THE TESTING EFFECT, INDIVIDUAL DIFFERENCES, AND TRANSFER: AN INVESTIGATION OF LEARNING STRATEGIES USING EDUCATIONAL MATERIALS

A Dissertation

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

By

Joseph W. Pirozzolo

May, 2019

THE TESTING EFFECT, INDIVIDUAL DIFFERENCES, AND TRANSFER: AN INVESTIGATION OF LEARNING STRATEGIES USING EDUCATIONAL MATERIALS

Joseph W. Pirozzolo

APPROVED:

Donald J. Foss, Ph.D.
Committee Co-Chair

David J. Francis, Ph.D.
Committee Co-Chair

Elena L. Grigorenko, Ph.D.

Arturo E. Hernandez, Ph.D.

Sascha D. Hein, Ph.D.
College of Education

Antonio D. Tillis, Ph.D.
Dean, College of Liberal Arts and Social Sciences
Department of Hispanic Studies

THE TESTING EFFECT, INDIVIDUAL DIFFERENCES, AND TRANSFER: AN INVESTIGATION OF LEARNING STRATEGIES USING EDUCATIONAL MATERIALS

An Abstract of a Dissertation

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

By

Joseph W. Pirozzolo

May, 2019

ABSTRACT

The positive effect of testing memory has been well demonstrated in laboratory settings and there is now a growing body of supporting evidence in real educational environments. However, whether and under what conditions testing facilitates transfer of learning is still somewhat unclear. Individual differences in learning from tests have also not been extensively studied. The aim of the current study is to further investigate the limits of transfer of learning via testing and explore the role of key cognitive abilities (i.e., reading comprehension, reasoning ability, and working memory). To accomplish this goal, we use an instance in the subject of Biology where we believe that background knowledge (i.e., the components of nucleic acids) is necessary for understanding of a subsequent related concept (i.e., DNA transcription). In a within-subjects experimental design with data from 153 undergraduate students, we examined the effect of testing over background knowledge on performance on subsequent related information. Our study provides evidence of the positive effect of testing on not only exactly repeated test items (d= 1.01), but conceptually related questions (near transfer; d = .60) and questions about a subsequent related passage (far transfer; d = .21). We also report that testing influences pre-test score predictions, such that repeated testing is associated with increased pre-test confidence, while varied testing is not. Finally, we report that individual differences in cognitive ability do not interact with testing effects, but transfer performance is correlated with reasoning ability. Overall, we conclude that retrieval practice with cued recall questions is a highly effective strategy for learning complex educational materials.

Keywords: Testing effect, transfer, individual differences, metacognition

ACKNOWLEDGEMENTS

There are *several individuals* who played critical roles in the successful completion of this project. Here, I would like to acknowledge:

First, my doctoral advisors, Drs. Donald J. Foss and David J. Francis, who have supported my academic and professional development throughout my graduate education. All that I have accomplished in the last six-years would not have been possible without their enduring support and generosity.

My committee members, Drs. Elena Grigorenko, Arturo Hernandez, and Sascha Hein, for their constructive feedback which considerably strengthened this project.

My undergraduate research assistants, who played a key role in data collection and execution of our studies, especially Haelim Jeong—whose dedication to the project was unmatched and greatly appreciated.

Key people in the Psychology Department Administration, especially Dr. Suzanne Kieffer and Linda Canales—who graciously offered their time and resources to the project on multiple occasions.

Finally, Dr. Robert A. Bjork—who planted the seed for my interest in studying human learning during my visit to UCLA in 2012 and who assisted with the design of this project, in addition to his longtime support of my work and collaboration with Dr. Foss and Dr. Francis.

TABLE OF CONTENTS

LIST OF TABLES

DEDICATION

To Marisol and Sofia, who are simultaneously my inspiration and my biggest supporters. And to my mother, Priscilla, whose unending faith and positivity has been critical to my success and has made a lasting impact on multiple generations.

The Testing Effect, Individual Differences, and Transfer: An Investigation of Learning Strategies Using Educational Materials

The goal of formal education is to facilitate learning that sustains time and can be applied to future problems in related fields. Recently, there has been an influx of research on the mnemonic benefits of the "testing effect" and subsequently there has been a push to effectively apply this research to educational settings. The testing effect (also called retrieval practice; henceforth the testing effect and retrieval practice are used interchangeably) refers to the finding that information is better remembered over time when attempts to retrieve previously learned material are made (e.g., taking a test, answering questions, or another form of recall). Although, the vast majority of testing effect studies show sizable increases in retention of material on which the learner made a retrieval attempt, less is known about the effectiveness of retrieval practice when the outcome measure differs somewhat from the retrieval practice conditions or requires transfer of learning. It is not entirely clear whether and under which conditions the testing effect facilitates transfer of learning (also called transfer), especially when using real educational materials (those that would generally be used in a classroom). In addition, few attempts have been made in the existent literature to explore possible individual differences in learning and transfer via the testing effect. Here, we review the literature and report on a study designed to (a) investigate the effects of retrieval practice on transfer of learning, (b) evaluate whether individual differences in cognitive ability moderate the effect of retrieval practice and transfer of learning, and (c) examine the effect of retrieval practice on pre-test score predictions. The studies proposed

here use authentic educational materials that have been used in college courses and transfer is defined in a way that is particularly relevant to students and instructors.

The Testing Effect

For teachers and instructors, the general conception of tests or examinations is that they assess learning that has been acquired through instruction and study. Although the general public is becoming more aware of research on the retention benefits of retrieval, educators have historically disregarded the learning benefits of testing (Clark & Bjork, 2014). Although the earliest studies on the testing effect date back to the early 1900s (Gates, 1917; Roediger & Karpicke, 2006a; Spitzer, 1939), one of the most significant studies on testing was carried out by Tulving (1967). In Tulving's (1967) study, participants learned a word list using both study trials where participants attempted to learn the words on the list and test trials where participants attempted to recall the words they had studied. Participants were assigned to one of three conditions. In the repeated study condition participants studied the word list for three consecutive trials before testing. If S stands for study and T stands for test, participants in the repeated study group carried out the learning task as SSST for 24 trials (each S or T represents one trial). A repeated testing group was given one study trial followed by three consecutive test trials (STTT), and a standard learning condition alternated study and test trials throughout the experiment (STST). The results showed that by the 24th trial (a test trial for all participants) all three groups showed similar levels of recall performance, despite the repeated test group being exposed to one-third of the study trials that the repeated study group received. These results clearly indicated that a test is more than merely an assessment—tests can be meaningful learning experiences.

In a study using a nearly identical design to Tulving (1967), Karpicke and Roediger

(2007) demonstrated that the standard group (STST) may actually outperform the repeated study condition (SSST) when the final test is delayed (one-week). Increasing the retention interval, the time between exposure to learning material and the final test, seems to increase the effect size of retrieval practice. Manipulating the retention interval has been a common method for (a) examining the effectiveness of retrieval practice when different lengths of retention are required and (b) investigating the cognitive mechanisms of retrieval. Toppino and Cohen (2009) found a significant interaction between the retention interval and retrieval practice conditions. In their study, improvement of a testing condition over a restudy group depended on whether the retention interval was a few minutes or 48-hours. In two separate studies the testing group performed better than a restudy group at a retention interval of 48-hours, however, no difference between groups was found at retention intervals of two and five-minutes (Toppino & Cohen, 2009). The consistent finding of an interaction between testing and the retention interval suggests that retrieval of learned information prevents or slows the rate of forgetting. Bjork (1975) postulated that testing can be thought of as a memory modifier—that is, each time information is retrieved, the structure of that information is altered and strengthened. Under this framework, information that is frequently retrieved is more likely to be remembered and information that is not frequently retrieved is adaptively forgotten in order to make remembering more important information easier (Schacter, Norman, & Koutstaal, 1998).

Other explanations of the testing effect include those that attribute memorial benefits to elaboration or the creation of more detailed memories. Similar to Bjork's (1975) hypothesis, the *elaboration* theory holds that retrieval (as opposed to rereading or studying) creates the opportunity for additional memories and mental networks to be made, therefore

increasing the likelihood that information will be remembered at a later point (Carpenter, 2009; Eisenkraemer, Jaeger & Stein, 2013; Roediger & Butler, 2011). Another explanation for benefits of retrieval practice revolves around the amount of effort that is made during the learning process. The *effort* hypothesis argues that because it is likely that more effort is made when answering questions or attempting to recall information than when reading learning material, this additional effort is responsible for improved memory of information for which retrieval attempts have been made (Pyc & Rawson, 2009). A final common explanation for the effects of retrieval practice contends that performance on a final test is improved (at least partially) as a result of the similarity between the retrieval practice context and the final test context (Morris et al., 1977; Roediger, 1990; Roediger & Butler, 2011). This explanation is called the *transfer-appropriate* theory, as the logic follows that learning is more likely to translate to a final test when the conditions under which participants initially trained are more similar to those on which they are evaluated.

Recently, dozens of studies have demonstrated the benefits of retrieval practice in many different contexts (e.g., Rawson & Dunlosky, 2011). Most testing effect experiments feature the contrasting of conditions that require taking a memory test of learned information and conditions where rereading (sometimes combined with highlighting learning materials) or restudying is the primary learning strategy. Although other types of control conditions are used (e.g., concept mapping; Karpicke & Blunt, 2011), rereading or restudying is generally used because it best represents how most students prepare for exams (Bjork, Dunlosky, & Kornell, 2013). The majority of the literature on the benefits of retrieval practice have been controlled laboratory studies using word-lists and word-pair stimuli; however, studies have also shown strong effects of retrieval practice when using educational stimuli. More recently,

significant effects of retrieval practice have been applied in classrooms using educational materials. In the following sections studies demonstrating the positive effects of testing will be reviewed. We first discuss laboratory studies that use non-educational stimuli and then review studies that are more directly relevant to educational settings.

In an important investigation of the role of retrieval, Jacoby (1978) had participants learn word-pairs (e.g., foot-shoe), by either studying the intact pair or by attempting to retrieve or generate the right-hand word when given the cue and only a fragment of the right-hand word (e.g., foot- s_ _e). In addition, participants saw each word-pair either once or twice (repeated). If a word-pair was in a repeated condition, it was presented either immediately after the first exposure or after a delay with 20 intervening pairs (spaced). Results indicated a massive relative performance increase on word-pairs that were initially read intact and later generated after a 20-pair delay. These results suggest that generating information rather than studying it can benefit memory, especially after a delay. These results laid the foundation for a phenomenon similar to the testing effect called the generation effect. Testing effects have been found using cued-recall, where like Jacoby (1978) participants are asked to recall a part of material that has previously been studied (Estes, 1960; Kornell, Hays, & Bjork, 2009; McDaniel & Mason, 1985), and free-recall where participants must recall learned information without a cue (Hogan & Kintsch, 1971; Tulving, 1967; Zaromb & Roediger, 2010).

The testing effect with educational materials. Although many of the most famous testing effect studies use word-list and paired associate stimuli in free recall and cued recall paradigms, there is substantial evidence that the testing effect translates to educational materials, such as prose passages and learning units. In one study, participants learned

information from a prose passage and then either restudied the passage or took an open-book (where participants had access to notes and the textbook) or closed-book practice test (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). At a one-week delayed final test, results showed significant performance increases when participants had previously taken a practice test. Initial practice test performance was higher in open-book testing conditions than closed-book conditions, but this increase in performance was not maintained over the one-week retention interval. Furthermore, final test performance was highest in a group that took a closed-book practice test and received feedback. Roediger and Karpicke (2006b) showed that at delayed intervals (2-days and 1-week) testing produced better results than additional studying when using educational prose passages. Similar effects of testing have also been found using textbook passages and both multiple-choice and short-answer tests (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Kang, McDermott, & Roediger, 2007). Several other studies have shown testing effects using educational material (e.g., Butler & Roediger, 2007; Carpenter & Pashler, 2007; McDaniel & Fisher, 1991; Smith & Karpicke, 2014; Roediger & Karpicke, 2006b).

Education makes frequent use of multiple-choice question tests—a recognition memory task. Because multiple choice tests' lures (incorrect answers) are used to provide students the opportunity to select the correct answer, one fear is that retrieval practice will cause test takers to remember incorrect answers instead of the correct ones. This is a logical concern, as commonly used four-alternative multiple-choice questions have three incorrect answers and only one correct answer. As a result, students are exposed to many more lures than correct answers. Roediger and Marsh (2005) found that exposing participants who took practice tests to a greater number of incorrect multiple-choice lures increased the likelihood

that lures would be selected as answers at a later test. A positive effect of testing was still observed by Roediger and Marsh (2005), although increasing the number of lures led to smaller testing effects. Some studies, however, have found that exposure to competitive lures can actually have a facilitative effect, as additional retrieval attempts must be used to determine why a lure is incorrect (Little & Bjork, 2015). Corrective feedback is thought to be important for reducing the negative effects of multiple-choice testing (Butler & Roediger, 2008).

In addition to laboratory studies, researchers have also used more applied research designs to test the effectiveness of retrieval practice in the classroom. Bangert-Downs, Kulik, and Kulik (1991) conducted a meta-analysis of 35 classroom studies that examined the effects of testing versus additional studying. Bangert-Downs et al. found that 83% of the studies analyzed reported positive effects of testing. This analysis also suggests that an increase in the number of tests provides further retention benefits. Roediger, Agarwal, McDaniel, and McDermott (2011) conducted a study in a sixth-grade Social Studies class, in which students were quizzed over some topics but not others. In experiments 1 and 2 students received three multiple-choice quizzes on each chapter throughout the term. Results showed that on chapter and end-of-semester exams students performed better on material that appeared on quizzes than material that had not. Additionally, students performed better on questions that had appeared more frequently on quizzes. Testing effects were seen even at a substantial delay (1-2 months for some chapters) on the end-of-semester exam. These results show that the testing effect can be successfully applied in real educational settings and sizeable gains can be seen at relatively long retention intervals. Other studies have found similar effects of retrieval practice in real classrooms using young school-aged (Rohrer,

Taylor, & Sholar 2010), middle school (Lipko-Speed, Dunlosky & Rawson, 2014; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011, McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Metcalfe & Kornell, 2007) and college-aged populations (Bjork, Little, & Storm, 2014; Foss & Pirozzolo, 2017).

The preponderance of the studies discussed above suggest that (1) retrieval practice is a valuable learning experience, (2) the testing effect is evident across populations (adults and children), settings (laboratory and classroom), materials (word-lists, prose passages, and textbook chapters) and test formats (free-recall, multiple-choice, and short-answer) and (3) the testing effect is most robust when retrieval is delayed. Despite the amount of knowledge available on the testing effect, relatively little is known about transfer of learning from testing.

Transfer of Learning

Researchers define transfer in many different ways. Most studies that investigate transfer use the general definition that a task requires transfer when the learner must apply knowledge to a new problem. In their comprehensive review of research on transfer, Barnett and Ceci (2002) discuss studies that have been conducted on transfer over the last 100 years. Despite the abundance of writing on the topic, we still are sure of very little about transfer. One of the reasons for the lack of understanding of transfer is that few researchers have settled on a definition of transfer. Barnett and Ceci (2002) subsequently developed a taxonomy for transfer designed to provide at least some guidelines for research on a concept that has been elusive. In the past, researchers have thought of transfer as either near or far—near transfer representing less change from one context to the next and far transfer being characterized by more change. In their taxonomy Barnett and Ceci (2002) place transfer on a

continuum from near to far, justly characterizing the many ways in which tasks can vary in distance. Importantly, to evaluate transfer, the authors also define two major categories on which transfer occurs, content and context, and nine dimensions within those categories. In their definition, Barnett and Ceci (2002) explain that content describes "what transferred," including: the type of skill (principle or heuristic), how performance is evaluated (e.g., accuracy, use of the correct approach), and the demand on memory (e.g., recall, recognize, and execute). In this taxonomy, the context component consists of: the knowledge domain (e.g., math, science), physical context (e.g., location), temporal context (e.g., retention interval), functional context (e.g., academic, non-academic), social context (e.g., group or individual), and modality (e.g., written, oral). Within these context domains, transfer can vary from nearer (e.g., mouse vs. rat) to farther (e.g., science vs. art). The taxonomy developed by Barnett and Ceci (2002) allows researchers a framework with which to (a) conceptualize the degree of transfer present in two tasks and (b) design experiments on transfer.

