ABSTRACT

Title of Thesis:	THE BENEFITS OF TESTING: INDIVIDUAL DIFFERENCES BASED ON STUDENT FACTORS
	Alison Marie Robey, Doctor of Philosophy, 2017
Thesis Directed By:	Professor, Michael Dougherty, Department of Psychology

The testing effect, the notion that retrieval practice compared to restudying information leads to greater and longer retention, is one of the most robust findings in cognitive science. However, not all learners experience a benefit from retrieval practice. Many manipulations that influence the benefits of the testing effect have been explored, however, there is still much to learn about potential individual differences in the benefits of retrieval practice over restudy. As the testing effect grows in popularity and increasing numbers of classrooms begin implementing retrieval practice, it is essential to understanding how students' individual differences and cognitive abilities contribute to the effect. For my dissertation, I explore how students' cognitive abilities, specifically, episodic memory, general fluid intelligence, and strategy use, relate to the benefit of retrieval practice. In Study 1, I developed a new measure to simultaneously capture two aspects of strategy use: variation in *what* strategies learners use and variation in *how*

learners use strategies. In Study 2, I examine how these two types of strategy use, along with episodic memory and general fluid intelligence can be used to predict the magnitude of the testing effect. Converging evidence from multiple analyses suggests variation in *how* learners use strategies was the only individual difference to influence the benefit learners receive from retrieval practice. More specifically, learners who are less adaptive and flexible in their strategy use show a greater benefit than more skilled strategy users. These findings have implications both for improving existing theories of the mechanisms of the testing effect and for determining how to best incorporate retrieval practice into classroom settings.

THE BENEFITS OF TESTING: INDIVIDUAL DIFFERENCES BASED ON STUDENT FACTORS

by

Alison Marie Robey

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
Of the requirement for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Michael Dougherty, Chair
Professor Donald J. Bolger
Professor Tracy Riggins
Dr. Scott Roberts
Professor L. Robert Slevc

Acknowledgements

I would like to thank my graduate school advisors, Dr. Michael Dougerty & Tracy Riggins, all of my lab mates, Leslie Rollins, Sarah Blankenship, Rose Nguyen, & David Ampofo, and all of the undergraduate research assistants who helped with data collection and processing for my dissertation, especially Carrie Aposporos and Victoria Owns. Additionally, Oluwafunmilayo Ayeni, a summer research student who collected and processed all of the data from Study 1b. Finally, I would like to thank my friends, family, and boyfriend, Andrew Lawson, for all of their support.

Table of Contents

Acknowledgements....ii
Table of Contents....iii
List of Tables....v
List of Figures....vi
Chapter 1: Introduction....1
Theories of the Testing Effect....2
Individual Differences and the Testing Effect....6
Individual Differences Academic Skills....6
Individual Differences in Cognitive Abilities....8
Individual Differences in Strategy Use....9
Current Study....13
Chapter 2: Study 1 – Development of Strategy Variability Measure....16
Chapter 3: Study 1a....17
Participants....17
Materials....17
Procedure....18
Retrospective Strategy Report Scoring....19
Coding of other responses....19
Variability in what strategies learners used....20
Variability in how strategies are used....20
Results....21
Chapter 4: Study 1b....23
Participants & Materials....23
Procedure....23
Retrospective Strategy Report Scoring....24
Coding of other responses....24
Variability scoring....24
Results....24
Chapter 5: Study 1c....26
Participants & Materials....26
Procedure....27
Retrospective Strategy Report Scoring....27
Coding of other responses....27
Variability scoring....28
Results....28
Chapter 6: Study 1 Combined....30
Results....30
Chapter 7: Study 1 Discussion....31
Chapter 8: Study 2 – Individual Differences in Retrieval Practice....32
Participants....32
Materials....32
Paired Associates Testing Effect Task....32
Episodic Memory Measures....34
General Fluid Intelligence Tasks....35
Retrospective Strategy Report....36

Procedure....37
General Analysis Plan....38
Results....39
Study 2 Discussion....52
Chapter 9: General Discussion....55
Implications for theories of the testing effect....57
Real World Applications....59
Appendices....61
References....73

List of Tables

Table 1. Descriptive comparisons of the generated word lists

Table 2. Correlation matric and descriptives of all measured variables

Table 3. Results of dominance analysis both with and without outliers

List of Figures

Figure 1. The design of a typical testing effect paradigm

Figure 2. Individual difference patterns found in previous studies

Figure 3. Scatterplots showing the relationship between measures of strategy variability and recall with outliers removed, Study 1a.

Figure 4. Scatterplots showing the relationship between strategy variability and recall accuracy with outliers removed, Study 1b.

Figure 5. Scatterplots showing the relationship between strategy variability measures and recall accuracy with outliers removed, Study 1c.

Figure 6. Scatterplots showing the relationship between strategy variability and recall accuracy with outliers removed, Study 1 Combined.

Figure 7. Boxplots showing final recall accuracy for items that were restudied and items that received practice retrieval

Figure 8. The relationship between episodic memory and final recall accuracy for items receiving restudy and retrieval practice

Figure 9. The relationship between general fluid intelligence and final recall accuracy for items receiving restudy and retrieval practice

Figure 10. The relationship between strategy composite scores an final recall accuracy for items receiving restudy and retrieval practice

Figure 11. The relationship between CWS ratios and final recall accuracy for items receiving restudy and retrieval practice without outliers

Figure 12. Structure SEM model standardized results with outliers included. Solid line represent significant paths

Figure 13. The relationship between all four cognitive abilities and the testing effect represented by the difference in final retrieval accuracy for items that received retrieval practice and the items that received restudy

Figure 14. The testing effect for each of the five word pair categories

Figure 15. RT distributions for correct and incorrectly recall items after restudy or retrieval practice

Chapter 1: Introduction

For much of formal education's history it was thought that learning occurred during study and that tests provided a method for evaluating or assessing what a student had learned. Research in cognitive science however has shown that testing may in fact be one of the best methods for ensuring the long-term retention of material. The testing effect, the notion that retrieval practice compared to restudying information leads to greater and longer retention, is one of the most robust findings in cognitive science (see Roediger & Butler, 2011 for review). Beyond the laboratory, recent research on the testing effect confirms that it effectively improves students' academic outcomes in authentic educational settings (e.g., Lyle & Crawford, 2011; Carpenter, Pashler, & Cepeda, 2009). The testing effect leads to improvements on a variety of educational material including: foreign-language equivalent pairs (e.g., Carpenter, Pashler, Wixted, & Vul, 2008), multiple choice tests (e.g., Marsh, Roediger, Bjork, & Bjork, 2007), and complex reading passages (Roediger & Karpicke, 2006). Many factors influencing the benefits of the testing effect have been explored, such as, the optimal amount of retrieval practice (e.g., Rawson & Dunlosky, 2011), the optimal delay between retrieval attempts (e.g., Landauer & Bjork, 1978), and whether or not feedback should be provided (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005). However, factors associated with the students themselves and individual differences in the benefits of retrieval practice over study have not received the same attention. While the testing effect is a robust finding in general, not all learners show improvements as a result of retrieval practice. For example, Brewer and Unsworth (2012) found that only two-thirds of students benefited from testing, whereas 12% showed no difference between restudy and test, and 21% preformed **worse** with testing compared to restudy. It is currently unknown however if these differences in effectiveness are due to random chance or something systematic. As the testing

effect grows in popularity and increasing numbers of classrooms begin implementing practice retrieval, it is essential to understand how students' individual factors and cognitive abilities contribute to the effect. For my dissertation I explore whether a set of cognitive abilities, specifically episodic memory, fluid intelligence, and variation in strategy use, can be used predict the magnitude of the testing effect. These findings have implications both for improving existing theories of the mechanisms of the testing effect and for determining how to best incorporate retrieval practice into classroom settings.

