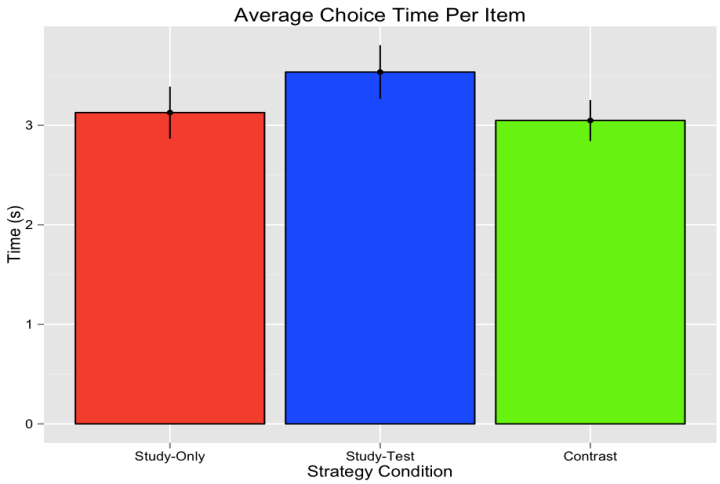


Figure 24. "Choice" latency by group.

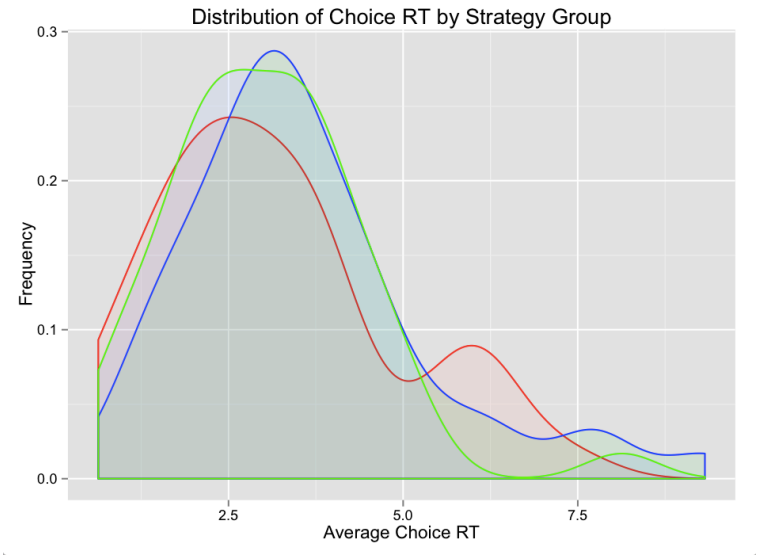


Figure 25. Distribution of "Choice" latency between Study-Only (red), Test-Only (blue), and Contrast (green) groups

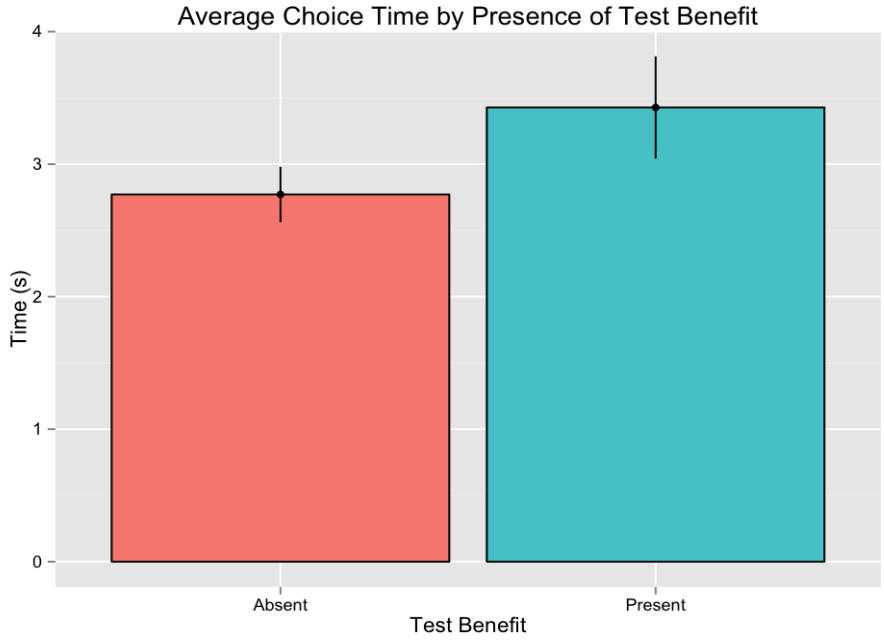


Figure 26. "Choice" latency by presence or absence of testing effect.