Transfer may be one of the most important concepts in education (Roediger, 2007). If transfer were not achieved, a student would be forced to treat even similar problems as new experiences. Likewise, if knowledge and skills do not transfer, it is also likely that much of the time spent learning is wasted (Druckman & Bjork, 1994). Given this context, it could be said that the goal of education or learning is to transfer or apply previously acquired knowledge to create meaningful learning. Karpicke (2012) argues that meaningful learning "is thought to produce organized, coherent, and integrated mental models that allow people to make inferences and apply knowledge" (Karpicke, 2012, p. 160). Whereas meaningful learning is flexible and withstands time, rote-learning, characterized by studying and

memorization leads to forgetting and does not facilitate transfer (Mayer, 2008). Although much is still unclear, testing has been one learning method that has been shown to help facilitate transfer.

The testing effect and transfer. Most testing effect studies carried out in the past have featured final test items that are identically repeated from previous practice test items—therefore involving retention, not transfer (Pan & Rickard, 2018). While the finding that testing improves retention of explicitly learned material is important, at this stage in the study of testing we believe that it is most critical to understand the extent to which testing yields knowledge of information that is generalizable. As discussed above, transfer may be one of the most critical aspects of learning. To make use of Barnett and Ceci's taxonomy, testing effect studies have generally addressed transfer across (a) temporal contexts, (b) test formats, (c) and knowledge domains (Carpenter, 2012). Various retrieval practice studies have examined transfer across time (e.g., Carpenter, Pashler, & Cepeda, 2009; Toppino & Cohen, 2009), which can also be conceptualized as examining performance at different retention intervals. Transfer across test format has been evaluated in two main ways: asking different questions on the same topic and asking questions that require subjects to make an inference on the final test (Karpicke, 2012). Karpicke and Blunt (2011) had participants read science passages, and subsequently either engage in retrieval practice, create a concept map, or restudy passages. A final assessment tested participants on questions that came directly from the text (verbatim questions) and questions that required participants to make an inference. The results indicated that participants in the retrieval practice condition performed better than control on both types of questions.

Other efforts have been made to examine transfer from tests across formats. For

example, McDaniel, Thomas, Agarwal, McDermott, and Roediger (2013) quizzed middle-school students over some concepts but not others. Quiz items were designed to either give a definition and ask for the appropriate term (definition-response), or vice-versa, present the term and ask for the correct definition (term-response). On the final test, students performed better on information that was quizzed in either fashion than on non-quizzed material. Interestingly, students performed equally on term-response items even when quizzed by the corresponding definition-response item. The authors conclude that this finding represents successful transfer as participants showed improved performance (relative to no-quiz control questions) on quizzed items that appeared in a different format. Similarly, Foss and Pirozzolo (2017) manipulated the frequency of exams in a college Methods in Psychology course. A "standard class" received two long exams throughout the semester, while a "frequent class" received eight short exams. All exams were comprised of half multiple-choice and half short-answer items. The final test given at the end of the semester consisted of items that had appeared identically on a previous exam, flipped items— asking the same question as a previously tested item but in a different format (multiple-choice or short-answer)— and novel items. Results showed that, while the frequent class generally performed better on the final exam, both groups performed better on flipped than novel items. Similar results of "flipped" items have also been observed in other studies (Bjork et al., 2014; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014).

Other laboratory studies have found successful transfer from retrieval practice. Rohrer, Taylor, and Sholar (2010) found transfer from testing with fourth-grade students. In their study, Rohrer et al. found that participants who practiced retrieval of the names and location of fictional cities on a map more accurately identified the cities on a final test 24-

hours later than participants who were only allowed to restudy the map. In addition, participants who practiced retrieval of the cities performed better than control, regardless of whether the final test questions were identical to those previously practiced. In possibly the most educationally relevant study showing successful transfer from testing, Butler (2010) had participants read prose passages and then either restudy the passage (control condition), take the same practice test questions repeatedly (same test condition), or take a set of reworded practice test questions (variable test condition). In study 1a, Butler (2010) found that participants performed better on both factual and conceptual questions in a testing condition than the restudy condition. Final test questions in study 1, were exact repeats of previously taken practice test questions for participants in one of the two testing conditions, and therefore the results of study 1a tested the retention of information over a 1-week delay. The final test in study 1b consisted of questions that required an inference to be made (thus defining near transfer in the context of the study). Results from study 1b indicated that participants also performed better in a testing condition (relative to restudy) on factual and conceptual final test questions that required an inference. No difference in performance was observed between the two testing groups (same test vs. variable test) in either of study 1a and 1b. In study 2, Butler (2010) demonstrated a positive effect of testing even when comparing same test group performance to a control condition where participants were allowed to study isolated sentences where test relevant information was contained. Possibly most interestingly, Butler showed in experiment 3 that participants performed better in the testing condition than in the control condition on questions that required inferences to be made to answer questions about another knowledge domain (thus defining "far transfer"). For example, questions about the construction of airplane wings were determined to be related questions from a passage on

the anatomy of bats, but from another knowledge domain. The results of Butler (2010) suggest that testing improves retention of material over time and produce superior ability to make inferences within and across domains.

Contrary to some of the results previously discussed, Wooldridge, Bugg, McDaniel and Liu (2014) conducted a study using authentic educational materials in which they did not observe transfer effects from testing. Wooldridge et al. had laboratory participants read a college Biology chapter. After initial exposure to the reading material, participants either reread and highlighted the chapter or took two practice tests on the material. Two days later all participants returned to take a final test. For participants in a testing condition, practice tests were constructed such that they either consisted of items that would be repeated identically on the final test (repeated test condition) or topically related items—those that were from the same chapter section as items that would eventually be on the final assessment (related test). On the final test participants in the repeated test condition, but not in the related test condition, performed better than restudy control. The authors note that the lack of a transfer effect in this study may be due to the fact that the materials came directly from the textbook publisher and were not as strictly controlled as the materials used in other studies reporting successful transfer from tests (e.g., Butler, 2010; Carpenter & Kelly, 2012). Using a similar design, Pirozzolo and Foss (2017) replicated the results of Wooldridge et al., finding a testing effect for identically repeated final test items, but no superior performance of testing for conceptually related items. In another study investigating retrieval practice and transfer, Tran, Rohrer and Pashler (2014) exposed participants to a series of premises or conditional statements. Participants then either reread the premises or took fill-in-the-blank practice tests. Results indicated that taking tests improved final test performance but restudy and testing

groups did not differ in performance on questions that required deductive inferences. The authors speculate that to achieve transfer from testing, learning materials or tests must require challenging retrieval and foster the connection between initial questions and the desired inference or target material.

In the most comprehensive study of the effect of testing on transfer to date, Pan and Rickard (2018) carried out a meta-analysis of 192 effect sizes of testing on transfer present in 67 papers—an analysis containing data from over 10,000 participants. Pan and Rickard were interested in both reporting an average effect size of testing on transfer and identifying important factors that may moderate these effects, such as features of the testing procedures and final test format. Results of the meta-analysis showed an average effect size of $d = .40$ for all effect sizes reported in the analysis. Additionally, Pan and Rickard reported that effect sizes were largest when the transfer test (a) represented the same material in different format (e.g., cued-recall to multiple choice), (b) consisted of an application of previously learned material (e.g., application final test questions), (c) involved making medical diagnoses, (d) or used related word-cues. The effect of testing on transfer was weakest when the transfer test consisted of rearranged stimulus-response items (e.g., term-definition facts), or involved untested material (seen during the initial study phase) and worked examples. Furthermore, Pan and Rickard reported that there are potentially three key predictors of the magnitude of testing-transfer effect sizes: response congruency, elaborated feedback, and initial test performance. (Response congruency in this context was defined as whether correct answers for practice test questions were retained as the correct answers for final test questions.) Taken together these results suggest that the effect of testing on transfer is most prominent when 1) the same correct answers that are used for initial tests are used on the final test, 2) correct

answer feedback is detailed (elaborate feedback), and 3) initial test performance is high.

To summarize, studies have shown that retrieval practice can facilitate transfer of learning, however, many of the studies that report successful transfer have defined transfer narrowly (e.g., Foss & Pirozzolo, 2017; McDaniel et al., 2013) or have used learning materials that differ from those that are likely to be used in secondary and post-secondary education (e.g., Rohrer et al., 2010). Some studies have also indicated that the relationship of the retrieval practice material and the target transfer material must be specifically known and carefully designed (Pan & Rickard, 2018; Tran et al., 2014; Wooldridge et al., 2014). Additionally, while Butler (2010) observed successful transfer, there was no effect of test variation. Studies on the benefits of interleaving learning material (e.g., Kornell & Bjork, 2008) suggest that variation in the learning or testing process creates challenges but leads to improved memory and performance. We contend that individual differences in cognitive abilities (i.e., reading comprehension, reasoning ability, and working memory) are an important factor that is frequently omitted from studies in the testing effect literature, and that these abilities may interact with learning conditions—especially when conditions are challenging.

Individual Differences and the Testing Effect

To date very few attempts have been made to examine whether individual differences exist in learning from retrieval practice. The most basic question of importance on the topic is: do all students benefit equally from retrieval practice? In one study investigating the relationship between the testing effect and individual differences, Bouwmeester and Verkoeijen (2011) had children (age 7-13) learn Deese-Roediger-McDermott (DRM) word lists. In the DRM paradigm, participants are presented with several words related to a

common, but not explicitly stated, theme. After initial exposure to DRM word lists, participants either restudied the list or made retrieval attempts. All participants then returned one-week later for a final test. A latent class analysis on final test performance identified three distinct groups—importantly, one group which did not evidence benefits of retrieval practice. Results indicated that participants who are low on gist trace processing, a general mental representation of past events, did not show the traditional retention benefits of retrieval practice. These results suggest that at least one cognitive construct moderates the effects of the testing, however, this finding could be limited to studies that use the DRM word list paradigm.

In a study specifically aimed to investigate the relationship between cognitive abilities and testing effects, Brewer and Unsworth (2012) reported larger testing effects for participants with lower episodic memory and general-fluid intelligence. This finding suggests that lower "ability" students could potentially make larger gains when using retrieval practice and close the gap with higher ability students. While the findings of Brewer and Unsworth (2012) seem promising in terms of working towards an individual based model of prescribing learning strategies, other studies have failed to replicate these results (Pan, Pashler, Potter & Rickard, 2015). In another study by Kern (2014), participants read History passages and then either reviewed the passages or took practice tests. Individual differences in language comprehension and background knowledge were also assessed. Although results on the final test did not show typical positive testing effects, language comprehension predicted performance on practice and final tests. Background knowledge also predicted performance in the testing groups, but not the review group. Other attempts have been made to investigate whether testing effects differ as a function of cognitive abilities and personality

characteristics (e.g., Bertilsson, Wiklund-Hörnqvist, Stenlund, & Jonsson, 2017). In a traditional testing effect study using word-pair stimuli, Bertilsson et al. (2017) failed to find both a correlation between final test performance and working memory, grit, or need for cognition (NFC) and interactions between these individual difference constructs with testing effects.

Education is not naïve to individual differences. Although, research on such constructs as learning-styles has generally failed to show strong empirical support (e.g., Pashler, McDaniel, Rohrer, & Bjork, 2008), studies have found significant effects of individualized instruction (e.g., Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007). However, the testing effect literature has not fully investigated the potential role of individual differences. The most complete attempts to examine the possible interaction of individual differences with the testing effect are represented here by Bouwmeester and Verkoeijen (2011) and Brewer and Unsworth (2012). These studies, however, do not consider transfer of learning. We contend that a more comprehensive investigation of retrieval practice and individual differences is needed, especially one that assesses transfer of learning in an educationally relevant context.

Effect of Testing on Metacognition

We are also interested in the effect that testing has on metacognitive judgments. In our review of the literature we have demonstrated that there is much evidence that retrieval practice is associated with improved retention and in some cases transfer of learning. But are students aware of the benefits of testing as they learn target material? Studies have shown that retrieval practice improves the accuracy of metacognitive judgments, reducing the "foresight bias"—or subjects' overconfidence in their level of knowledge when they have

relatively unlimited access to information during reading and studying (Soderstrom & Bjork, 2014). Along these lines, the predominate hypothesis is that using retrieval practice (with feedback) supplies students with information about their performance, which they can then use to develop future strategies for improving their learning. Some studies have shown that not only are metacognitive judgments more accurate when testing on the material has occurred, but judgments made under testing conditions are higher—indicating more confidence in participant responses (Barenberg & Dutke, 2018). In a somewhat inverse manner, some studies have also shown that the degree of confidence in responses can moderate the testing effect, such that testing effects are only evidenced when participant confidence is high (Zhang, Chen & Liu, 2018).

While some studies have reported that pre-test score predictions are both higher and more accurate under testing conditions, still others have reported that testing reduces learner confidence. Miller and Geraci (2016) executed an interesting study that sought to understand participant strategies when making performance judgments. Results revealed that participants who received retrieval practice lowered their score predictions, but participants who did not undergo retrieval practice but had access to a peer's retrieval practice scores increased their predictions (Miller & Geraci, 2016).

The variance in results in the research reported here is likely due to differences in study design, materials, and the difficulty of practice and final tests. Despite the abundance of studies on metacognitive judgments and the moderately large literature on the relationships between testing and confidence judgments, we are not aware of any studies that have compared confidence judgments between repeated testing and more varied testing groups. In the present study, we seek to understand the effect that testing (repeated and

varied) has on both judgments made about the practiced material, and those made on materials that requires the application of practiced material (transfer).

The Present Study

Retrieval practice has been shown to improve learning in a multitude of conditions using various materials. The positive effect of testing memory has been well-demonstrated in laboratory settings and there is now a growing body of supporting evidence in real educational environments. However, the extent to which testing benefits transfer to scenarios that depart from the initial testing conditions is still unclear. Studies that have shown successful transfer from testing have either defined transfer narrowly (e.g., McDaniel et al., 2013) or used materials that are not likely to be used in secondary and post-secondary education (e.g., Rohrer et al., 2010). The aim of the current study is to further investigate the limits of transfer of learning via testing and explore the role of key cognitive components in transfer from practice tests (i.e., science reasoning, reading comprehension, non-verbal reasoning, and working memory). No study to date has simultaneously examined transfer via retrieval practice and individual differences in cognitive ability.

To accomplish the goal of elucidating the effects of testing and the role of cognitive ability on transfer, we use an instance in Biology where experts believe that background knowledge (i.e., the components of nucleic acids) is necessary for understanding a future concept (i.e., DNA transcription). We employ this hypothesis in order to examine whether learning gains made from testing on background knowledge will transfer to the target material. Specifically, we contend that learning background knowledge to a higher level of mastery predicts greater transfer to the target material. As discussed in detail above, testing is believed to lead to greater mastery—therefore we believe that testing on background

information will aid transfer. We are also interested in how testing (as a learning method) influences metacognitive judgments and their accuracy. We contend that the proposed study contributes to the existing literature, as the current study provides a more complete and ecologically valid investigation of the factors that are believed to influence transfer of learning via testing than has previously appeared in the literature.

In this study we executed a randomized split-plot experimental design to examine the effects of testing on transfer of learning, its role in metacognitive judgments, and the potential moderating effects of cognitive ability. In this study we seek to evaluate the following questions:

1. Does testing improve performance on the final test?
2. Does testing improve transfer of learning?
3. Does the testing method (repeated or varied) moderate the testing effect and its effect on transfer of learning?
4. Does cognitive ability moderate the testing effect and the effect of testing on transfer?
5. Does testing influence pre-test score predictions and prediction accuracy?