Theories of the Testing Effect

The design of a traditional testing effect study is outlined in Figure 1. Participants are first presented with to-be-learned material, such as word pairs in an initial learning phase. After initial learning, participants begin a study phase where they are given the opportunity to restudy one half of the material and are given retrieval practice on the remaining half. Studies vary as to whether the correct answer is provided on retrieval practice trials (Pashler et al, 2005). Although the testing effect is typically observed both with and without feedback, effects are larger when feedback is provided (McDermott, Kang, & Roediger, 2005). A final test is given on all material, and performance is compared for material that was restudied versus material that was tested. Traditionally, it is found that material that has been given practice retrieval is better retrieved at final test than the material that was restudied. It should be noted however that when the delay between the Study and Final Test phases is exteremly short (i.e. less then 5 minutes), some findings have been reversed with better preformance observed for restudied items, likely due to the effects of massed-practice (e.g., Roediger & Karpicke, 2006; Balota, Duchek, & Paullin, 1989).

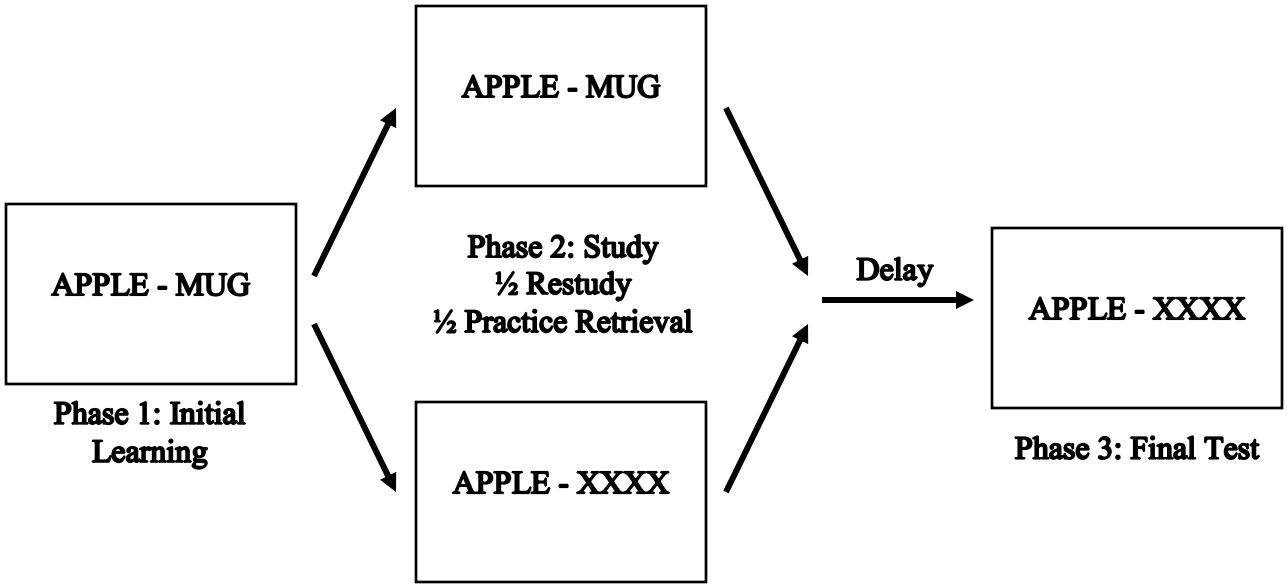


Figure 1. The design of a typical testing effect paradigm.

Three predominant theories of the testing effect have been proposed. It should be noted that these theories may be overlapping rather than competing. First, Morris and colleagues' (1977) Transfer-Appropriate Process theory states that retrieval will be more successful when conditions at test are similar to those experienced during learning. Applied to the testing effect, learning that occurs during practice retrieval would be more similar to later testing than learning that occurs during restudy. To test the validity of the Transfer Appropriate Process Theory as an explanation of the testing effect, Carpenter and DeLosh (2006, Study 1) manipulated the type of retrieval task participants experienced during retrieval practice and final test to be one of three types: free recall, cued recall, or recognition. The match between practice retrieval task and final retrieval task had no influence on the benefit of retrieval practice. Instead, it was found that free recall retrieval practice always led to greater benefits than recognition retrieval practice, regardless of final test type (Also see Glover, 1989).

In their study, Carpenter & DeLosh (2006) found that retrieval practice tasks involving less accessible material (i.e., free recall) lead to greater testing effects than tasks involving more

accessible material (i.e., recognition). This leads to the second theory of the testing effect, the Elaborative Retrieval Hypothesis (ERH, also see McDaniel, Kowitz & Dunay, 1989). This hypothesis posits that retrieval of information from memory leads to an elaboration of the existing memory trace and the less accessible the information is at retrieval, the more elaboration will occur. Elaboration of a memory trace may occur by adding additional retrieval cues, leading to a greater number of routes for the material to be successfully retrieved in the future.

Related to the ERH, the Retrieval Effort Hypothesis proposes that difficult but successful retrieval leads to a greater strengthening of the memory trace than easy but successful retrieval, suggesting that both the accessibility of the material and the difficulty of retrieving the material play an important role determining the benefits of retrieval practice (Pyc & Rawson, 2009; also see Desirable Difficulties). Tests of the ERH and retrieval effort hypothesis have found that manipulations that increase the ease of practice retireval, such as increasing the number of cues provided at practice retrieval (Carpenter & DeLosh, 2006, Study 2), decreasing the interval between initial learing and practice retrieval, and increasing the initial learning criterion (Pyc & Rawson, 2009), decrease the long term benefits of the testing effect.

To directly test whether the benefits of practice retrieval were comparable to the benefits of using an elaboration strategy during encoding, Karpicke & Smith (2012) made a small modification to the traditional testing effect paradigm. During the study portion of the design, participants were instructed to use an elaborative encoding strategy on the restudy trials in the form of either visual imagery or verbal elaboration. They found that retrieval practice improved later memory preformance above and beyond an elaborative strategy, suggesting that the testing effect must occur due to mechanisms other than just improved elaboration. Instead, they proposed that the benefits of retreival practice are due to a process that is unique to retrieval such

as the generation of potential retrieval cues. The third theory of the testing effect is the Episodic Context Account of retrieval practice, which proposes that testing enhances later memory performance by reinstating and updating contextual representations of items in memory, making future memory search processes more efficient (Karpick, Lehman, & Aue, 2014). To test the Episodic Context Account of retrieval practice, Whiffen & Karpicke (2017) had particpants complete a retrieval practice version of a list discrimination task. All participants were first presented with two lists of six words. Then, half of the participants passively restudied all of the words and half completed list discrimination where they were asked to idenitfy whether each word came from the first or second list. This task is novel, in that it does not require a memory test per se, but it does require the participant to remember and reinstate the intial context to complete the list discrimination decision. At final test, participants who completed the list discrimination task not only remembered more words from the lists, but also used more temporal information when retrieving the words. These findings support the theory that retrieval enchances later memory preformance by reinstating and updating contextual information (in this case temporal information) related to the original study episode (also see Lehman, Smith, & Karpicke, 2014).