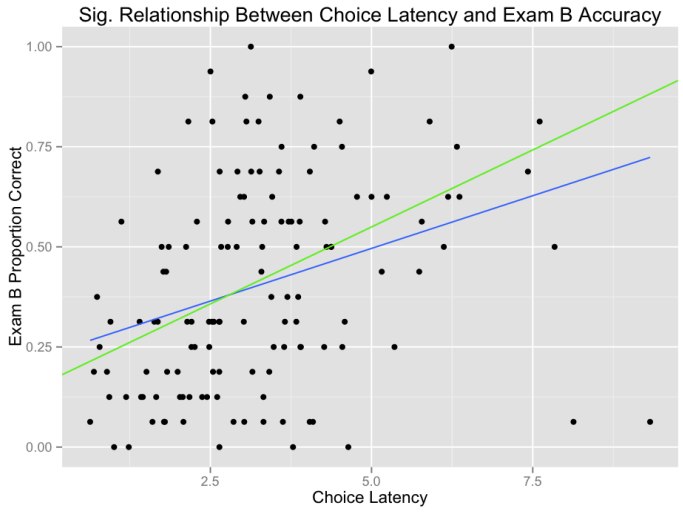


Figure 27. Relationship between "Choice" latency and Exam B accuracy. Regression lines shown with (blue) and without outliers (green).

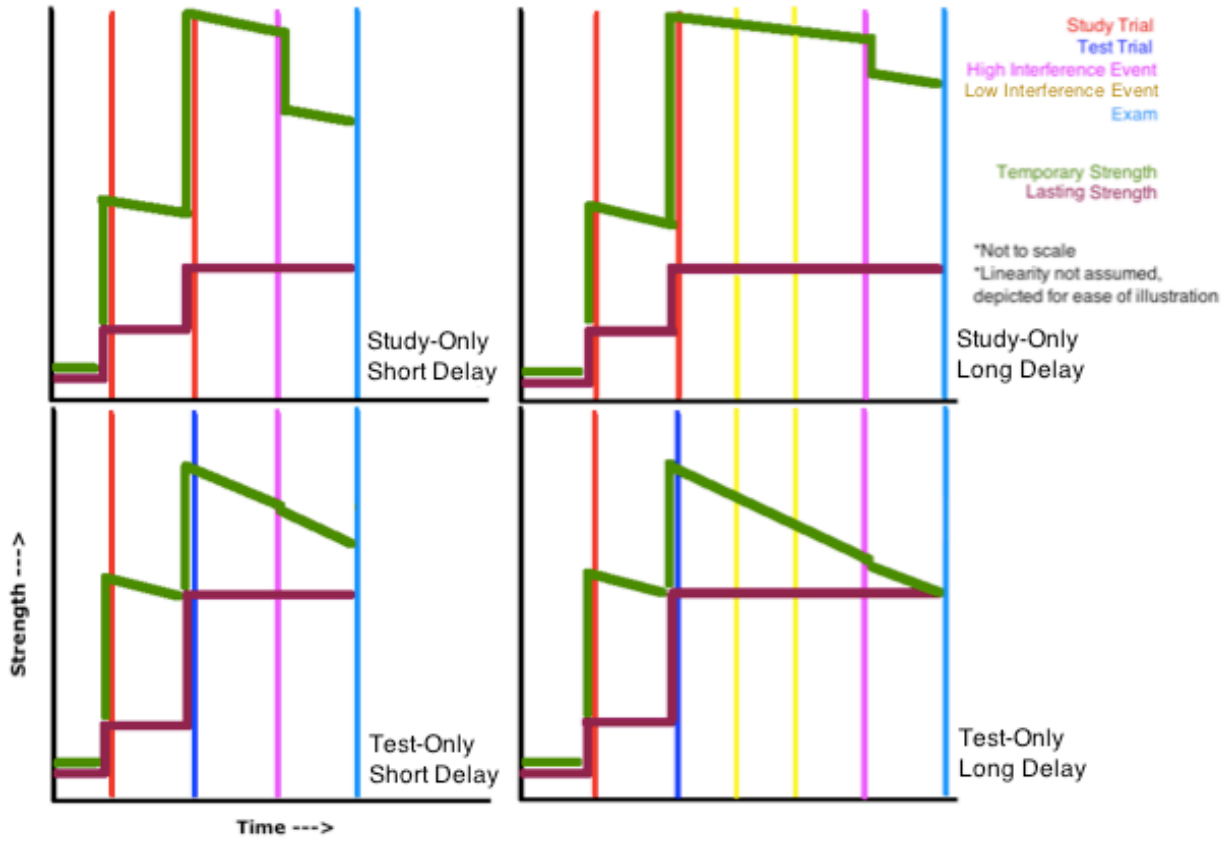


Figure 28. Illustration depicting theoretical processes underlying interaction between group and delay. For each panel, time moves from left to right, and increases in strength (Y-axis) correspond to faster reaction time. High interference event refers to encoding of List B. Low interference event refers to tasks intervening between lists. For clarity, retrieval strength has been referred to as temporary strength, and storage strength referred to as lasting strength.