We predict that in a scenario where the target material requires prior learning (our selected Biology passages), testing will aid transfer especially when a variety of test questions are used during training. We also expect that testing will lead to lower and more accurate pre-test score predictions. Finally, we predict that cognitive ability is positively related to the likelihood of achieving successful transfer, as we believe that students who are higher in cognitive ability are better equipped to make connections among learning experiences and therefore generalize their knowledge more effectively.

Method

Participants

One-hundred fifty-three undergraduate students (117 female) at the University of Houston participated in this study for partial course credit. Participants who completed all parts of the study were additionally rewarded with a $15 gift card. Participant age ranged from 18 to 43 with an average age of 21.50 years (*SD* = 4.05).

Design

The study utilized a 2 (testing method: repeated or varied) X 2 (condition: test, restudy) X 3 (question type; Passage A Repeated; Passage A Transfer; Passage B transfer) mixed design with one between (testing method) and two within-subjects (condition and question type) factors. Topics in Biology were randomly assigned to learning condition (test or restudy) and manipulated within-subjects, such that each participant was exposed to both a testing and restudy condition on separate (non-overlapping) Biology topics. Testing method (repeated or varied) was manipulated between subjects, such that in the testing condition participants assigned to repeated condition received more repetition of practice test questions, whereas participants in the varied condition received more variation in practice test questions[1]. Specifically, participants in the test-repetition (TR) condition answered the same eight practice test items three times, while those assigned to the test-variation (TV) condition answered questions on the same eight facts and concepts, but saw three variations of each item. Thus, this method ensured that participants in the TR and TV test conditions were

[1] Here we delineate between the categories of the testing method variable (repeated and variated) and the combination of learning condition and testing method variable which creates four possible conditions: test-repetition, test-variation, restudy-repetition, and restudy-variation. Since there was no measurable difference between the tasks in the restudy-repetition and restudy-variation conditions, we focus on comparing the test-repetition and test-variation groups and, henceforth, refer to these sub-groups as TR and TV respectively.

exposed to the same volume of practice test questions—but the TV condition experienced more variety in testing on the same material. On the final test, participants answered three types of questions for each Biology topic to which they were exposed. Final test questions were either a passage A question taken on Day-1 (passage A repeated), a novel question on a passage A concept or fact (passage A transfer), or a novel question on a passage B concept or fact (passage B transfer). All test questions used in the study were cued recall questions that could typically be answered in one sentence. Both practice test questions presented on Day-1 and final test questions taken on Day-2 were written and scored (using a premade answer key) by the experimenters.

To examine the effect of individual differences in cognitive ability on the effect of testing and transfer of learning, participants took three cognitive measures in-lab and self-reported and or provided a FERPA release for SAT and ACT test scores registered with the University of Houston. Cognitive measures administered in-lab included: Lawson's Classroom Test of Scientific Reasoning (Lawson, 1978; Lawson, Alkhoury, Benford, & Falconer, 2000), the Symmetry Span task testing working memory capacity (Foster, Shipstead & Harrison et al., 2015), and Raven's Advanced Progressive Matrices test—a test of non-verbal reasoning (Raven, 2009).

Materials

Biology topics. Three topics in Biology taken from *Campbell Biology: Concepts and Connections* (Reece & Campbell, 2011) were selected specifically because we believed they possessed a general hierarchical structure, such that mastery of certain concepts was necessary in order to comprehend more complex subsequent concepts. It is in this way that we operationalized a meaningful level of "far transfer." Selected initial passages read on

Day-1 (passage A) included *the components of nucleic acids*, *foodwebs in the ecosystem*, and *osmosis and diffusion.* Each of these primary topics was paired with a more complex subsequent passage (passage B): *transcription*, *energy flow throughout the ecosystem*, and *active and passive transport.* Passage B topics came from the same chapter as Passage A topics, often in the following chapter section. Each passage was between 800 and 1000 words. Pilot testing suggested that each passage could be easily read in 8-10 minutes.

Similar to procedures reported by Butler (2010), in each passage we identified four concepts. Concepts were defined as consisting of multiple key facts. After identifying four concepts within each passage, we wrote four cued-recall questions for each concept. Within each concept, we identified what we believed to be the most important fact and wrote four questions about each fact. In total, each passage contained four concepts and four facts (facts nested within concepts), with four questions written about each. When writing multiple questions about each concept and fact, we made significant efforts to write questions that were closely related, but questions often were asked in different ways or about different aspects of the same concept or fact (i.e., questions were not merely reworded). For example, two questions on the same concept, *similarities and differences between DNA and RNA,* are shown below:

> *Raul is working in the lab and finds a double-stranded nucleic acid that contains a deoxyribose sugar. What molecule has Raul found? (Answer: A DNA Molecule)*

What are two differences between DNA and RNA? (Answer: DNA is double stranded while RNA is single stranded; DNA contains a deoxyribose sugar while RNA contains a ribose sugar)

Fact-based questions were structured in a similar way. Two questions on the same fact were closely related but were not necessarily answerable with the same word or phrase. Below are two fact-based questions on *DNA complementary bases*:

What are the four complimentary nitrogenous bases that comprise a DNA molecule. (Answer: Adenine, Guanine, Cytosine, and Thymine)

DNA consists of compounds Adenine, Guanine, Cytosine, and Thymine, which are known as _____ . (Answer: Complementary bases)

Importantly, materials were counterbalanced to ensure equal assignment of Biology topics to learning condition, and to eliminate order effects due to the order of learning condition and Biology topic. Using 12 "tracks" of Biology topic and learning condition presentation, we ensured that each of the three Biology topics appeared equally in both the testing and restudy condition. We also designed the study such that each Biology topic and learning condition had an equal number of instances in the first and second (last) position during the learning phase of the study. A schematic of the counter-balancing procedures can be seen in appendix A.

Measures

Final test. The final test consisted of 64 cued-recall questions on the two Biology topics to which participants had been exposed over two days of study participation. Half of the questions (32) were on the Biology topic learned under the Restudy learning condition and the other 32 questions were on the Biology topic learned under the Testing learning condition. For each of the two Biology topics, eight final questions were repeated identically from the practice tests taken on Day-1 (passage A repeated), eight were novel questions about concepts and facts that were tested on Day-1 (passage A transfer), and 16 were novel questions about passage B concepts and facts (passage B transfer).

Reading comprehension. To examine the role of reading comprehension we received actual SAT and ACT admissions test scores through a student FERPA release signed during the informed consent process on Day-1 of the study. Specifically, we were interested in the SAT and ACT subtests that most directly measure reading comprehension: SAT Verbal, SAT Reading Test, and ACT Reading Test[2]. As can be seen in Table 1, not all students had SAT and ACT scores registered with the University. To deal with the issue of different patterns of test taking, we created a single reading comprehension variable, which was constructed by calculating the standard score (z-score) for each subtest and averaging across all reading related measures if participants had more than one. Some participants also took the same admissions exam multiple times. If a student had multiple observations for the same subtest, we standardized the average of their multiple scores. Of the 153 participants included in the final analysis, 48 (31.4%) were missing ACT and SAT data—therefore we

[2] In our sample two versions of the SAT admissions exam were taken by participants. In order to eliminate extraneous variance due to which version of the exam participants took, we treated the two tests as unique measures and standardized each separately.

were not able to calculate a reading comprehension score for these participants. Multiple imputation was therefore used in analyses involving the reading comprehension measure to avoid potential bias that might have resulted from exclusion of cases with missing data. The full description of procedures for multiple imputation are described in our data analytic section below.

Science reasoning. In order to examine the role of science reasoning abilities in learning from tests and transfer of learning, we again created a composite variable consisting of the standard scores of two measures: 1) Lawson's Classroom Test of Scientific Reasoning (LCTSR; Lawson, 1978; Lawson et al., 2000), administered in-lab following completion of Day-1 learning tasks and 2) the ACT science reasoning test, acquired through the same FERPA release used for reading comprehension data. The LCTSR is a 24-item multiple-choice test that measures baseline scientific reasoning ability (Jensen, McDaniel, Woodard, & Kummer, 2014). The score on the LCTSR is the total number of items correct, ranging from 0-24. All participants included in the final analysis had LCTSR scores, while 39 participants had ACT Science Reasoning scores. In order to improve the accuracy of our science reasoning construct and to deal with missing ACT data, we again standardized each participants LCTSR and ACT Science Reasoning score and averaged across the two scores (for those who had both) to create a single science reasoning score.

Non-verbal reasoning. The Raven's advanced progressive matrices (RAPM; E-prime 2.0.10) is a test of non-verbal reasoning. The RAPM in our study (Foster et al., 2015) consisted of the 18 odd-numbered items from the longer, 36-item RAPM (Raven, 2009; Raven, Raven, & Court, 1998). In the RAPM, subjects were shown a 3×3 grid of shapes with the bottom right shape missing. The shapes follow a logical pattern from left to right, and

from top to bottom. Participants were then asked to choose the shape the completes the pattern by selecting one of eight alternatives. Participants were given 10-minutes to complete the 18-item RAPM measure. Scores were calculated by summing the number of correct answers. For the purposes of analysis discussed below, non-verbal reasoning was centered at the mean.

Working memory. The Symmetry Span (Foster et al., 2015; E-prime 2.0.10) is a complex working memory span task. In the task, participants were first shown a figure created by several blocks and were asked whether the figure was symmetrical left versus right. After making the symmetry judgment, participants were then shown a single red box in a 4 x 4 grid of boxes and asked to remember the location of the box. Each presentation of a symmetry judgement and red box was considered a trial and a set of trials consisted of two to five trials. At the end of each set of trials, participants were asked to recall the correct location and order of the red boxes by clicking the grid on their computer screen. During the task participants also had to maintain 85% accuracy on the symmetry judgments. Each symmetry span session consisted of several sets of trials. At completion, two scores were calculated: the "absolute score," where points were only awarded if a participant correctly recalled all (two to five) red boxes, and the "partial score," where participants received a point for each individual correct box recall, independent of whether all boxes were correctly recalled. For the purposes of analysis, we used the partial score (centered at the grand mean) because it tends to discriminate better between low and high ability participants (Conway, Kane & Bunting et al., 2005).

Procedures

Day-1 procedures. Figure 1 shows the order of the specific Day-1 tasks for participants assigned to either the repeated or the varied testing method. Participants were told that they were participating in a study that examines college level science learning. Participants were randomly assigned to a testing method at arrival on Day-1 and assigned to a desktop computer where they would complete all study tasks. Demographics, learning tasks, and the LCTSR were completed using pre-programmed Qualtrics Software. The Symmetry Span and RAPM tasks were administered using E-Prime (E-prime 2.0.10) software on the same desktop computer. Following informed consent, participants were given the option to grant the researchers access to SAT and ACT data registered with the University of Houston using a FERPA release form.

Next participants completed a demographics questionnaire. Information reported in the demographics questionnaire included age, gender, university classification, and major area. We also asked participants whether they had taken eight specific high school and college level Biology courses. If students had taken a Biology course, we additionally asked them to report the grade they received in the course. Lastly, we asked participants to report their language history. Specifically, we were interested in their self-reported English proficiency and whether they speak languages other than English[3].

After completing the demographics questionnaire, participants began the learning phase of the study. Participants were first presented with passage A of the first Biology topic to which they had been assigned and were allowed 10 minutes to read the passage. Passages were presented as full-screen view PDF documents and participants were allowed to navigate

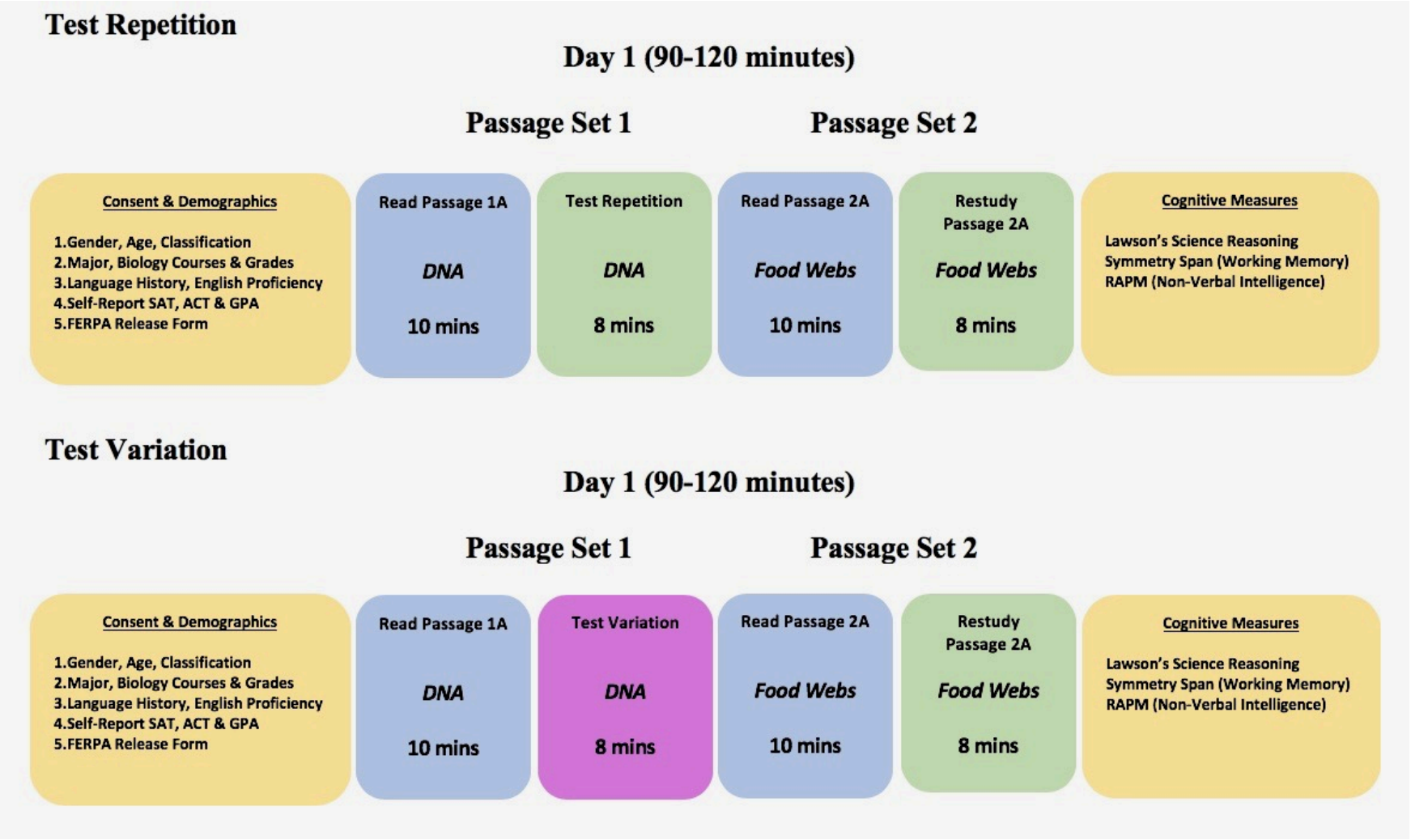
[3] Measuring language history and English proficiency is particularly important at the University of Houston, which has no ethnic majority and over 3000 international students from more than 50 different countries.

through the document at their own pace. After reading passage A on Biology topic 1 for 10 minutes, participants assigned to the restudy condition were immediately prompted to reread and restudy the same passage for an additional eight minutes. Alternatively, participants assigned to a testing condition were prompted to take 24 cued recall questions over passage A. Participants in the TR condition took the same eight questions three times, while those in the TV sub-group took three different questions on the same eight concepts and facts. In both testing conditions, participants were given correct answer feedback immediately following each individual submitted response.