Although these theories provide insight into why the testing effect may occur, much work is still needed as no existing theory accounts for potential individual differences and how characteristics of the learner may impact the magnitude of the testing effect. Past studies of these hypotheses have focused on manipulations of the stimuli (i.e., word pairs that are semantically related will be easier to retrieve than unrelated pairs; Pyc & Rawson, 2009); however, both differences across items *and* differences between learners will likely impact what benefits are observed. Individual differences are an important, and relatively unexplored, element of retrieval

practice and it is critical to determine if differences between learners impact the benefits of retrieval practice.

Individual Differences and the Testing Effect

Several studies have explored individual differences in student factors and their impact on the testing effect (cf. Bouwmeester & Verkoeijen, 2011; Brewer & Unsworth, 2012; Carpenter, Lund, Coffman, Armstrong, Lamm, & Reason, 2015; Carrol, Campbell, Ratcliffe, Murnane, & Perfect, 2007; Chan, 2009; Karpicke, Blunt, & Smith, 2016; Ma, Yang, Yanru, & Zhao, 2016; Pan, Pashler, Potter, & Rickard, 2015, Spitzer, 1939). These studies have found mixed results with some studies suggesting only high ability students show a testing effect (Carpenter et al., 2015), some finding benefits for all students, though some greater than others (Spitzer, 1939; Brewer & Unsworth, 2012), and some finding equal benefits across all students (Carroll et al., 2007; Chan, 2009; Karpicke et al., 2016; Ma et al., 2016; Pan et al., 2015).

Additional research has explored the impacts of learners' age (e.g., Tse, Balota, & Roediger, 2010) and learners with brain injuries versus healthy learners (e.g., Balota, Duchek, Sergent-Marshall, & Roediger, 2006). These factors, however, are outside the scope of this paper and will not be discussed further. Below I summarize the current findings on individual differences within two categories, the impact of students' academic related skills and the impact of students' cognitive abilities. Although the focus of my dissertation is undergraduate learners, several of the studies described below make use of younger samples.

Individual Differences in Academic Skills. One interesting question is whether higher or lower achieving students benefit more from retrieval practice. Carpenter and colleagues (2015) attempted to answer this question in an introductory undergraduate biology course. All students completed an in-class activity that involved either copying definitions (restudy) or

retrieving definitions (retrieval practice) of vocabulary words that appeared in an assigned text book chapter. The students were tested on this material later in the semester and performance on copying versus retrieving was compared. Based on their class performance up until that point, students were split into three equally sized groups, classified as either low, medium, or high achievers. Researchers found that high achieving students benefited more from retrieval practice, whereas low achieving students benefited more from copying definitions. These results suggest that high achieving students benefit more from retrieval practice, however there was a flaw in the design that could have led to this effect. This design did not contain a true initial learning phase, and instead the material was from a chapter listed on the course syllabus. It is not surprising to assume that many students, particularly low performing students, may not have read the text book chapter prior to class. If this occurred, then the students would not have had previous exposure to the material, and successful retrieval would not have been a feasible expectation. As a key aspect of retrieval practice is being able to retrieve the information, no benefit would be expected without initial learning. When examining performance on the in-class activity the high achieving students who completed retrieval practice exercise were able to recall almost 40% of the definitions whereas the low achieving students recalled less than 10%. In contrast, all three groups performed equally well on the activity that required copying the definitions.

Although not the goal of their study, Spitzer (1939), found that though all 6th graders benefited from retrieval practice compared to study, students with higher reading comprehension scores benefited more from retrieval practice than students with lower scores. In contrast, Karpicke et al. (2016) found that reading comprehension did not predict the magnitude of 10-year old's testing effect. Additionally, Carroll and colleagues (2007) have shown that the

benefits of retrieval practice are equivalent across differing levels of prior knowledge (also see Ma et al., 2016).

Individual Differences in Cognitive Abilities. Brewer and Unsworth (2012), had participants complete a large battery of cognitive assessments, including measures of working memory (WM), episodic memory (EM), attention, and general fluid intelligence (gF). Using composite measures of these four abilities, they found that both EM and gF predicted the magnitude of the testing effect, as defined by the difference in accuracy for restudied items versus tested items. Participants who scored higher on EM and gF measures showed smaller benefits of testing. Brewer and Unsworth explained their results using the Elaborative Retrieval Hypothesis, suggesting that participants with higher EM abilities were better able to use elaborative strategies during both restudy and testing, minimizing the elaborative benefits typically observed with retrieval practice. No relations were found for either WM (also see Chan, 2009) or attention. Additionally, in a sample of 10-year old children, Karpicke and colleagues (2016) found that processing speed did not influence the magnitude of children's testing effect.

Pan and colleagues (2015) attempted and failed to replicate Brewer and Unsworth's (2012) finding that EM was related to the magnitude of the testing effect in two independent studies. Although the designs were similar, Pan and colleagues (2015) made two key modifications. First, Pan and colleagues (2015) only examined the relationship between EM and the testing effect; WM, attention, and gF were not included because they were either not related to the testing effect in the original study (WM, attention) or because they could not determine a theoretical explanation (gF). Second, Pan and colleagues (2015) inter-mixed restudy and test trials, whereas Brewer and Unsworth (2012) administered the restudied and tested items as two separate blocks. Inter-mixing trials may dilute the benefits of retrieval practice if participants

inadvertently use strategies in one condition to complete trials in the other condition. This is especially likely to happen if participants discover that retrieval practice improves learning. However, recent work by Abel and Roediger (2017) has shown that the testing effect is equivalent across blocked and intermixed restudy. Given the mixed findings of the Brewer and Unsworth (2012) and the Pan et al (2015) studies, the question of whether EM abilities are related to the magnitude of the testing effect is still largely unanswered. In addition, the finding that gF is related to the testing effect has yet to be replicated. Study 2 of my dissertation explores the role of EM and gF as individual differences explaining the magnitude of the testing effect to attempt to replicate either Brewer & Unsworth (2012) or Pan et al. (2015). Additionally, I examine the role of a currently unexplored cognitive factor, variation in strategy use, using a new method developed in Study 1.

Individual Differences in Strategy Use

An additional individual difference that has yet to be explored in relation to the testing effect is student strategy use. As alluded to earlier, one potential explanation for why some learners benefit more from retrieval practice than others is due to differences in strategy use. One could even view retrieval practice as a forced strategy, and therefore it would be unsurprising that students who naturally engage in other beneficial strategies would not benefit as much from retrieval practice as those who do not, as they would be expected to perform better in the restudy condition.