After completing either a restudy or test condition for Biology topic 1, participants were presented with passage A of Biology topic 2. Again, participants were allowed 10 minutes to read the passage. After reading passage A of Biology topic 2, participants completed the opposite learning condition (test or restudy) from the one they had completed for Biology topic 1. After completing the second and final learning condition, participants were given a 5-minute break.

Following the break, participants completed the 24-multiple choice question LCTSR. Participants were given unlimited time to complete the LCTSR. Participants then completed the Symmetry Span and RAPM tasks. After completion of all cognitive measures, participants were dismissed for the day.

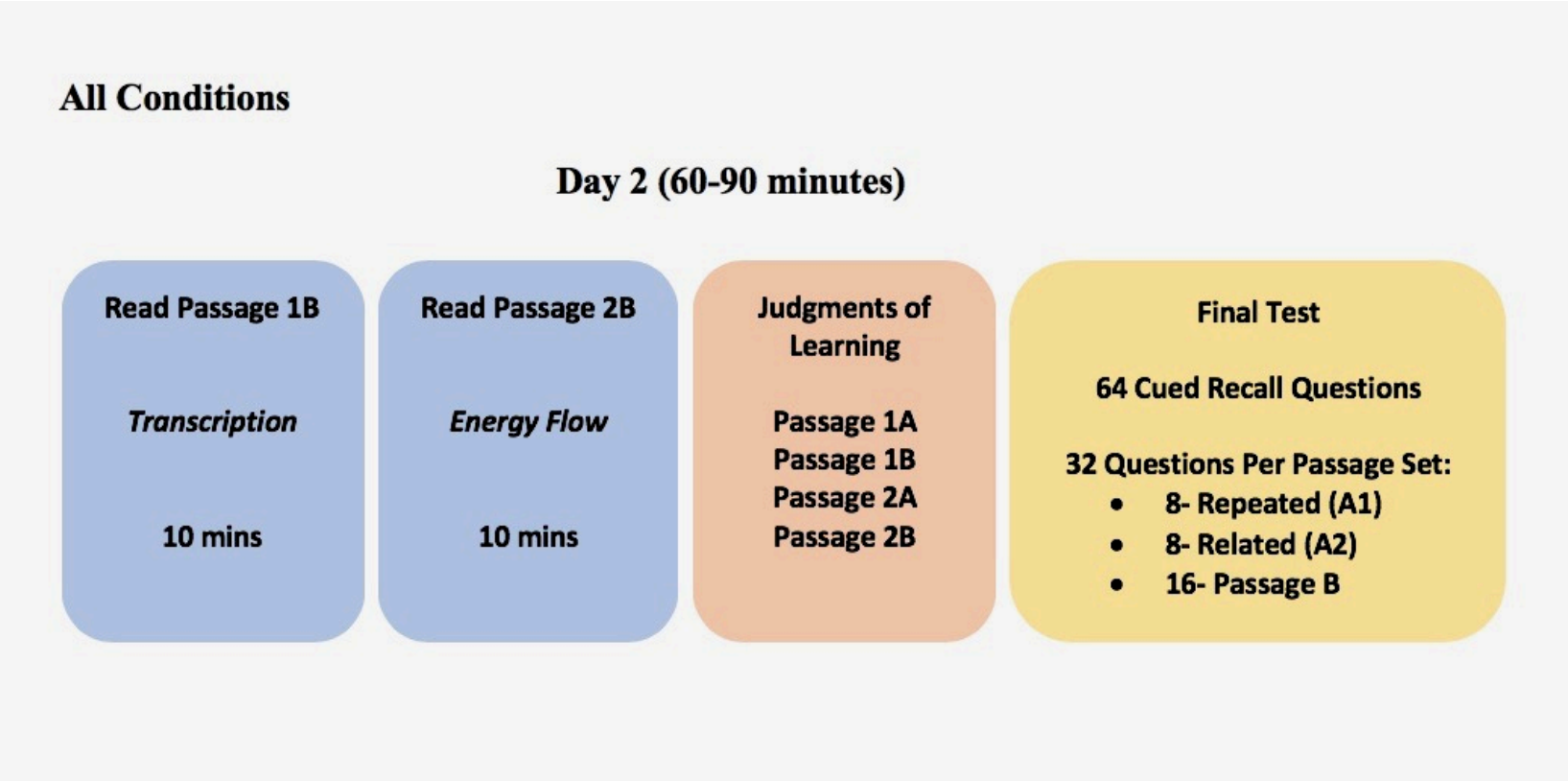
Figure 1. Schematic depicting study procedure on Day-1.



Day-2 learning tasks. Figure 2 shows the order of tasks participants completed on Day-2. After completing all tasks on Day-1, participants returned 24 hours later to read passage B for the respective Biology topics they learned on Day-1 and to take a final assessment over all material learned during the two-day study. Participants began Day-2 by presentation with passage B for each of the two Biology topics learned during Day-1. Participants were instructed that they would be reading and studying a passage that is closely related to a passage they read yesterday and asked to attempt to make a connection between the two passages. After reading instructions for the Day-2 tasks, participants were allowed 10-minutes to read passage B for each of the two Biology passages they were assigned to. Participants were presented with passage B in the same manner in which they received passage A.

Day-2 score prediction and final test. After reading passage B for both Biology topics on Day-2, participants were prompted to begin the final test. Before taking the final test, participants were asked to predict the percentage score they would get on questions from passage A and passage B for both of the Biology topics they learned (four total judgments). After making a score judgment for each passage, participants began the 64 cued recall question final test. Final test questions were grouped by topic, such that questions about Biology topic 1 (e.g., *components of nucleic acids* and *transcription*) were not interleaved with questions about Biology topic 2 (e.g., *foodwebs in the ecosystem* and *energy flow throughout the ecosystem*). Participants were allowed an unlimited amount of time to complete the final test. After completing the final test, participants were informed of the aims and hypotheses of the study and ultimately dismissed.

Figure 2. Schematic depicting study procedure on Day-2



Hypotheses

The following hypotheses were made of the results of our experiment:

H_1: Participants will perform better overall in a testing condition than in the restudy condition.

H_2: Participants will perform better on questions requiring transfer of learning (passage A transfer and passage B transfer) in a testing condition than in the restudy condition.

H_3: While the TR group will perform better than the TV group on passage A repeated items, the TV group will perform better on passage A transfer and passage B transfer items.

H_4: Participants will make lower and more accurate predictions of final test performance in a testing condition than in the restudy condition.

H_5: Cognitive ability will moderate the effect of testing, such that testing will aid participants with lower cognitive ability scores more than participants with higher cognitive ability scores.

H_6: Participants with higher scores on cognitive ability measures will perform disproportionately better on final test questions involving transfer (passage A transfer and passage B transfer) than participants with lower cognitive ability scores.

Data Analysis

Three types of analyses were carried out: 1) analysis of the effects of experimental manipulations on final test data (also called the original analysis), 2) analysis of the effects of experimental manipulations on pre-test score predictions, and 3) investigation of potential moderating influences of individual differences in cognitive ability on effects of testing on

learning and transfer. Our data analytic approach in this study calls for reporting the results of eight statistical models, each of which simultaneously tests multiple hypotheses. Because multiple hypotheses are tested in each of our analyses, we chose to control the false discovery rate (FDR; Benjamini & Hochberg, 1995; Maxwell & Delaney, 2004) at $p = .05$. Unless otherwise specified, all hypothesis tests reported in this manuscript with p values of .05 or lower are considered statistically significant findings after controlling the FDR at $p = .05$ for each model.

Final test data. Hypotheses 1-3 were tested using a 2 X (2 X 3) split-plot, repeated measures ANOVA (SAS 9.4; PROC MIXED). The outcome measure in this analysis was mean proportion correct on the three types of final test questions: passage A repeated (A1 items), passage A transfer (A2 items), and passage B transfer (passage B items) for both Biology topic 1 and 2 (six mean scores total). Variables in the analysis included: testing method (repeated or varied), learning condition (testing or restudy), and question type (passage A repeated, passage A transfer, or passage B transfer). H_1 was evaluated using the main effect of learning condition. H_2 was examined by the interaction of learning condition and question type and the results of specific contrasts testing mean differences between testing and restudy conditions on passage A transfer and passage B transfer items. H_3 was evaluated using a three-way interaction between learning condition, testing method, and question type and the results of three specific contrasts testing the mean difference between TR and TV groups on each of the three types of questions.

Score prediction data. To examine the effect of our experimental design features on pre-test score predictions and their accuracy (H_4), we carried out two repeated measures ANOVAs (SAS 9.4; PROC MIXED): one which used raw score predictions as the outcome

measure and one that used a prediction-observed difference score as the outcome—a measure of prediction accuracy. In the score prediction model, we regressed participant raw prediction scores on factors: learning condition, testing method, and Biology passage (passage A or passage B). We controlled for Biology content (DNA, Foodwebs, or Osmosis/Diffusion) as we believed that the perceived difficulty of the specific Biology content may have influenced predictions. In evaluating whether participants reported lower score predictions in a testing condition than in a restudy condition, we were primarily interested in the main effect of learning condition. Two-way interactions between learning condition, testing method, and Biology passage and the three-way interaction tested whether the effect of learning condition on score predictions was moderated by other design features.

To examine the effect of experimental manipulations on prediction accuracy we constructed a prediction-observed difference score. Difference scores were derived by subtracting participants' actual score on a set of materials (passage A or passage B test questions) from the predicted score. Using this procedure, positive values indicate overconfidence (i.e., a participant scored worse than expected) and negative values indicate under-confidence (i.e., a participant scored better than expected). We used a repeated measures ANOVA to examine the effects of learning condition, testing method, and Biology passage on prediction accuracy, while controlling for the type of Biology content. Again, we were primarily interested in the main effect of learning condition to evaluate whether participants made more accurate score predictions in a testing condition than in the restudy condition. Interactions between learning condition, testing method, and Biology passage were used to test whether the effect of learning condition was moderated by testing method and Biology passage.

Individual differences in cognitive ability. To accomplish our exploratory aim of investigating the potential moderating effects of individual differences in cognitive ability, we constructed a series of five repeated measures ANCOVAs (SAS 9.4; PROC MIXED). In Models 1-4 we took the original model described above and added one of the four cognitive ability constructs. Specifically, we added the main effect of the cognitive ability construct, two-way interactions of the construct with learning condition, testing method, question type, and the three-way interaction between the construct, learning condition, and question type. We followed this procedure, such that each of Models 1-4 represented an investigation of the effect of a singular cognitive ability construct (e.g., science reasoning) and its interaction with the experimentally manipulated variables. Finally, in Model 5 we combined all significant cognitive ability constructs and significant interactions to build a joint model. The joint model was used to investigate the extent to which the observed result in Models 1-4 were due to unique effects of specific cognitive ability constructs. H_5 was evaluated using the interaction between the constructs and learning condition, while H_6 was tested by looking at the interaction between cognitive ability constructs and question type. Although the hypotheses for each singular construct are examined in Models 1-4 respectively, Model 5 represents the most complete and ecologically valid test of our hypotheses because it controls for correlations among the individual difference measures.

Treatment of missing data. We observed missing data on the reading comprehension measure for 48 participants in our study (out of 153; 31.3%). As an alternative to omitting these participants from the individual differences in cognitive ability analyses, we performed multiple imputation for the analyses involving the reading comprehension measure. Multiple imputation involves three steps: step 1 provides multiple plausible values for the missing

information for each participant with missing data, step 2 involves fitting the analytic model to each of the data sets created through the imputation process, and step three involves combining the information across the separate analyses into a single estimate, standard error, and inference that takes into account the results from the separate analyses. With those steps in mind, we used SAS PROC MI (SAS 9.4) to produce 20 reading comprehension scores for each of the participants who were missing the score, leading to 20 separate datasets. An estimate of reading comprehension was obtained by fitting an imputation regression model using testing method (the only between-subjects manipulation) and each of the remaining cognitive ability constructs (science reasoning, non-verbal reasoning, and working memory) as predictors of the missing reading comprehension score. This process produced 20 datasets, which were then analyzed using the PROC MIXED model described above. The results from each of these models were output to a separate dataset of estimates and standard errors. Following procedures for pooling the results of each our 20 models outlined by Rubin (1987), SAS PROC MI ANALYZE was used to pool the parameter estimates for each of our effects across the 20 models. Methods for pooling the results of ANOVAs put forth by van Ginkel and Kroonenberg (2014) were used to produce *F* statistics and *p* values for each of our categorical predictors. All of the other individual difference variables had complete data. Thus, multiple imputation was only used for analyzing models involving reading comprehension.

Results

Demographics and Descriptive Statistics

Given that this study was executed using college Biology materials and that some participants may have been more familiar with the topics used in the study, we asked

participants to report their undergraduate major and experience with high school and college biology courses. The 153 participants whose data were used the final analysis included 86 Psychology majors, 31 other liberal arts majors (e.g., English), 15 Biology majors, 14 other STEM majors (e.g., Engineering), and 4 participants who had not yet declared a major. We asked participants whether they had taken eight specific high school and college Biology courses and to report the grade they received in the course, if taken. Participants in the study reported having taken an average of 2.25 (*SD* = 1.24) of the selected biology courses with a mean GPA of 3.23 (*SD* = .60). We also asked participants about their language history. Sixty participants reported that their first language was something other than English (participants reported 15 different first languages). We then asked participants to rate their proficiency in each language that they currently use on a scale 1 to 7, with higher scores indicating higher proficiency. In using this procedure, we were primarily interested in the level of English proficiency for participants who reported a language other than English as their first language. The 60 participants who reported a non-English first language had an average English proficiency of 6.57 (SD = .79).

Table 1 shows the descriptive statistics for cognitive ability and other measures collected in the study by testing method. Separate t-tests were carried out to test for mean differences between the repeated and varied groups on each of the measures listed in Table 1. No significant differences between repeated and varied groups were observed on any of the 12 continuous measures. Chi squared tests also revealed that gender and university major had comparable distributions in the repeated and varied groups. Table 2 shows the means and standard deviations for final test data by each of our three manipulated variables: learning condition, testing method, and question type.

Table 1

Demographics by Testing Method

	Repeated			Varied		
	n	*M*	*SD*	*n*	*M*	*SD*
Age	74	21.3	3.41	79	21.68	4.59
Biology Courses	73	2.21	1.29	72	2.29	1.2
Biology Courses GPA	73	3.22	0.61	72	3.23	0.59
English Proficiency	74	6.8	0.55	79	6.82	0.55
Symmetry Span Absolute	74	18.41	8.39	79	17.33	9.15
Symmetry Span Partial	74	27.59	7.44	79	26.96	7.99
RAPM	74	9.89	3.12	79	9.25	3.01
Lawson's	74	0.57	0.2	79	0.54	0.19
SAT Verbal	25	498.2	104.6	20	488.5	68.37
SAT Reading Test	21	29.64	4.47	21	30.48	3.6
ACT Reading	21	23.33	5.76	18	25.86	5.13
ACT Science Reasoning	21	21.98	5.11	18	24.08	3.67

Note: SAT Verbal and SAT Reading Test indicate different versions of the Reading Component of the SAT exam. A composite "reading comprehension" score was derived by taking the standard score of the SAT and ACT reading components and averaging across them for participants who had multiple scores. n = sample size; M = mean score; SD = standard deviation; RAPM = Raven's Advanced Progressive Matrices; Lawson's = Lawson's Classroom Test of Scientific Reasoning.

Table 2
Final Test Performance by Learning Condition, Testing Method and Question Type

	M	*SD*
Learning Condition		
Restudy	0.50	0.26
Test	0.66	0.26
Testing Method		
Repeated	0.58	0.28
Varied	0.58	0.26
Question Type		
A1	0.66	0.27
A2	0.60	0.27
B	0.48	0.24

Note: M = Mean; SD = Standard Deviation; A1 = Passage A Repeated; A2 = Passage A Transfer; B = Passage B Transfer.

Final Test Data

A 2 X (2 X 3) split-plot, repeated measures ANOVA revealed significant main effects of learning condition, $F(1,151) = 108.33$, $p < .001$, $d = .70$, and question type, $F(2,151) = 94.30$, $p < .001$. Testing method, $F(1,151) = .28$, $p = .60$, $d = .07$, was not a statistically significant predictor in the model, as participants in the repeated testing condition (adjusted mean = .59; henceforth LS Mean) performed comparably to participants in the varied testing condition (LS Mean = .60), averaging across question types and learning conditions. However, testing method interacted with question type (see below). Overall, participants performed better in a testing condition (LS Mean = .67) than in the restudy condition (LS Mean = .52), although the magnitude of this difference also varied across question type (see below). Participants performed better on A1 (repeated material; LS Mean = .66) than A2 (related material; LS Mean = .63) items, $F(1,151) = 8.18$, $p = .005$, $d = .16$., and better on A2 than passage B items (LS Mean = .51), $F(1,151) = 116.81$, $p < .001$, $d = .58$.

We observed significant two-way interactions between learning condition (i.e., testing versus restudy) and question type, $F(2,151) = 50.42$, $p < .001$, and between testing method (i.e., repeated versus varied) and question type, $F(2,151) = 4.83$, $p = .01$. The two-way interaction between learning condition and testing method, $F(1,151) = 1.07$, $p = .30$, and the three-way interaction between learning condition, testing method, and question type, $F(2,151) = 1.04$, $p = .36$, were not significant.

To better understand the two-way interaction involving learning condition, specific simple effects contrasts were used to examine the effects of testing versus restudy within question type. These revealed that participants performed better in the testing condition than in the restudy condition on A1 items (test LS Mean = .78 ; restudy LS Mean = .54), $F(1,151) = 149.87$, $p < .001$, $d = 1.01$, A2 items (test LS Mean = .70; restudy LS Mean = .55), $F(1,151) = 49.25$, $p < .001$, $d = .60$, and passage B items (test LS Mean = .53; restudy LS Mean = .48), $F(1,151) = 10.61$, $p = .001$, $d = .21$. Figure 3 shows learning condition performance by question type. By examining Figure 3 it is apparent that, while participants performed significantly better in the testing condition on each of the three types of questions, the magnitude of the testing effect varies across question type, with the effect being larger on A1 items than on questions involving transfer (A2 and passage B items).

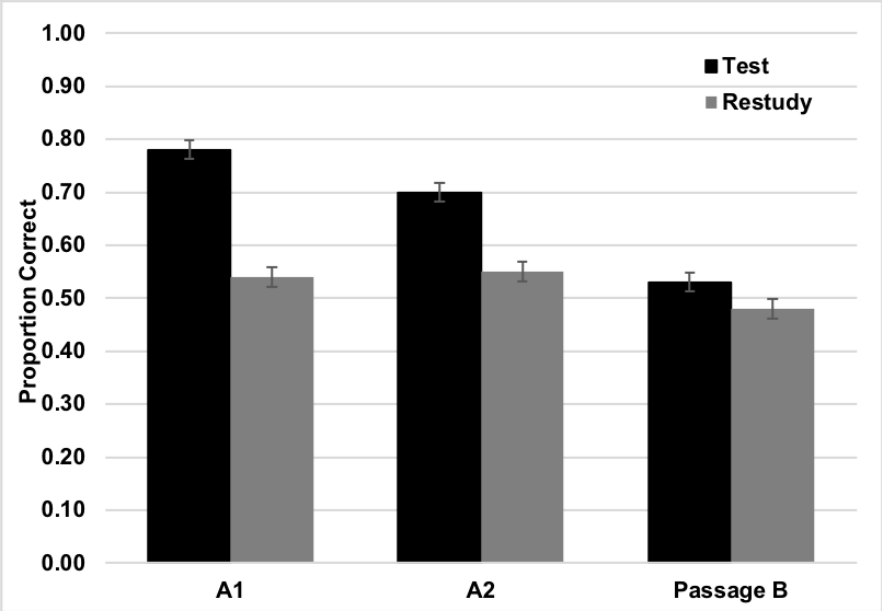


Figure 3. Learning condition as a function of question type.

Figure 4 shows the disordinal interaction between testing method and question type. To better understand the significant interaction between testing method and question type, we constructed two interaction contrasts and three simple effects contrasts. An interaction contrast testing the difference between A1 and A2 items for repeated and varied groups, $F(1,151) = 8.17$, $p = .005$, was significant. We also observed a significant interaction contrast comparing the A2 and B items difference between repeated and varied groups, $F(1,151) = 5.55$, $p = .02$. Simple effects contrasts showed that repeated and varied conditions did not differ on A1 questions (repeated LS Mean = .66; varied LS Mean = .65), $F(1,151) = .12$, $p = .73$, $d = .04$, A2 (repeated LS Mean = .60; varied LS Mean = .65), $F(1,151) = 3.27$, $p = .07$, $d = -.20$, or passage B items (repeated LS Mean = .51 ; varied LS Mean = .51), $F(1,151) = .00$, $p = .97$, $d = .01$. These effects are in the hypothesized direction for transfer items on passage A, with varied testing outperforming repeated testing, but the difference is not statistically significant. Taken together, these results indicate that, while the repeated and

varied groups did not differ on any one of the three item categories, testing method moderates the effect of question type. The specific interaction we observed between testing method and question type suggests that varied testing yields comparable performance on both A1 and A2 items, whereas repeated testing is associated with a non-trivial decrease in A2 (near transfer) performance relative to A1. Testing method does not seem to impact far transfer performance, as both the repeated and varied groups comparably on B items.

Although the three-way interaction between condition, testing method, and question type was not significant, we sought to directly test for differences in performance between participants assigned to the TR and TV conditions in a way that is similar to our reporting above in the two-way interaction between testing method and question type. Interaction contrasts revealed a significant difference between A1 and A2 means between the TR and TV groups, $F(1,151) = 7.86$, $p = .006$. However, the parallel interaction contrast comparing the difference between A2 and B means for the TR and TV groups was not significant, $F(1,151) = 3.21$, $p = .08$. We alo constructed simple effects contrasts testing mean differences between the TR and TV groups on each of the three types of questions. Our contrasts revealed that the TR and TV groups did not differ on A1 (TR LS Mean = .80; TV LS Mean = .76), $F(1,151) = 1.43$, $p = .23$, $d = .20$, A2 (TR LS Mean = .68; TV LS Mean = .73), $F(1,151) = 1.85$, $p = .18$, $d = .21$, or passage B items (TR LS Mean = .53; TV LS Mean = .53) , $F(1,151) = .03$, $p = .87$, $d = .03$. Similar to results reported on the two-way interaction between testing method and question type, here we observed that the TR and TV groups do not show the same difference in their performance on A1 and A2 items. Again, despite the lack of significant mean differences between the TR and TV groups on the three type of final exam question, it should be noted that the trend in the data aligns with our

hypothesis—specifically, superior TR performance on A1 items and superior TV performance on A2 items.

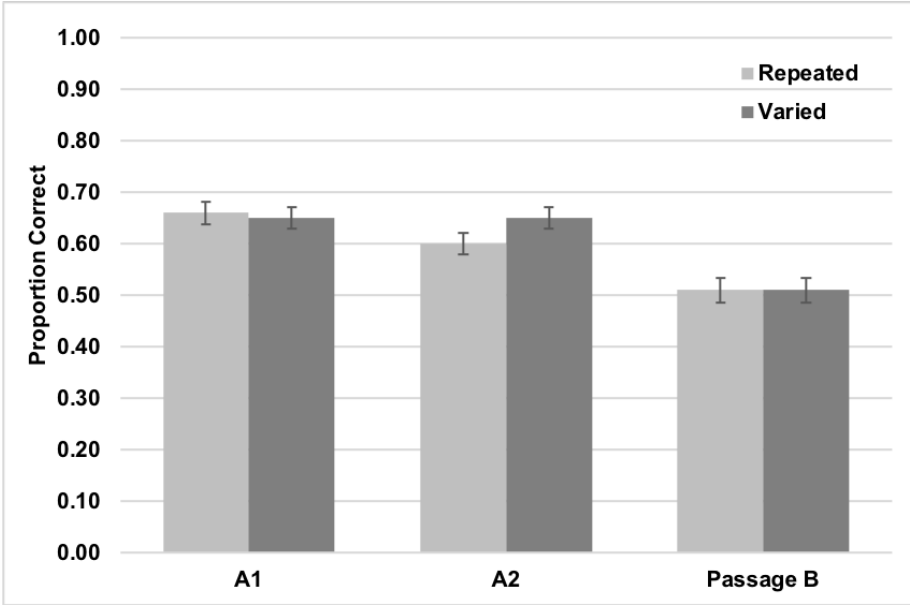


Figure 4. Testing method as a function of question type.

Score Predictions

Table 3 shows the raw prediction, actual, and difference score means for each learning condition and testing method.

Prediction. We analyzed data from score predictions made on Day-2 for material from all four passages using a repeated measures ANOVA. Results showed a significant main effect of Biology content, $F(2,149) = 14.54$, $p < .001$. Learning condition, $F(1,149) = 2.18$, $p = .14$, passage, $F(1,149) = 1.69$, $p = .20$, and testing method, $F(1,149) = .06$, $p = .81$, were not significant factors in the model. Two-way interactions between learning condition and passage, $F(1,149) = 1.05$, $p = .31$, learning condition and testing method, $F(1,149) = 2.45$, $p = .12$, and testing method and passage[4], $F(1,151) = 5.44$, $p = .02$, were not statistically

[4] The two-way interaction between testing method and passage was not significant after applying FDR at $p = .05$

significant. However, we observed a significant three-way interaction between learning condition, testing method, and passage, $F(1,149) = 8.92$, $p = .003$.

Specific contrasts showed that participants made significantly lower score predictions on Osmosis/Diffusion content (LS Mean = .52) than on DNA content (LS Mean = .57), $F(1,149) = 8.63$, $p = .004$, and lower predictions on DNA than on Food Webs content (LS Mean = .62), $F(1,149) = 6.72$, $p = .01$. Test (LS Mean = .59) and restudy (LS Mean = .57) conditions did not differ on passage A, $F(1,149) = 2.89$, $p = .09$, or passage B predictions (test LS Mean = . 57; restudy LS Mean = .56), $F(1,149) = .55$, $p = .46$. Contrasts also showed no difference in score predictions between TR (LS Mean = .62) and TV (LS Mean = .56) groups on passage A, $F(1,149) = 2.20$, $p = .14$, or passage B judgements (TR LS Mean = .55; TV LS Mean = .59), $F(1,149) = .89$, $p = .35$. Figure 5 shows score predictions by learning condition, testing method, and Biology passage. The three-way interaction seems to be due to the finding that participants in the TV group made lower score predictions than the TR group for passage A material, but higher predictions than TR on passage B.

Table 3. Prediction, Actual, and Difference Scores by Condition and Passage.

		Passage A			Passage B		
	n	Predict	Actual	Difference	Predict	Actual	Difference
Restudy-Repetition	74	0.55 (.27)	0.51 (.26)	0.04 (.30)	0.54 (.26)	0.46 (.24)	0.08 (.27)
Restudy-Variation	77	0.58 (.24)	0.54 (.21)	0.04 (.27)	0.57 (.25)	0.47 (.24)	0.10 (.26)
Test-Repetition	74	0.62 (.24)	0.74 (.20)	-0.12 (.23)	0.55 (.25)	0.51 (.23)	0.04 (.26)
Test-Variation	77	0.56 (.27)	0.75 (.19)	-0.19 (.26)	0.59 (.26)	0.51 (.23)	0.08 (.26)

Note: Results reported represent raw means and standard deviations by group. For difference scores, positive values represent expecting a higher score than observed, and negative values represent expecting lower score than observed. Predict- mean prediction score made before taking the final test. Actual- observed final test score. Difference- sum of the difference between predicted and actual score.

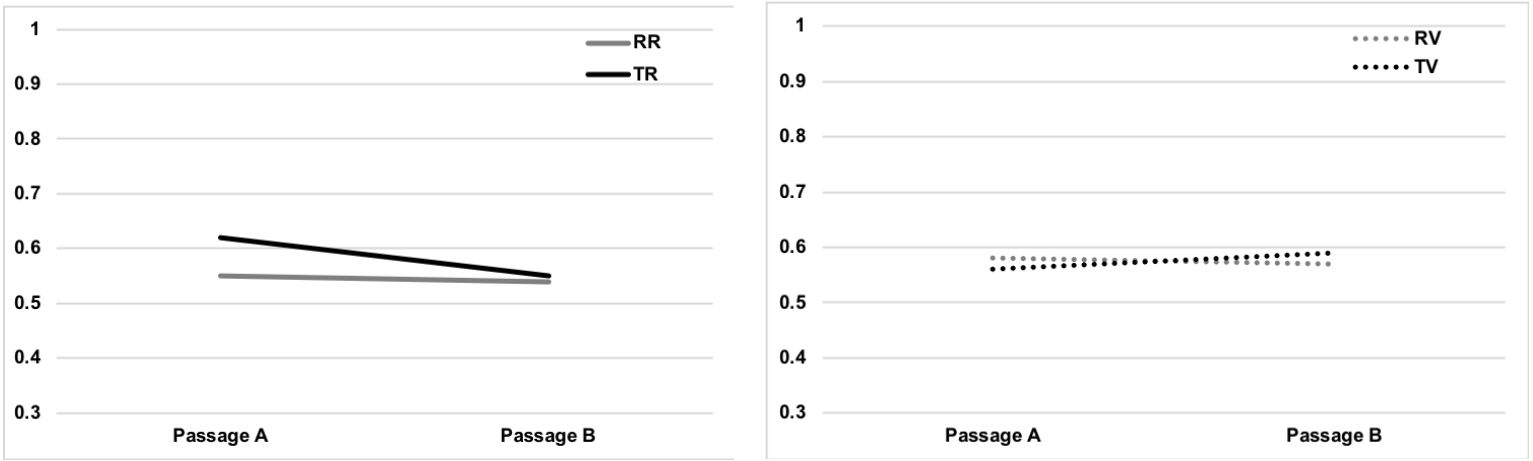


Figure 5. Score predictions by learning condition, testing method, and Biology passage. Data points in Figure 5 represent regression adjusted means (LS Means). The panel on the left shows the restudy (RR) and test (TR) groups within the repeated testing method, while the right panel shows the restudy (RV) and test (TV) groups within the varied testing method.

Prediction accuracy. In a separate analysis of the effect of experimental design features on prediction accuracy, we observed significant main effects of learning condition, $F(1,149) = 70.50$, $p < .001$, and Biology passage, $F(1,149) = 63.56$, $p < .001$. Participants had lower (negative) difference scores in the testing condition (LS Mean = -.04) than in the restudy condition (LS Mean = .08), suggesting that participants were under-confident in the testing condition, but over confident in the restudy condition. Participants also had lower difference scores for passage A (LS Mean = -.03) than for passage B material (LS Mean = .07). Testing method, $F(1,149) = .09$, $p = .77$, and Biology content, $F(1,149) = .71$, $p = .49$, were not significant predictors of the difference score measure.

We also observed significant two-way interactions between learning condition and passage, $F(1,149) = 45.50$, $p < .001$, and between testing method and passage, $F(1,149) = 14.32$, $p = .002$. Figure 6 shows the difference between predicted and observed scores by learning condition and passage. Specific contrasts revealed a significant difference between test (LS Mean = -.13) and restudy conditions (LS Mean = .06) on passage A, $F(1,149) = 104.33$, $p < .001$, and passage B difference scores (test LS Mean = .05; restudy LS Mean = .09), $F(1,149) = 5.24$, $p = .02$. The two-way interaction between condition and testing method, $F(1,149) = 2.11$, $p = .15$, and the three-way interaction between learning condition, testing method, and passage[5], $F(1,149) = 4.48$, $p = .04$, were not significant.