It has long been known that engaging in memory strategy use during encoding improves later memory performance (e.g., Bower, 1970), and that training in new strategies can improve performance in under-performing groups (e.g., Naveh-Benjamin, Brav, & Levy, 2007). It is also known, however, that there is great variation in strategy use between individuals (Dunlosky &

Hertzog, 1989; Kirchhoff, 2009). Variation in strategy use can be conceived in two different forms: (1) variation in *WHAT* strategies are used and (2) variation in *HOW* strategies are used. However, to date, there are no available methods that will simultaneously measure both of these factors.

When learning simple material such as word pairs there are many potential strategies learners may engage in such as rote rehearsal, visual imagery, or verbal elaboration. The type of strategy a learner chooses to engage in will impact what they later remember. For example, deeper encoding strategies, such as those that tap into words' semantic meaning (e.g., verbal elaboration), lead to better remembering than shallow strategies such as rote rehearsal (Dunlosky & Hertzog, 2001). Two common methods currently exist for measuring *what* strategies learners use on a particular task (for Review see Kirchhoff, 2009; Richardson, 1998). First, self-report methods allow learners to directly report to the experimenters what strategies they believed they used. Learners can report their strategies by selecting from a list of pre-selected strategies (Camp Markley, & Kramer, 1983; Dunlosky & Hertzog, 2001; Paivio, & Yuille, 1969; Paivio, Yuille, & Smythe, 1966; Richardson 1978) or describe their strategies for each stimuli in an open-ended fashion (Martin, 1965; Roberts, 1968; Stoff & Eagle, 1971). Additionally, either of these options can either occur concurrently, after each stimuli during the initial learning phase, or retrospectively, at the end of the task after final retrieval is complete.

An alternative method for measuring what strategies participants use is to manipulate the stimuli between trials and then conduct clustering analyses. The most common use of this method occurs with verbal list learning paradigms. In these tasks participants hear lists of many items and then after a delay are asked to recite as many items as they can remember. The experimenters can manipulate factors such as whether the words fall into easily identifiable

categories. For example, if one set of words contains all unrelated words and another set contains words from four distinct categories, experimenters can observe the order and number of words retrieved and if more words are retrieved when the words are from categories the experimenters can conclude those subjects were using a semantic clustering strategy (Bousfield, 1953). Similar manipulations can be done to capture the use of visual imagery. Although this method is able to capture spontaneous strategy use without prompting, it is limited to strategies that can be targeted by stimuli manipulation and only one strategy can be tested at a time.

In addition to variation in *what* strategies learners use, learners also vary in *how* they use strategies, and depending on the to-be-learned material, different strategies may be more or less beneficial (McGee, 1980). For example, a visual imagery strategy might be highly beneficial for a task that requires remembering highly concrete words (e.g., throw, car), but potentially detrimental for a task that requires remembering more abstract words (e.g., valor, honest). A learner who is flexible with their strategy use will be able to adapt and use different strategies depending on the situation to gain the most benefit.

Currently, no methods exist to measure *how* learners engage in strategy use during encoding, but similar methods exist in other domains, specifically expertise. Within the expertise literature many characteristics have been used to define what it means to be an expert at a particular skill. As an example, consider sommeliers (i.e., expert wine tasters). One characteristic an expert is thought to have is consistency (Einhorn, 1972, 1974). A sommelier rates many wines and is expected to be consistent in their ratings in that if they taste the same wine multiple times they will give it the same rating each time. Additionally, experts are also expected to be discriminating (Hammond, 1996), wherein a sommelier would discriminate between different wines and give them different ratings. Both of these abilities (consistency and discrimination) are

characteristics of experts, but neither one can independently be used to categorize someone as an expert (Shanteau, Weiss, Thomas, & Pounds, 2002). A non-expert could either use the same strategy for every stimuli and appear highly consistent, or use a different strategy for every stimuli and appear highly discriminant. To overcome this obstacle, Shanteau et al., (2002) developed the Cochran – Weiss – Shanteau (CWS) ratio that incorporates both discrimination and consistency into a single metric, see equation (1).

$$CWS\ Ratio = \frac{Discrimination}{Consistency} \quad (1)$$

This metric has been useful in describing and predicting skilled performance across a number of domains, including auditing, livestock judging, hiring (Shanteau et al., 2002), aviation performance and weather prediction (Pauley, O'Hare, & Wiggins, 2009; Roth & Mavin, 2015; Wiggins, 2014), medical and clinical diagnoses (Loveday, Wiggins, Searle, Festa, & Shell, 2013; Witteman & Tollenaar, 2012; Witteman, Weiss, & Metzmacher, 2012) and teacher's grading (Canal, Bonini, Miccioli, & Tentori, 2012) to name a few. I argue that variation in strategy use should be characterized in a similar way. For example, skilled learners will consistently select the most appropriate strategy for the task at hand, applying their strategies both consistently and discriminately. To illustrate, when given a set of items to learn, such as word pairs, not all items will be equal. Different word pairs will have different characteristics and a skilled strategy user should be able to pick up on these characteristics and use strategies in an adaptive way. Items that share characteristics will likely benefit most from the same strategies (i.e. consistency), and items with different characteristics will likely benefit most from different strategies (i.e., discrimination). A skilled strategy user should be consistent and for a particular type of stimuli use the same strategy reliably, while also differentiating between similar cases and for different types of stimuli use different strategies. For example, a skilled strategy user might realize that

half of the word-pairs contain highly concrete items and choose to consistently use a visualization strategy on all of these items. Additionally, this same leaner may notice that the other half contain abstract words and decide to use a verbal association strategy for all of those pairs. This would result in the skilled strategy user having high discriminability (they used different strategies for different types of stimuli), high consistency (they used the same strategies within their defined categories), and therefore, a large CWS ratio. A learner who was not as discriminate or not as consistent would have a lower CWS ratio.

Additionally, this measure has the benefit that it does not require a definitive "best" strategy to be used for a particular situation. Instead, it allows for individuals to select what strategies work best for them in specific situations without assuming the same optimal strategy should be used by everyone. One learner may prefer to use a visualization strategy and another may prefer to use verbal associations, however both of these learners could have the same CWS ratio.

In Study 1 of my dissertation I test the hypothesis that that CWS ratio can be applied to the domain of strategy use. It is expected that learners with higher CWS ratios will have better memory accuracy at retrieval. In Study 2 I test both variation in *what* strategies are used and variation in *how* strategies are used as individual differences in the testing effect. It is expected that learners who engage in more beneficial strategies would show smaller testing effects, because their strategies allow them to perform better on items that are restudied compared to learners who use no or less beneficial strategies.

Current Study

The current project explores how episodic memory, general fluid intelligence, and variability in strategy use relate to the benefits received from retrieval practice. In Study 1, I

tested whether a retrospective strategy report could be used to capture variability in strategy use and predict memory performance. Specifically, I explored both variability in *what* strategies are used with a strategy composite score, and variability in how strategies with the CWS ratio. This metric has been used to measure expertise in many domains, but has never been applied to the domain of strategy use. Participants completed a standard cued-recall task. During encoding they were presented with word pairs then during retrieval they were presented with the first word (the cue) and asked to retrieve the second (the target). Participants then viewed all word pairs an additional time and made retrospective strategy judgments related to what strategy they thought they used for each pair (e.g., Dunlosky & Hertzog, 2001). Word-pairs were manipulated to consist of 5 different types: Related – High Imageability, Related – Low Imageability, Unrelated – High Imageability, Unrelated – Low Imageability, and Nonsense Words, to allow for measures of consistency and discrimination. Measures of *what* and *how* participants used strategies were then used to predict final memory performance.