The TV (LS Mean = -.18) group was significantly more under-confident than the TR group (LS Mean = -.08) on passage A material, $F(1,149) = 6.23$, $p = .01$, but the two groups (TV LS Mean = .07; TR LS mean = .03) did not differ in accuracy on passage B material, despite having means in opposite directions for the two types of passage.

[5] The three-way interaction between learning condition, testing method, and Biology passage was not significant after controlling FDR at $p = .05$—despite showing an individual p value of less than .05.

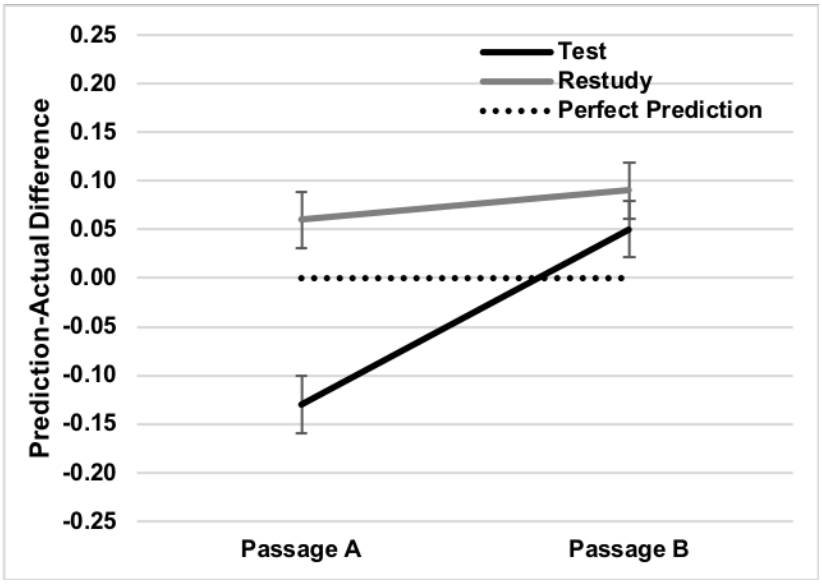


Figure 6. Prediction-actual difference score by learning condition and Biology passage. Positive difference score values indicate the degree of overconfidence and negative difference score values indicated the degree of under-confidence. Error bars represent standard errors.

Individual Differences in Cognitive Ability

To evaluate whether the effects of testing on learning and transfer of learning are equal for students with varying abilities, we examined four individual difference constructs—science reasoning, reading comprehension, non-verbal reasoning, and working memory—to the data analytic approach used in the original analysis. A correlation table showing the relationships between performance and cognitive ability constructs can be seen in Appendix B. This approach was executed using four separate models, each examining whether a single individual difference construct moderated the effects observed in the original analysis. Finally, a fifth (joint) model was used to elucidate whether the effects found in each of the four separate models are the result of unique effects specific to a single construct or the results of the covariance across the multiple constructs.

Model 1: Science reasoning. Results from models 1-4 can be seen in Table 4. In Model 1, the effects of our experimentally manipulated variables (i.e., leaning condition, question type, and testing method) and their interaction were largely unchanged relative to our true experiment model. Results from Model 1 indicated that science reasoning ability, $b = .14$, $F(1,149) = 61.59$, $p < .001$, had a strong positive relationship with final test performance. We observed a significant interaction between question type and science reasoning, $F(2,149) = 8.73$, $p < .01$. Two-way interactions between learning condition and science reasoning, $F(1,149) = 1.15$, $p = .29$ and testing method and science reasoning, $F(1,149) = 3.25$, $p = .07$, were not significant. The three-way interaction between learning condition, question type, and science reasoning, $F(1,149) = 1.62$, $p = .23$, was also not significant.

Figure 7 shows final test performance by question type and science reasoning ability. Through examining Figure 7 it is clear that there are disproportional differences between participants with the low, medium, and high science reasoning ability across the three types of final test items. Specifically, participants with higher ability show a decrease in performance across the three types of questions (transfer) at a slower rate than the low and medium science reasoning groups. The finding of this interaction between question type and science reasoning suggests that participants higher in science reasoning are better able to transfer their knowledge from passage A to passage B regardless of the condition under which that material is learned and transferred. The lack of a significant interaction between learning condition and science reasoning suggests that the effect of retrieval practice is equal across participants with varying science-reasoning abilities. Likewise, science-reasoning ability does not moderate the effect of retrieval practice on transfer. While non-significant,

the results of Model 1 suggest that science reasoning may moderate the effect of testing method given the *p*-value between .05 and .10.

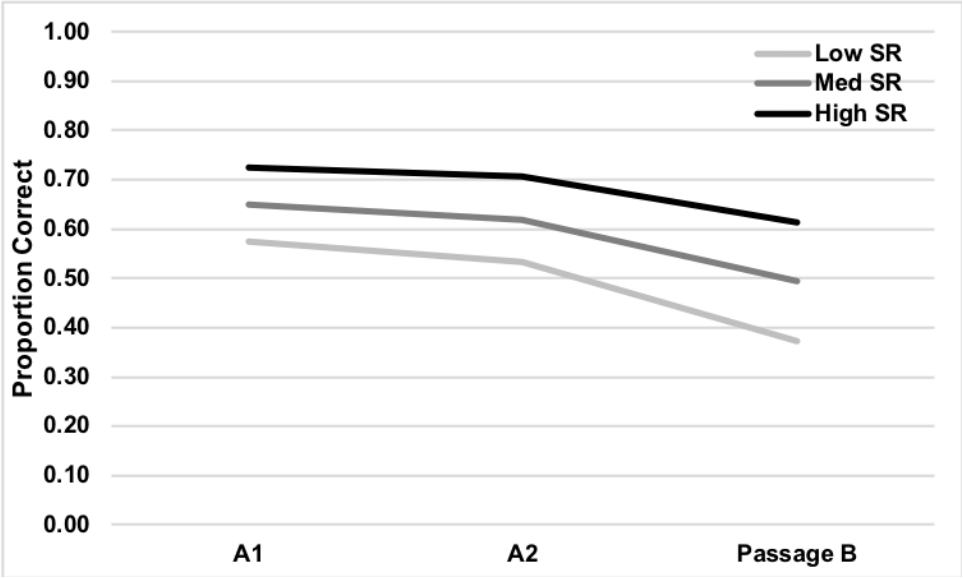


Figure 7. A line plot showing final test performance by question type and science reasoning (SR) ability. Low SR = SR scores 1 *SD* below the mean, med SR = SR score at mean, high SR = SR score 1 *SD* above the mean.

Table 4

Results for Models 1-4

	Model 1: SR		*Model 2: RC*		*Model 3: NVR*		*Model 4:WM*	
	F	*p*	*F*	*p*	*F*	*p*	*F*	*p*
Learning Condition	104.91	< .001	106.36	< .001	112.86	< .001	108.09	< .001
Question Type	111.35	< .001	99.90	< .001	98.00	< .001	96.01	< .001
Testing Method	0.66	0.42	0.85	0.36	1.31	0.25	0.32	0.58
Learning Condition x Question Type	50.36	< .001	50.97	< .001	51.43	< .001	51.06	< .001
Learning Condition x Testing Method	0.62	0.43	1.12	0.29	1.10	0.30	1.20	0.28
Testing Method x Question Type	4.70	0.01	4.80	0.01	5.03	0.01	5.24	0.01
Learning Condition x Testing Method x Question Type	1.10	0.34	1.28	0.28	1.04	0.36	0.99	0.37
Model Specific Individual Difference Construct (ID)	61.59	< .001	40.34	< .001	31.36	< .001	5.77	0.02
Learning Condition x ID	1.15	0.29	0.85	0.36	0.59	0.46	0.29	0.59
Question Type x ID	8.73	< .001	5.45	0.01	2.17	0.12	0.23	0.80
Testing Method x ID	3.25	0.07	0.30	0.58	0.18	0.67	0.93	0.34
Learning Condition x Question Type x ID	1.62	0.23	0.37	0.69	0.44	0.66	1.71	0.19

Note: Results in Table 5 reflect the results of fixed effects (main effects and interactions) across Models 1-4. Main effects and interactions of learning condition, testing method, and question type were used in each model. Models 1-4 added one individual difference construct at a time. Multiple imputation was use for model 2. Results in model 2 represent the average effects from the analysis of 20 datasets produced from multiple imputation Procedures for pooling the results of these 20 analyses were derived from Rubin (1987) and van Ginkel and Kroonenberg (2014). SR = Science Reasoning; RC = Reading Comprehension; NVR = Non-Verbal Reasoning; WM = Working Memory.

Model 2: Reading comprehension. In Model 2, we observed a significant main effect of reading comprehension, $b = .09$, $F(1,149) = 40.34$, $p < .001$, and a significant interaction between reading comprehension and question type, $F(1,149) = 5.45$, $p = .005$. Figure 8 shows the interaction between reading comprehension and question type—a result that resembles the interaction between science reasoning and question type (reported in Model 1)—where participants with higher reading comprehension ability showed better transfer than participants with lower reading comprehension. Because of the similar trend in results between Model 1 and Model 2 (i.e., strong main effects of a cognitive ability measure and its interaction with question type) we pay particular attention to the results of Model 5 (reported below) where we include all significant factors from Models 1-4 in a joint model in an attempt to replicate these effects in the most complete model possible.

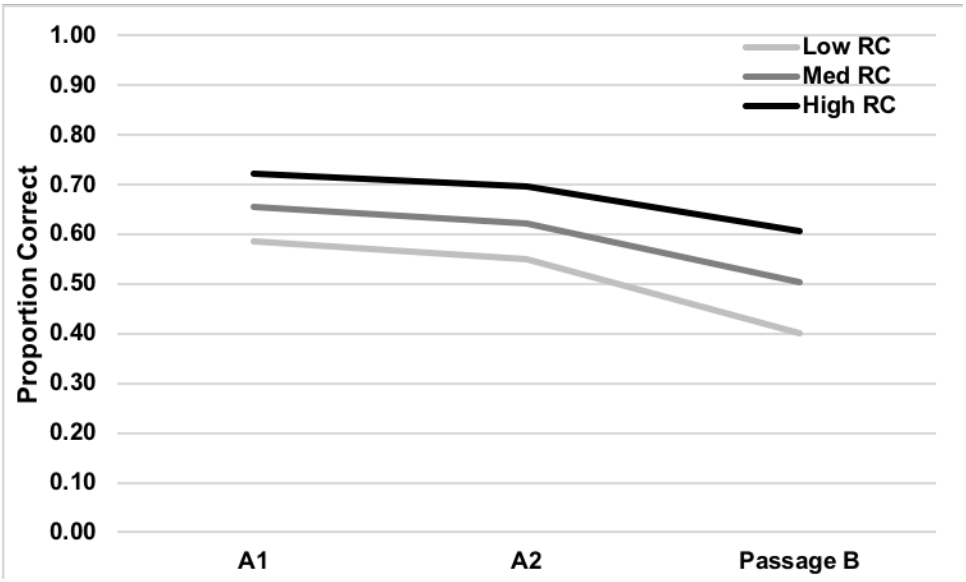


Figure 8. A line plot showing final test performance by question type and reading comprehension (RC). Low RC = RC scores 1 *SD* below the mean, med RC = RC score at mean, high RC = RC score 1 *SD* above the mean.

Model 3: Non-verbal reasoning. Results for Model 3 showed a significant main effect of non-verbal reasoning, $b = .03$, $F(1,149) = 31.36$, $p < .001$. Commensurate with

Models 1 and 2, we also sought to evaluate whether non-verbal reasoning moderates the effects of design factors discussed in the original analysis. However, results revealed that non-verbal reasoning does not interact with learning condition or testing method ($F < 1.0$). Non-verbal reasoning also did not interact with question type, $F(2,149) = 2.17$, $p = .12$. The three-way interaction between learning condition, non-verbal reasoning, and question type, $F(2,149) = .44$, $p = .65$, was also not significant. Collectively, these results suggest that the effects of learning condition, testing method, and question type are unchanged by non-verbal reasoning ability.

Model 4: Working memory. The results from Model 4 show a significant main effect of working memory, $b = .001$, $F(1,149) = 5.77$, $p = .02$, however, this effect is markedly weaker than seen with the main effects of individual difference constructs in the Models 1-3. Working memory did not moderate the effects of learning condition, testing method, or question type ($F < 1.0$). The three-way interaction between learning condition, question type, and working memory, $F(2,149) = 1.71$, $p = .19$, was not significant. Results from Model 4 suggest that working memory is weakly correlated with final test performance and, furthermore, does not moderate the design effects in the original analysis.

Model 5: Joint effects of individual differences. In Models 1-4 we replicated the results of our analysis of the experimentally manipulated study variables, that is, Models 1-4 did not negate a previously reported significant effect of experimental manipulations from the original analysis. In addition, in Models 1-4 we also observed significant main effects of science reasoning, reading comprehension, non-verbal reasoning, and working memory, and significant interactions between science reasoning and question type and reading comprehension and question type. In Model 5 we included each of the individual differences

constructs and the previously reported significant interactions in addition to the factors in our original experimental model. The goal of Model 5 was to investigate whether the effects found in Models 1-4 represent effects specific to each of the four individual difference constructs or whether specific results were due to the covariance between two or more of our constructs. For this reason, Model 5 is considered the most complete and ecologically valid of the five models.

The results of Model 5 can be seen in Table 5. Once again, the results reported in the original model were replicated in Model 5. We observed significant main effects of reading comprehension, $b = .07$, $F(1,147) = 9.82$, $p = .002$, science reasoning, $b = .05$, $F(1,147) = 7.75$, $p = .006$, and non-verbal reasoning, $b = .01$, $F(1,147) = 7.10$, $p = .009$. Working memory, $b = .000$, $F(1,147) = .04$, $p = .85$, was not a significant predictor in Model 5. Interestingly, the interaction between science reasoning and question type, $F(1,147) = 4.38$, $p = .02$, was replicated in Model 5 while the interaction between reading comprehension and question type, $F(1,147) = 1.00$, $p = .37$, was not. A line plot of the interaction between science reasoning and question type can be seen in Figure 9, which can be considered a more ecologically valid and updated version of Figure 7. As reported in Model 1, the interaction between science reasoning and question type is characterized by better transfer for participants with higher science reasoning abilities. The lack of a significant interaction between reading comprehension and question type in Model 5, suggests that the previously reported significant interaction in Model 1 was not due to superior reading comprehension among students with stronger scientific reasoning, but was specifically attributable to science reasoning.