In Study 2 I explored how cognitive abilities (episodic memory, general fluid intelligence, and strategy use) relate to the benefits of retrieval practice. Different from the replication attempt by Pan and colleagues (2015), I used blocked restudy and retrieval trials rather than random in order to explore the impact of task order on the magnitude of the testing effect. The new method developed in Study 1 was used to collect measures of strategy variability related to both *what* strategies learners use and *how* learners use strategies (i.e., CWS ratio). Previous research suggests three potential patterns of results (See Figure 2). First, retrieval practice benefits all learners equally and none of the selected variables will serve as individual difference measures. Second, retrieval practice benefits lower skilled learners more, decreasing the gap in overall performance. Third, retrieval practice benefits higher skilled learners more,

increasing the gap in overall performance. Additionally, it is possible that different individual factors (i.e., EM, gF, *what* strategies are used, and *how* strategies are used) may relate to the testing effect in different ways. Previous research examining individual difference in students' factors has showed all three of the above patterns, therefore, no explicit hypothesis for the expected relation of these individual differences was made.

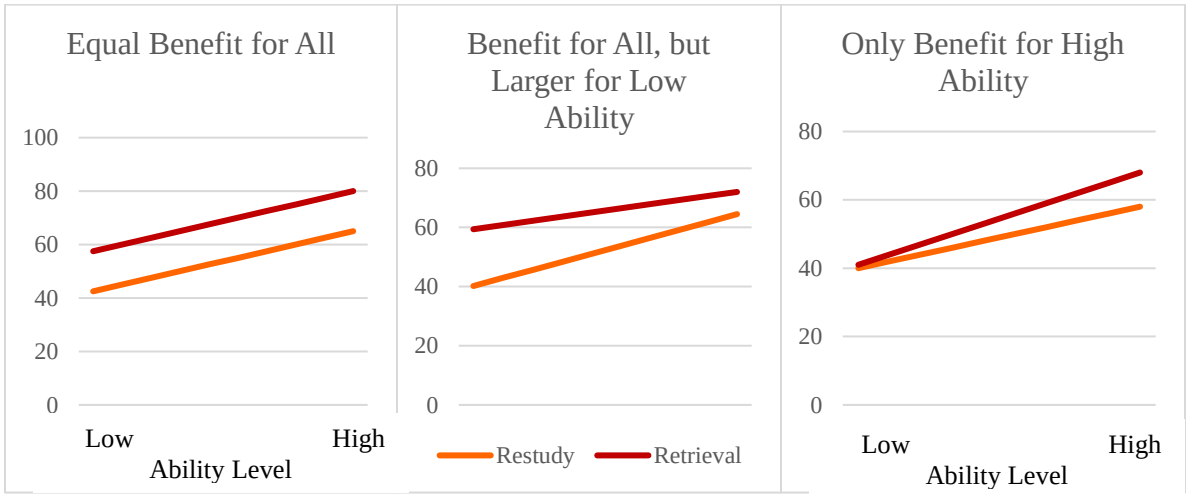


Figure 2. Individual difference patterns found in previous studies.

Chapter 2: Study 1 – Development of Strategy Variability Measure

The goal of Study 1 was to explore a newly developed retrospective report designed to assess strategy variability. Over a series of three studies, strategy variability was assessed in two ways. First, *what* strategies participants reported using during learning, and second, *how* learners used strategies. This new method combines previously used measures to allow for the measurement of two types of strategy variability simultaneously. Variability in *what* strategies participants use was captured with a strategy composite score calculated by averaging the benefit scores of reported strategies (a method similar to that used by Dunlosky & Hertzog, 2001). Variability in *how* participants use strategies was captured using a modified CWS ratio, a metric that has been used to measure expertise in a variety of domains (e.g., Canal et al., 2012; Loveday et al., 2013; Pauley et al., 2009; Roth & Marvin, 2015; Shanteau et al., 2002, Witteman, 2012; Witteman & Tollenaar, 2012; Wiggins, 2014). The CWS ratio has been used in several domains, but never before applied to the domain to strategy variability. This series of studies examines how both variation in *what* strategies are used and variation in *how* strategies are used related to final memory performance.

Chapter 3: Study 1a

Participants

Thirty-one participants were recruited from the University of Maryland psychology subject pool. All participants received course credit for participation.

Materials

One key aspect of the CWS ratio is examining strategy use both within and between categories. In order to provide differing categories for strategies to be used on, the word pairs of a standard cued-recall tasks were manipulated to consist of five different types: Related – High Imageability, Related – Low Imageability, Unrelated – High Imageability, Unrelated – Low Imageability, and Nonsense Words. A total of 40 word pairs were created, with 8 pairs per category.

All words were generated using the MRC Psycholinguistic Database (Wilson, 1988). Two sets of word lists were generated using the following specifications; number of letters: 3 – 7, number of syllables: 1 – 2; Familiarity: 488 – 636 (mean to + 1.5sd), and words were limited to nouns. One set of words contained only high imageability words (Imageability rating 588 to 669, +1sd to the maximum). The other set contained only low imageability words (Imageability rating 200 to 343, the minimum to -1 sd). Kucera-Francis Frequency values were also included in the output.

Next, all generated words from the low imageability list were entered into the Edinburgh Associative Thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973). If one of the top three associated responses was also on the generated word list, those words were matched as a related pair. Once all low imageability words that could be matched were matched, the 8 pairs with the highest associative strength were selected to be the Related-Low Imageability pairs. The same

process was then repeated with the high imageability list. Eight associated word pairs were selected for the Related-High Imageability group, with the stipulation that the mean associative strength, familiarity, frequency, and number of letters and syllables needed to be comparable to the selected Related-High Imageability pairs.

All words from the generated lists that were not assigned to a related word pair set were then randomly matched. Eight pairs from each list were selected to complete the Unrelated-High Imageability and Unrelated-Low Imageability lists, again with the stipulation that the mean familiarity, frequency, and number of letters and syllables needed to be comparable to the already created lists. This process resulted in four word pair lists that were equivalent in familiarity, frequency, number of letters, and number syllables, but differed in imageability and associative strength (See Table 1). Additionally, a set of 8 pairs of nonsense words was generated using the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). A complete list of word pairs is provided in Appendix A.

Table 1. Descriptive comparisons of the generated word lists

	Familiarity	Frequency	Letters	Syllables	Associative Strength	Imageability
Related-Low Imageability	575.44	208.56	5.19	1.63	0.22	300.63
Unrelated-Low Imageability	546.88	209.13	5.00	1.81	0	310.75
Related-High Imageability	573.44	176.44	4.69	1.38	0.21	601.75
Unrelated-High Imageability	574.06	168.19	4.56	1.44	0	602.25

Note: All statistics were taken from the MRC Database or Edinburgh Associative Thesaurus

Procedure

Participants completed a standard cued recall paradigm with an encoding portion, where all word pairs were presented, one at a time, for 6 seconds, immediately followed by a self-paced retrieval portion, where participants were presented with the cue word and tasked with retrieving the target. During encoding word pair categories were blocked so that participants were always

presented with 4 pairs from the same category in a row to encourage participants to notice the similarity and promote consistent strategy use within a category. The blocks were randomized such that participants were presented with two sets of 4 random pairs from each of the 5 categories, with the order of the categories randomized. During retrieval, word pair presentation was random.