Table 5

Results for Model 5: Joint Effects of Individual Differences

	F	*p*
Learning Condition***	109.75	< .001
Question Type***	109.49	< .001
Testing Method	1.84	0.18
Reading Comprehension (RC)**	9.82	0.002
Science Reasoning (SR)**	7.75	0.006
Non-Verbal Reasoning (NVR)**	7.10	0.009
Working Memory (WM)	0.04	0.85
Learning Condition x Question Type***	52.24	< .001
Learning Condition x Testing Method	0.93	0.34
Testing Method x Question Type*	4.68	0.01
Learning Condition x Testing Method x Question Type	1.32	0.27
Question Type x RC	1.00	0.37
Question Type x SR*	4.38	0.02

Note: *p < .05, **p < .01, ***p < .001;

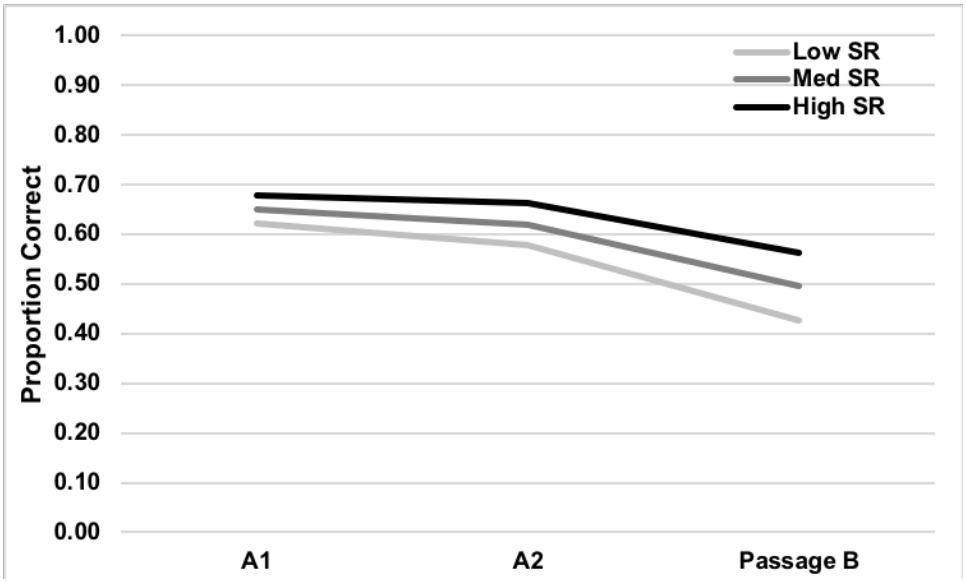


Figure 9. A line plot showing final test performance by question type and science reasoning (SR) ability. Low SR = SR scores 1 *SD* below the mean, med SR = SR score at mean, high SR = SR score 1 *SD* above the mean.

Discussion

In this study we simultaneously investigated 1) the effects of testing on learning and transfer of learning, 2) the effects of testing on pre-test score predictions and their accuracy, and 3) whether individual differences in cognitive ability interact with the effects of testing on learning and transfer of learning. We examined these three main lines of research questions in a controlled laboratory study using college level biology materials—which we deemed to be highly educationally relevant and complex.

Final Test Data

The primary design feature of this study is that we defined meaningful levels of near and far transfer. Specifically, we identified topics in Biology which could be broken down to primary (passage A) and secondary (passage B) sub-topics. As the goal of this study was to determine the extent to which testing leads to improved learning and transfer of learning, we further operationalized transfer by constructing three levels of final test questions (A1, A2, and passage B) and compared participants performance in the testing and restudy conditions on a set of questions of each type. While we expected that participants would perform better on A1 items (i.e., those seen exactly at least once on Day-1) in the test condition as compared to the study condition, we paid particular attention to testing versus restudy performance on A2 (i.e., novel items about passage A concepts and facts tested during Day-1) and B items (i.e., questions about passage B—a passage with a hierarchical relationship to passage A).

First and foremost, we found that participants performed better on the final test in the test condition than in the restudy condition across A1, A2, and passage B test items. This finding is significant, because it suggests that taking tests leads to better learning (A1) and transfer of learning (A2 and B), even for complex college-level educational materials. Not

only did participants perform better in the testing condition on questions that they had seen before (A1 items), they performed better (in the testing condition) on conceptually related items (A2) and items about a topic on which they did not test, but that would benefit from better learning of the passage A material (passage B). Additionally, our final test was sensitive enough to detect these differences in performance at a 24-hour delay, whereas, other studies have only found effects this large at up to a 7-day delay (e.g., Butler, 2010).

The effect sizes of our transfer effects (d = .60 for A2 items; d = .21 for B items) are within range of the effects reported by Pan and Rickard (2018). As a cautionary note, we point out that our study differs in its definition of transfer (especially for passage B items) and materials from many of the studies analyzed in Pan and Rickard. In the terminology put forth by Pan and Rickard it could be argued that the type of transfer format in our study could be described as both application/inference questions (A2 and B items) and untested material (B items). While some of the A2 items had response congruency, many of the A2 items required slightly or wholly different answers. Additionally, feedback during the practice test phase in our study only provided correct answers and did not elaborate. In terms of qualifying our results under the framework of Pan and Rickard, the effects reported here are generally in line with a trend of positive transfer for testing conditions, however we argue our study is the first to report successful transfer of practice tests to subsequent untested educational Biology material.

Although we expected participants to perform well on A1 items in the testing condition, it is somewhat surprising that the performance in the testing condition only decreased by about 10% from A1 to A2. One reason for this relatively robust “near transfer” effect could be that the items within each concept (some randomly assigned to be A1 and

some to be A2 items) were closely related to one another. As reported in our procedures, we began by identifying key concepts in each of the Biology passages that were to be used in the study. We wrote multiple cued-recall items for each concept we identified in each passage in order to 1) be able to present participants in the test-variation condition with multiple "related" practice test items and 2) to be able to present both TR and TV conditions with novel items about a concept they had taken practice test questions on, thus defining "near transfer" within the context of the study. Additionally, we argue that the comparison between test and restudy at A2 is more of a fair comparison than on A1, because for both groups A2 items have never been seen before.

In analyzing the results of this study, we were very interested to see whether participants performed better on passage B items in the testing condition. To review, on Day-1 participants in our study read passage A and either restudied or took practice test questions on passage A. Then 24-hours later on Day-2, participants read passage B, which was constructed to require the use of information learned from passage A. The results of our study indicated that participants performed better on passage B final test questions in the test condition than in the restudy condition. While we observed a modest Cohen's *d* effect size of .21, these results suggest that testing over primary material (passage A) yields measurable performance improvements on subsequent material (passage B). We argue that the positive effect of testing on subsequent material could be due to both: 1) better direct (transfer) of the information learned in passage A when answering passage B test questions, and/or, 2) better comprehension of passage B material during reading, gained through a better understanding of the pre-requisite knowledge (passage A). Although it is interesting and important to

consider the mechanism for the positive effect of testing that we found in this study, our study was not primarily designed to investigate these questions empirically.

Taken together, the results of our study suggest that a relatively short retrieval practice session (24 practice test questions and correct answer feedback) was associated with improved performance on conceptually related material and subsequent material which participants did not test on or restudy. Furthermore, we found positive effects of testing and evidence suggesting that testing leads to better transfer of learning using a within-subjects design, where all participants were exposed to a test and control condition. Using the within-subjects design in this study possibly makes the results clearer about the "direct effects of testing." While there is a significant literature on "indirect testing effects," (e.g., Arnold & McDermott, 2013; MacLeod & Daniels, 2000) we argue that performance in both the testing and restudy conditions was equally influenced by test-expectancy effects and effects of familiarity of the testing format. In our study it seems reasonable that participants were more familiar with the content of questions that might be asked on the final test after seeing practice test questions on Day-1 for the topic that they tested on (testing condition); however, all participants were exposed to cued recall practice test questions on one of the two passages, and therefore, were familiar with our study's particular style and procedure for testing. Thus, test format familiarity effects discussed in detail by some researchers (e.g., Morris et al., 1977; Roediger, 1990; Roediger & Butler, 2011) cannot explain the effects of testing observed in our study.

We also designed this study to investigate the effect of variability in testing. A theory of encoding variability (e.g., Estes, 1955) suggests that more variation in the learning process might increase the number of unique learning experiences, creating a stronger network of

knowledge and leading to improved performance. While not a direct investigation of test variability, the principle of encoding variability discussed and applied in Kornell and Bjork (2008) was a strong motivation for our pursuit of the potential benefits of varied testing. Butler (2010) was the first to specifically test whether taking "re-worded" questions (test variability) was better than taking the same questions repeatedly. Butler reported no difference between test variability and test repetition in his study, however, participants performed better in any testing condition relative to control in many scenarios. In a similar pattern, our study also did not observe any significant differences between the TV and TR sub-groups. However, we did observe a significant interaction between testing method and question type, such that the repeated condition appeared to perform better than the varied condition on A1 items, but there was a larger difference between the two groups on A2 items in the opposite direction (better performance under the varied testing condition than under the repeated testing condition on A2 questions). However, pairwise comparisons between the two groups were not significant. While the interaction between testing method and question type is interesting, the most direct tests of the utility of variation in practice test questions are in the three-way interaction between learning condition, testing method, and question type and pairwise comparisons between the TV and TR sub-groups on A1, A2, and B final test items. Neither the three-way interaction between learning condition, testing method, and question type nor the pairwise comparisons between the TV and TR groups were significant. Despite these non-significant effects, and in line with predictions made in H_4, we observed a Cohen's *d* effect size of .20 in favor of the TR group on A1 items and an effects size of *d* = .21 in favor of the TV group on A2 items. For these comparisons, we estimate that

achieved power (1-β) was approximately .60—slightly below the desired power to detect a significant effect of this magnitude.

The finding of no significant difference between the TR and TV groups is interesting despite our failure to reject the null hypothesis. This finding means that there was not a large difference between the TR and TV groups on A1 despite the fact the that TV group saw A1 questions with feedback only once, while the TR groups made a retrieval attempt and saw feedback for these items three times. Furthermore, the TV group performed at least comparably to (if not better than) the TR group on A2 items. The TR and TV groups did not seem to differ whatsoever on passage B items, which may mean that the extent to which variability in practice test questions is beneficial is limited to some level of "near transfer" and may not aid students (relative to test repetition) in learning subsequent material.

In designing our study, we paid particular attention to the results of Butler (2010), who also investigated the effects of testing and test variability on transfer using educational materials. Herein, we have reported some similar results to those observed in Butler, most notably that we found evidence that testing improves transfer of learning and that variation in testing is at least comparable to test repetition for transfer. In the interest of comparing our results to Butler one difference in our study is that we defined "near transfer" as performance on A2 items, which were novel but closely related to A1 items; whereas Butler defined near transfer as performance on the questions requiring inferences in his studies 1b and 2. We also defined "far transfer" as performance on passage B items, while Butler looked at performance on questions from another related knowledge domain that required the use of information learned during initial reading and practice tests. In looking at the effects of test repetition, we specifically designed practice test questions in the TV condition to be closely

related to one another, but not just reworded versions of the same questions as done in Butler. We wondered if increasing the figurative "distance" in similarity between questions taken in a practice test session would increase the effect of testing on transfer. While, neither our study nor Butler found a positive effect of test variability over test repetition, mean performance in both our study and Butler's study 1b (same test vs. varied test performance on conceptual questions) was in the hypothesized direction—superior transfer performance in the varied condition. Future research should re-examine the roles of variation in testing and power experiments to detect relatively small effect sizes.

Score Prediction Data

We also investigated the effects of testing and test variability on pre-test score predictions and their accuracy. Procedures for participants' score prediction began by asking participants to predict the percentage score they expected to receive on passage A and passage B content for both of the Biology passages they were exposed to. We executed two statistical analyses using these predictions: one looking at the experimental design effects on the predictions themselves and one looking at our measure of prediction accuracy (obtained by subtracting each participant's actual score from their prediction score for each passage). In examining the level of accuracy of participant score predictions we look at both the distance from zero (a literal definition of accuracy) and the sign (+/-)—that is, the degree of overconfidence (i.e., expecting a higher score than received), or under-confidence (i.e., expecting a lower score than received).

The main finding in our analysis of participants *raw prediction scores* was a strong three-way interaction between learning condition, testing method, and Biology passage. This three-way interaction was characterized by a disordinal interaction between the TR and TV

groups—such that the TR group predicted higher scores on passage A than the TV group, but lower scores on passage B than the TV group. Although no study of which we are aware has investigated the effect of variable testing on metacognitive judgments, the finding that the TV group had comparable score predictions relative to the restudy condition is more in-line with the hypothesis that testing tempers metacognitive judgments (e.g., Soderstrom & Bjork, 2014). On the other hand, TR participants had the highest score predictions of any group, but actual performance data show that they were still under-confident (discussed below).

When we take into account participants' actual scores in the prediction-observed difference scores, we see very interesting results. Participants were generally *less* overconfident in the testing condition than in the restudy condition. However, the main effect of learning condition was superseded by significant interactions between learning condition and Biology passage and between testing method and Biology passage. As can be seen in Figure 6, in the testing condition, participants were severely under-confident on passage A but were slightly overconfident on passage B, while participants were overconfident at both time points in the restudy condition. We also observed an interesting disordinal interaction between testing method and Biology passage, where the participants in the varied condition were more under-confident than the repeated condition on passage A, but more overconfident than the repeated condition on passage B. While the three-way interaction between learning condition, testing method, and Biology passage in our accuracy models was not significant after controlling FDR at $p = .05$, pairwise comparisons revealed that the TV group was significantly more under-confident than the TR group on passage A, but not on passage B material.

As can be seen in Table 2, the TV group predicted a score of .56 on passage A, but achieved a score of .75 earning a prediction-observed difference score of -.19. The TR group predicted a passage A score of .62 and achieved an actual score of .74 (a difference score of -.12). While both the TV and TR groups are under-confident on passage A, it is interesting that completing practice tests (albeit in a more challenging, varied form) did not raise the TV group's pre-test predictions relative to restudy groups. Indeed, participants in the TV testing condition gave raw prediction scores of .58 in the restudy learning condition. In other words, participants in the varied testing method condition were equally unconfident in the restudy learning condition as in the testing learning condition, despite significantly better actual performance in the testing condition. By contrast, participants in the repeated testing condition were much more confident in the testing learning condition than in the restudy learning condition. Furthermore, it is also interesting that despite TV participants modest passage A predictions (.56), the TV group gave the highest passage B prediction (.59) and were the only group to predict higher scores for passage B than for passage A. It is possible that TV participants felt that passage A was more difficult because varied testing was more challenging, and could have rated passage B as comparable in difficulty or easier because they did not received feedback about their actual performance on passage B.

In interpreting this set of interesting data regarding participants metacognitive awareness of performance, we also need to consider the timing of predictions and exposure to passage A and passage B. At the time that score predictions were made, participants had been exposed to passage A 24-hours prior, but had just completed reading both passage Bs. While the timing of reading and additional exposure to the passages (re-reading and testing) may explain the main effect of Biology passage in the model on prediction accuracy, this

does not explain the interactions between learning condition, testing method, and Biology passage.

Overall, we draw three main conclusions from the score prediction data: 1) While test repetition increases confidence for the content that was tested (passage A) relative to restudy, test variation does not—despite equal actual performance between test repetition and test variation. Test variation seemed to decrease prediction scores even relative to the judgments made by the same participant on material that was restudied. It is likely the challenge of being exposed to a totally new question for each concept or fact tested during the practice test session decreased participant confidence on their level of mastery relative to the TR group. 2) Testing is associated with more under-confident score predictions relative to restudy. For both types of testing, participants had lower prediction-observed difference scores in the testing condition than in the restudy condition. Finally, 3) Test variation may artificially improve confidence for subsequent material (passage B). While the TV group greatly under-estimated their performance on passage A material, they also reported higher passage B predictions relative to their own passage A predictions and passage B predictions of all other groups. Overall, the TR group's passage B predictions were most accurate (difference score closest to zero).

Individual Differences in Cognitive Ability

As discussed, there have few attempts to investigate whether individual differences (specifically in cognitive abilities) moderate the effects of testing. In an exploratory aim of this project we identified four cognitive ability constructs that we believed influence science learning: science reasoning, reading comprehension, non-verbal reasoning, and working memory. We then analyzed the effect of each of these constructs and their interaction with

design features in our original experimental model separately in Models 1-4. Finally, we constructed a joint model consisting, of all original effects and significant cognitive ability main effects and interactions observed in Models 1-4. Our primary goals in these analyses were to evaluate whether individual differences in cognitive ability moderate the effect of testing (H_5) and the effect of transfer of learning (question type; H_6).