Immediately following the cued-recall task, participants completed a retrospective strategy report of all word pairs. Participants were presented with all word pairs again and made a forced choice decision regarding which of 5 memory strategies they felt they used most *when originally learning the word pair*. Strategy options included: Verbal Association, Visualization, Rote Rehearsal, None, and Other (Dunlosky & Hertzog, 2001). If participants select "Other", they were prompted to explain what strategy they felt they used.

Retrospective Strategy Report Scoring

Coding of other responses. A total of 86 other responses were given in Study 1a. All other responses were first independently coded by three coders to determine if the strategies reported were truly "Other" responses or if they belonged in one of the provided categories. For example, in Study 1a it was not uncommon for a participant to select "Other" and then respond that they repeated the word pair to themselves, a response that clearly belonged in the "Rote Rehearsal" category. As the coding scheme required categorical responses, Cohen's Kappa was calculated pair-wise, as a measure of inter-rater reliability (Cohen, 1960). Pair-wise Kappas were as follows: Kappa (95%CI); .93 (.86 to 1), .80 (.86 to .92), .78 (.65 to .90). Based on the magnitude guidelines of Landis and Koch (1977) all pair-wise rating had at least substantial agreement. For any responses did that did not receive full agreement among the three coders, a

code reported by at least two of the three coders was selected. There were no responses where all three coders disagreed.

For the responses that remained coded as "Other" responses, all three coders gave an additional score of either +1, -1, or 0, indicating where they believed the strategy reported was beneficial, inadequate, or if helpfulness was unclear. In Study 1a, six responses remained coded as "Other". All coders were in complete agreement on these scores.

Variability in what strategies learners used. Strategy composite scores were calculated for each participant by taking the average score of all strategies they reported using. Verbal Association and Visual Imagery responses were scored as a positive one (+1) as they are generally known to be beneficial to memory performance. Rote Rehearsal and None responses were scored as a negative one (-1), as they are either unhelpful for prevent learners from engaging the strategies that would be more beneficial (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Scores for "Other" responses were added in based on the results of the coding described above.

Variability in how strategies are used. A CWS ratio was calculated for each participant to assess variation in *how* strategies were used. First, measures of "unalikeability" in strategy use (i.e., variability of a categorical variable, Kader & Perry, 2007) were calculated within each word pair category for each participant using the equation below,

$$u_2 = 1 - \sum_i p_i^2 \qquad (2)$$

where p_i represents the proportion of a specific option within category i. Next the variability in the proportion of each strategy type between categories was calculated using a sample measure of standard deviation. Using the variability between categories as the measure of discriminability and the variability within as a measure of consistency, CWS ratios were calculated for each

participant. An example of this process on three simulated participants is provided in Appendix B.

Results

All analyses were run using Bayesian model comparison, with proportion data transformed using the logit transformation. This approach provides an index of the degree of support for any model of interest (the Bayes Factor, BF), including the null model. BFs provide the evidence of support for alternative or more complex models relative to a null or reduced model, such that BFs greater than 1 represent support for the more complex model and BFs less than one represent support for the reduced model. Additionally, BFs represent a continuum with values further from 1 representing greater support. A BF can be interpreted as an odds ratio with a BF of 2.0 representing 2 to 1 odds in favor of, or a 50% chance of, the more complex hypothesis. Although BFs are not meant to be interpreted with "cut-points", it is generally accepted that BFs between 3 and .33 are considered inconclusive results (Kass & Raftery, 1995).

Boxplots of all measured variables are provided in Appendix C. These boxplots were used to identify potential outliers. More specifically, data points were considered potential outliers if they were greater than 1.5 times the interquartile range. For variables that included potential outliers, analyses were run both with and without the outliers included.

Analyses revealed a strong correlation between Strategy Composite scores and final recall accuracy, $r = 0.49$, $BF = 8.97$, see Figure 3. Results were inconclusive regarding a correlation between CWS and final recall accuracy, $r = 0.41$, $BF = 2.92$. Removing potential outliers did not change the inconclusiveness of the results, $r = 0.31$, $BF = 0.94$.

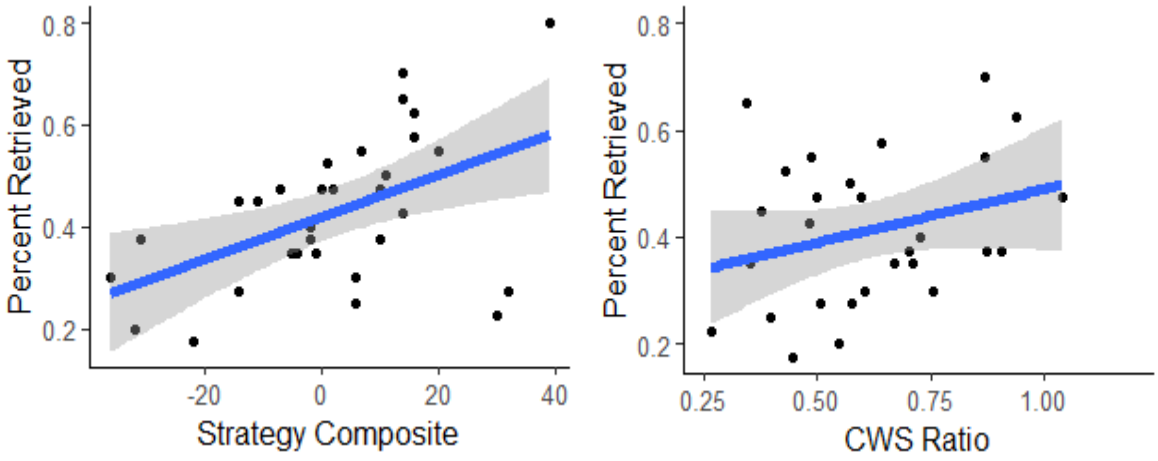


Figure 3. Scatterplots showing the relationship between measures of strategy variability and recall accuracy with outliers removed, Study 1a.

Results of Study 1a show support for a relationship between Strategy Composite scores and memory recall, however results related to CWS scores are inconclusive. One potential problem with the CWS scores in Study 1a however, is the large number of strategy responses participants reported as "Other". It is expected that this occurred due to a lack of explanation or definitions provided to the participants regarding what types of strategies should be included in each category. Study 1b served as a replication of Study 1a with the change that thorough definitions and examples of each strategy type were provided to participants before they made their judgments.

Chapter 4: Study 1b

All data for this study was collected and processed by our lab's summer research initiative student Oluwafunmilayo Ayeni, under the guidance and supervision of the author.

Participants & Materials

Forty-four participants were recruited from the University of Maryland psychology subject pool. All participants received monetary compensation for participation. All materials were the same as those used in Study 1a.

Procedure

The procedure was the same as that used in Study 1a, with the exception that after retrieval, but before giving their strategy responses, all participants were provided with definitions and examples for each strategy category. Definitions provided were as follows; Verbal Association: "Forming a relationship between the two words that is verbal in nature. This can include various relations such generating a sentence that uses the two words, thinking of the lyrics from a song, or linking the words verbally to a past experience"; Visual Imagery: "Using visualization to imagine the words in your mind. This can involve various things such as generating and visualizing a new scenario that involves the two words or imaging a situation you have been in before involving the two words"; Rote Rehearsal: "Repeating the two words to yourself over and over"; None: "There could be many reason that you don't engage in any strategy use. This includes thinking the word pair is so easy to remember that you don't need to use a deliberate strategy, trying to use a previously mentioned strategy and failing, or any other reason where you didn't engage a deliberate strategy"; Other: "This option should only be selected if you feel you used a strategy that does not fit into any of the previously listed strategies. Please note however that a strategy must be something you are actively doing to try

and remember the word pair. This means that simply stating a fact about the pair, for example "They are synonyms" does not count as a strategy and should instead be marked as None".

Retrospective Strategy Report Scoring

Coding of Other Responses. The strategy definitions helped to reduce "Other" responses as only 11 "Other" responses were given in Study 1b (in contrast to the 86 "Other" responses given in study 1a which had fewer participants). The coding of the other responses in this study followed the same guidelines as those described in Study 1a. Pair-wise Kappas were as follows: Kappa (95%CI), .66 (.27 to 1), .83 (.55 to 1), .63 (.17 to 1)[1]. The magnitude guidelines of Landis and Koch (1997) categorize all agreements as substantial. Only 1 response remained coded as other and all coders were in agreement as to its score.

Variability Scoring. Strategy composites and CWS ratios were calculated using the same methods described in Study 1a.

Results

The general data analysis plan was the same as that used in Study 1a. Boxplots of all variables are provided in Appendix C. Results were inconclusive regarding both relations between Strategy Composite and recall accuracy, r = 0.28, BF = 1.10, and CWS scores and recall accuracy, r = 0.10, BF = 0.35. When potential outliers were removed, the Bayes Factor shifted towards support for the null hypothesis, r = -0.04, BF = 0.32, see Figure 4.

[1] It should be noted that there were not enough "Other" responses in Study 1b or Study 1c to provide reliable Cohen's Kappa estimates (Cantor, 1996), however values were provided for comparison with other studies.

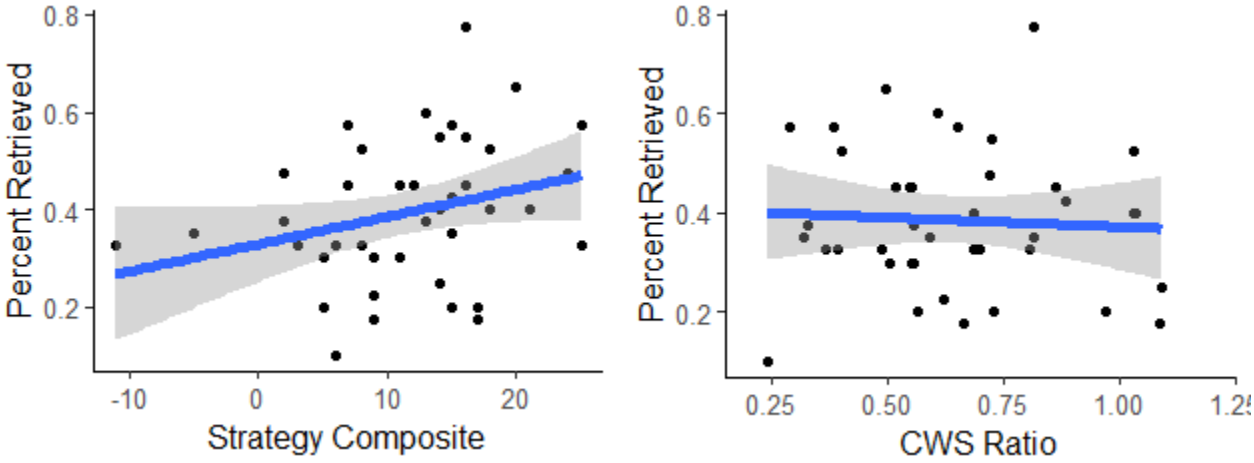


Figure 4. Scatterplots showing the relationship between strategy variability and recall accuracy with outliers removed, Study 1b.

Chapter 5: Study 1c

The goal of Study 1 was to develop a measure that could be used to capture both variability in *what* strategies participants use and variability in *how* they use their strategies. It was expected that participants who were more adaptable/flexible in their strategy use would be able to apply different strategies to different types of stimuli and improve their later memory retrieval, yet this pattern was not found in study 1a or 1b. In these studies we had hoped that participants would identify the fact that different categories of word pairs were being presented and use different strategies based on the type of pair. However, since participants were given no instructions to lead them to think there were differences between pairs it is possible they did not pick up on the word-pair categories until too far into the task. To try and alleviate this problem in Study 3c, participants completed two cued recall tasks, both with the 5 word-pair categories. It was expected that participants would learn during the first cued recall tasks which strategies were beneficial and which were not for specific categories, and then apply these strategies appropriately during the second task.

Participants & Materials

Seventy-Nine participants were recruited from the University of Maryland psychology subject pool. All participants received course credit for participation.

In addition to the cued recall task used in Study 1a and Study 1b, participants also completed an additional cued-recall task, along with a delayed free recall task and IQ task to see how strategy variability scores related to performance on other cognitive tasks.

Cued Recall Tasks. The additional cued recall task was identical to the task used in the two previous studies. Participants were presented with word pairs from 5 different categories for

6 seconds each, then presented with the first word of each pair and asked to retrieve the second. The method for constructing the new word pair list was the same as that describe in Study 1a.

Delayed Free Recall. Participants were presented with six, ten-word lists consisting of common nouns. All words were presented for 1 second each. The words were predetermined to be part of a specific list, but within each list the presentation order of the words was random. Immediately following the presentation of the 10^{th} word in a particular list, participants completed a distractor task where two sets of colored squares flashed on the screen and participants were asked to make a judgment whether the squares were the same or different. Participants repeated this colored squares matching task for 24 seconds. Following the distractor task participants had 45 seconds to retrieve as many of the words as they could in any order.

Ravens Advanced Progressive Matrices. Ravens advanced progressive matrices served as the IQ measure in this study. Participants were presented with up to 18 logic problems with increasing difficulty. Each problem presented a 3x3 matrix of complex geometric shapes displaying a pattern. The bottom right shape was always missing and participants had to select from 8 choices the shape they thought would complete the pattern. Participants were given 10 minutes to complete as many problems as they could.

Procedure

All participants completed the tasks in a set order: Cued Recall 1, Delayed Free Recall, Ravens, Cued Recall 2, and the Retrospective Strategy Assessment. As in Study 1b, thorough definitions of the strategy categories were provided before the assessment.