First, the effects found in the original model were replicated in each of Models 1-5. That is, including cognitive ability constructs and their interaction with the experimental design features did not change the effects reported in the original model regarding the effects of testing on learning and transfer. In Models 1-4 we observed significant main effects of all individual cognitive ability constructs (science reasoning, reading comprehension, non-verbal reasoning, and working memory). We did not observe a significant interaction between any cognitive ability construct and learning condition in any of our Models 1-4, which implies that the effect of testing (compared to restudy) is comparable across varying levels of our cognitive ability measures. In reconciling our results of Models 1-4 with others in the existent literature, we did not find evidence to suggest that testing effects differ across people with various profiles of cognitive abilities. Although it would be premature to conclude that individual differences in learning strategies (i.e., retrieval practice) do not exist, we report sizable testing effects but do not observe any interaction between learning condition and individual difference measures similar to the report by Bertilsson et al. (2017).

We did, however, observe two-way interactions between science reasoning and question type and reading comprehension and questions type. When the significant main effects and interactions were combined in a joint individual differences model (Model 5), main effects remained for science reasoning, reading comprehension, and non-verbal

reasoning, but not for working memory, and only the interaction of question type with science reasoning remained significant. The non-significant effect of working memory in Model 5 suggests that variance accounted for by working memory in Model 4 may be attributable to the other constructs, whereas the significant interaction of science reasoning and question type suggests that it is not an artefact of better reading comprehension among students with higher scientific reasoning. As can be seen in Figure 9, the interaction between science reasoning and question type seems to be due to a larger performance difference between participants with varying science reasoning abilities at further levels of transfer. For example, participants with a science reasoning score 1 *SD* above the mean showed a 2% relative decrease in performance from A1 to A2 items and a 15% relative decrease from A2 to B items. On the other hand, participants with a science reasoning score 1 *SD* below the mean had a 7% relative decrease from A1 to A2 and a 26% decrease from A2 to B items. Taken together, these results mean that higher science reasoning ability leads to better transfer on both A2 and B items. Practically, this is an important finding regarding transfer. These results may mean that, especially within science and other disciplines that require strong reasoning ability, students with lower reasoning ability may be disproportionally disadvantaged when seeking to apply their learning from primary material to more complex material. Future research should aim to confirm these effects and, if replicated, seek to investigate how to improve transfer for students with low reasoning abilities. The fact that the interaction remains significant in the presence of reading comprehension as a predictor implies that simply focusing on improving reading comprehension among lower ability students would not be sufficient.

Limitations and Future Directions

One limitation of this study is that it was not powered to detect relatively small effect sizes of test variability (relative to test repetition). Because our focus was the general effect of testing, we chose to manipulate testing (of either form) within-subjects and variation in testing between-subjects. This resulted in very good power to detect the effects of testing, but low power to detect small (d = .21 on A2 items) effects of variable versus repeated testing. While we report a significant interaction between testing method and question type and effect sizes of pairwise comparisons between the TR and TV groups on the three types of questions, we also would like to know if there are statistically significant differences between these groups. In line with the consequences of decisions made to maximize the likelihood of detecting the effects we were most interested in, we also acknowledge that we chose a relatively short retention interval (24-hours) and one fixed schedule for the timing of reading passages A and B. Previous studies have suggested that the testing effect is most powerful after at least a 24-hour delay (some studies have looked at delays of 1-week to several weeks). In our study, we were specifically interested in operationalizing transfer in a practical and educationally relevant way, that is, performance on novel passage A and passage B questions. Therefore, we designed passage B to be read on Day-2 after a 24-hour delay in order to maximize the effect that testing might have on passage B. However, we also were cautious not to increase the delay between Day-1 and Day-2 further at the risk of washing out potential transfer effects on passage B content. As a result, we were able to find positive effects of testing at different levels of transfer, but with our relatively short retention interval, we are unsure of how stable these effects are over time and whether the timing of reading passage A and passage B matters. Future research with the goal of validating

retrieval practice procedures for use in real educational settings should power studies to be able to detect significant differences between test repetition and test variation groups.

Conclusion

We set out to execute a highly educationally relevant investigation of the effect of testing on transfer of learning. We also designed our study to be able to consider the effects of testing on pre-test score predictions and whether individual differences in cognitive ability moderate the effect of testing and transfer of learning. Through our analyses we found significant effects of testing of varying magnitude on repeated items (d = 1.01), related items (d = .60), and questions on subsequent material (passage B; d = .21). We also reported that testing influences pre-test score prediction, such that repeated testing is associated with increased pre-test confidence, while varied testing is not. Additionally, both testing groups performed comparably on the final test, leading to large discrepancies in prediction-observed difference scores. Finally, we report that individual differences in cognitive ability do not interact with testing effects, suggesting that the testing effect is relatively equitable across people with varying levels of ability. In our study, science-reasoning ability moderated the effect of transfer, such that participants with higher science reasoning ability achieved better transfer performance on the final test. Overall, we conclude that retrieval practice with cued recall questions is a highly effective strategy for learning complex educational materials. Furthermore, when test questions are well-constructed and closely related to the target material, students who use retrieval practice can expect benefits of testing even when the final test questions are novel.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861-876. https://doi.org/10.1002/acp.1391

Arnold, K. M., & McDermott, K. B. (2012). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(3), 940–945. https://doi.org/10.1037/a0029199

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 89–99. https://doi.org/10.1080/00220671.1991.10702818

Barenberg, J., & Dutke, S. (2018). Testing and metacognition: Retrieval practise effects on metacognitive monitoring in learning from text. *Memory*, *27*(3), 269-279. https://doi.org/10.1080/09658211.2018.1506481

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? *Psychological Bulletin*, *128*, 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289-300.

Bertilsson, F., Wiklund-Hörnqvist, C., Stenlund, T., & Jonsson, B. (2017). The testing effect and its relation to working memory capacity and personality characteristics. *Journal of Cognitive Education and Psychology*, *16*(3), 241–259. https://doi.org/10.1891/1945-8959.16.3.241

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (ed.), *Information*

processing and cognition: The Loyola Symposium (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*, *3*, 165-170. https://doi.org/10.1016/j.jarmac.2014.03.002

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language, 65*(1), 32–41. https://doi.org/10.1016/j.jml.2011.02.005

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*(3), 407–415. https://doi.org/10.1016/j.jml.2011.12.009

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1118–1133. https://doi.org/10.1037/a0019902\

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918-928. https://doi.org/10.1037/0278-7393.34.4.918

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(45), 514-

527. https://doi.org/10.1080/09541440701326097

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563-1569. https://doi.org/10.1037/a0017021

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*, 279-283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review*, *19*, 443-448. https://doi.org/10.3758/s13423-012-0221-2

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*(3), 474-478. https://doi.org/10.3758/BF03194092

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U. S. history facts. *Applied Cognitive Psychology*, *23*, 760-771. https://doi.org/10.1002/acp.1507

Clark, C. M., & Bjork, R. A. (2014). When and why introducing difficulties and errors can enhance instruction. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), Applying the Science of Learning in Education: Infusing psychological science into the curriculum.

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, *315*(5811), 464–465. https://doi.org/10.1126/science.1134513

Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., & Engle, R.W.

(2005). Working memory span tasks: A review and a user's guide. *Psychonomic Bulletin & Review, 12*, 769-786. https://doi.org/10.3758/BF03196772

Druckman, D., & Bjork, R. A. (1994). Learning, remembering, believing: Enhancing human performance. Washington, DC: National Academy Press.

Eisenkraemer, R. E., Jaeger, A., & Stein, L. M. (2013). A systematic review of the testing effect in learning. *Paidéia*, *23*(56), 397-406. https://doi.org/10.1590/1982-43272356201314

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*, 369–377. https://doi.org/10.1037/h0046888

Estes, W. K. (1960). Learning theory and the new "mental chemistry." *Psychological Review*, *67*, 207–223. https://doi.org/10.1037/h0041624

Foss, D. J., & Pirozzolo, J. W. (2017) Four semesters investigating frequency of testing, the testing effect and transfer of training. *Journal of Educational Psychology*, *109*(8), 1067-1083. https://doi.org/10.1037/edu0000197

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–236. https://doi.org/10.3758/s13421-014-0461-7

Gates, A.I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*(40).

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562-567. https://doi.org/10.1016/S0022-5371(71)80029-4

Jacoby, L.L. (1978). On interpreting the effects of repetition: Solving a problem versus

remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*, 649–667. https://doi.org/10.1016/S0022-5371(78)90393-6

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*(2), 307-329. https://doi.org/10.1007/s10648-013-9248-9

Kang, S. H. K., McDermott, K. B., Roediger, H. L., III. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology*, *19*, 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*, 157-163. https://doi.org/10.1177/0963721412443552

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151-162. https://doi.org/10.1016/j.jml.2006.09.004

Kern, M. (2014). An Investigation of Individual Differences in the Testing Effect. (unpublished doctoral dissertation). Columbia University, New York, NY.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the enemy of induction? *Psychological Science*, *19*, 585-592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989-998. https://doi.org/10.1037/a0015729

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*(1), 11–24. https://doi.org/10.1002/tea.3660150103

Lawson, A. E., Alkhoury, S., Benford, R. B. C., & Falconer, K. A. (2000). What kinds of scientific concepts exist? Concept construction and intellectual development in college Biology. *Journal of Research in Science Teaching, 37*(9), 996–1018. https://doi.org/10.1002/1098-2736(200011)37:9<996::AID-TEA8>3.0.CO;2-J

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition*, *3*, 171-176. https://doi.org/10.1016/j.jarmac.2014.04.002

Little, J. & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*, 14-26. https://doi.org/10.3758/s13421-014-0452-8

MacLeod, C. M., & Daniels, K. A. (2000). Direct versus indirect tests of memory: directed forgetting meets the generation effect. *Psychonomic Bulletin & Review, 7*(2), 354-259. https://doi.org/10.3758/BF03212993

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Mayer, R. E. (2008). Learning and Instruction (2nd ed.). Upper Saddle River, NJ: Pearson.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L.

(2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399-414. https://doi.org/10.1037/a0021782

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201. https://doi.org/10.1016/0361-476X(91)90037-L

McDaniel, M.A., & Masson, M.E.J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 371–385. https://doi.org/10.1037/0278-7393.11.2.371

McDaniel, M. A., Thomas, R. C., Agarwal, P. K, McDermott, K. B, & Roediger, H. L., III (2013). Quizzing in middle school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*, 360-372. https://doi.org/10.1002/acp.2914

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20*, 3–21. https://doi.org/10.1037/xap0000004

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review*, *14*, 225-229. https://doi.org/10.3758/BF03194056

Miller, T. M., & Geraci, L. (2016). The influence of retrieval practice on metacognition: The contribution of analytic and non-analytic processes. *Consciousness and Cognition: An International Journal, 42*, 41–50. https://doi.org/10.1016/j.concog.2016.03.010

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519-533. https://doi.org/10.1016/S0022-5371(77)80016-9

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2009). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, *3*, 105-119.

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language*, *83*, 53–61. http://dx.doi.org/10.1016/j.jml .2015.04.001

Pan, S. C., & Rickard, T. C. (2018) Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710-756. https://doi.org/10.1037/bul0000151

Pirozzolo, J. W., & Foss, D. J. (2017). The testing effect and transfer using educational materials. Unpublished manuscript, University of Houston, Houston, Tx.

Pyc, M. A. & Rawson, K. A. (2009) Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*, 283–302. https://doi.org/10.1037/a0023956

Raven, J. (2009). The Raven Progressive Matrices and measuring aptitude constructs. International *Journal of Educational and Psychological Assessment*, *2*, 2-38.

Raven, J., Raven, J. C., & Court, J. H. (1998). Raven manual: Section 4, Advanced Progressive Matrices, 1998 Edition. Oxford, UK: Oxford Psychologists Press Ltd.

Reece, J. & Campbell, N. (2011). Campbell biology. Boston: Benjamin Cummings -Pearson.

Roediger, H. L., III. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*(9), 1043-1056. https://doi.org/10.1037/0003-066X.45.9.1043

Roediger, H. L., III. (2007). Transfer: The ubiquitous concept. In, H. L. Roediger, Y. Dudai, & S. M. Fitzpatrick (Eds.), Science of memory: Concepts (pp. 277-282). New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195310443.001.0001

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science*, *15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

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

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

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 1155-1159. https://doi.org/10.1037/0278-7393.31.5.1155

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal*

of Experimental Psychology: Learning, Memory, and Cognition, *36*, 233–239. https://doi.org/10.1037/a0017678

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York. https://doi.org/10.1002/9780470316696

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *49*, 289-318. https://doi.org/10.1146/annurev.psych.49.1.289

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*(7), 784-802. https://doi.org/10.1080/09658211.2013.831454

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99–115. https://doi.org/10.1016/j.jml.2014.03.003

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656. https://doi.org/10.1037/h0063404

Toppino, T. C. & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

Tran, R., Rohrer, D. & Pashler, H. (2014). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review*, *22*(1), 135-140. https://doi.org/10.3758/s13423-014-0646-x

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

https://doi.org/10.1016/S0022-5371(67)80092-6

van Ginkel, J. R., & Kroonenberg, P. M. (2014). Analysis of Variance of Multiply Imputed Data. *Multivariate behavioral research*, *49*(1), 78-91. https://doi.org/10.1080/00273171.2013.855890

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Lui, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*, 214-221. https://doi.org/10.1016/j.jarmac.2014.07.001

Zaromb, F. M., & Roediger, H. L., III (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*(8), 995-1008. https://doi.org/10.3758/MC.38.8.995

Zhang, M., Chen, X., & Liu, X. L. (2018). Confidence in accuracy moderates the benefits of retrieval practice. *Memory*. https://doi-org.ezproxy.lib.uh.edu/10.1080/09658211.2018.1529796

Appendix A

I. A schematic of condition and Biology Topic counterbalancing.

Repeated Testing	**Track 1**		**Track 2**		**Track 3**	
Position in Learning Phase	**1**	**2**	**1**	**2**	**1**	**2**
Learning Material	A	B	B	C	C	A
Condition	TR	R	TR	R	TR	R
	Track 4		**Track 5**		**Track 6**	
	1	**2**	**1**	**2**	**1**	**2**
	A	B	B	C	C	A
	R	TR	R	TR	R	TR
Varied Testing	**Track 1**		**Track 2**		**Track 3**	
	1	**2**	**1**	**2**	**1**	**2**
	A	B	B	C	C	A
	TV	R	TV	R	TV	R
	Track 4		**Track 5**		**Track 6**	
	1	**2**	**1**	**2**	**1**	**2**
	A	B	B	C	C	A
	R	TV	R	TV	R	TV

Note: Numerals represent the order (first, 1 or second, 2) of stimuli, while letters A (DNA), B (Food Webs), and C (Osmosis/Diffusion) represent specific passage A Biology topics that participants were exposed to. Finally, TR (test repetition), TV (test variation), and R (restudy) represent the learning condition for that particular Biology topic. This method for counterbalancing ensured equal Biology topic to learning condition assignment and controlled for order effects of learning condition and Biology topic.

Appendix B

I. A table showing the correlations between performance in the rest and restudy conditions and cognitive ability constructs.

Correlation Table for Outcomes and Cognitive Ability Constructs

	1	2	3	4	5	6	7	8
1. Test-Passage A	-							
2. Test-Passage B	.65***	-						
3. Restudy-Passage A	.39***	.53***	-					
4. Restudy-Passage B	.48***	.63***	.73***	-				
5. Science Reasoning	.42***	.48***	.30**	.49***	-			
6. Reading Comprehension	.41***	.51***	.45***	.54***	.59***	-		
7. Non-Verbal Reasoning	.26***	.31***	.35***	.40***	.55***	.33***	-	
8. Working Memory	.14	.12	.17	.22**	.30**	.02	.43***	-

Note: Test-Passage A = Performance in the test condition on passage A content; Test-Passage B = Performance in the test condition on passage B content; Restudy-Passage A = Performance in the Restudy condition on passage A content; Restudy-Passage B = Performance in the Restudy condition on passage B content; *p < .05, **p < .01, ***p < .001.