Retrospective Strategy Report Scoring

Coding of Other Responses. A total of 22 "Other" responses were given in this study. The coding of other responses in this study followed the same guidelines as those described in

Study 1a and Study 1b. Pair-wise Kappas were as follows: Kappa (95%CI), .64 (.39 to .89), .92 (.87 to 1), .70 (.46 to .94). The magnitude guidelines of Landis and Koch (1997) categorize all agreements as substantial. Only 1 response remained coded as other and all coders were in agreement as to its score.

Variability Scoring. Strategy composite and CWS ratios were calculated using the same methods described in Study 1a.

Results

The general data analysis plan was the same as that used in Study 1a and Study 1b. Boxplots of all variables are provided in Appendix C. There was strong support for a relation between Strategy Composite scores and memory performance on the second cued recall task, r = 0.56, BF = 136327.9. There was no support for a relation between CWS scores and recall accuracy on the second cued recall task, r = 0.06, BF = 0.265. When potential outliers were removed, the Bayes Factor shifted towards greater support for the null hypothesis, r = .16, BF = 0.25, see Figure 5.

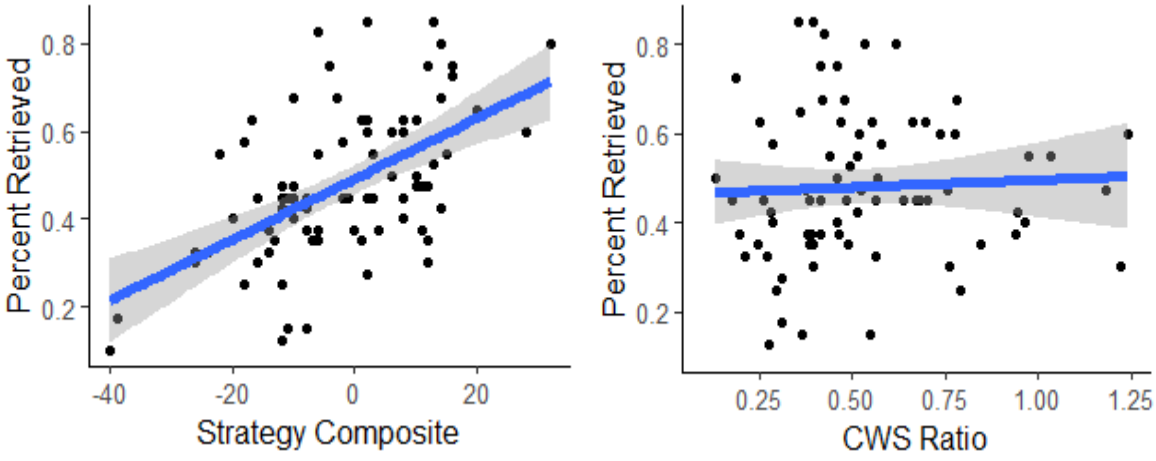


Figure 5. Scatterplots showing the relationship between strategy variability measures and recall accuracy with outliers removed, Study 1a.

Similar correlations were run between Strategy Composites and CWS score with performance on the first cued recall task, delayed free recall, and Ravens. There was again strong support for a relationship between Strategy Composite scores and cued recall accuracy, $r = 0.42$, $BF = 205.21$. All other comparisons had either support for no relationship or were inconclusive: Strategy Composite and DFR $r = 0.19$, $BF = 0.85$; Strategy Composite and Ravens $r = 0.03$, $BF = 0.24$, CWS and cued recall $r = 0.12$, $BF = 0.39$, CWS and DFR $r = -0.06$, $BF = 0.27$; CWS and Ravens $r = -0.03$, $BF = 0.24$.

Chapter 6: Study 1 Combined

In order to observe the relations pooled across all three studies, analyses were rerun on the combined cued recall accuracy (Study 1c Cued Recall 2), Strategy Composites, and CWS scores. Boxplots of the combined data can be found in Appendix C.

Results

Again there was strong support for a relationship between Strategy Composites and cued recall accuracy r = 0.35, BF = 1910.31. Data were inconclusive regarding a relationship between CWS scores and recall accuracy, r = 0.11, BF = 0.38, however with outliers removed support greatly shifted towards the null hypothesis, r = 0.004, BF = 0.18, see Figure 6.

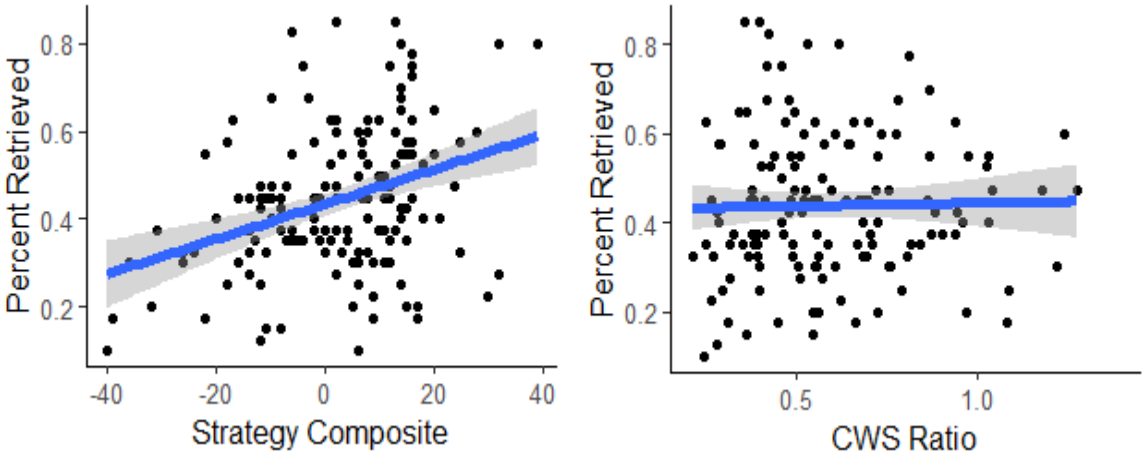


Figure 6. Scatterplots showing the relationship between strategy variability and recall accuracy with outliers removed, Study 1 Combined.

Chapter 7: Study 1 Discussion

The goal of Study 1 was to explore if a relationship could be found between two measures of strategy variability and cued-recall accuracy. The first measure assessed variability in what strategies learners used. Strategy composite scores were calculated by averaging strategy scores across strategies that are known to be more or less beneficial in improving memory performance. Across all three studies there was strong support that strategy composite scores could be used to predict memory retrieval, with learners who use a greater number of strategies known to be beneficial retrieving a greater number of items.

The second measure of strategy variability assessed variation in how participants used strategies. Modified CWS ratios were calculated for each participant to assess their variation in strategy use between word-pair categories relative to variation within a category. Across studies there was no support that CWS ratios could be used predict retrieval accuracy. In fact, when all three studies were combined there was very strong support for no relation between these two variables. There are several potential explanations for the lack of relation between these two variables which will be discussed in the general discussion. Because support was found for the relation between strategy composite scores and recall accuracy, the retrospective strategy report was included in Study 2 and both measures of variability were explored as individual differences in the testing effect.

Chapter 8: Study 2 – Individual Differences in Retrieval Practice

The goal of Study 2 was to examine if individual difference in student factors, specifically cognitive abilities and variability in strategy use, relate to the magnitude of the testing effect. The cognitive abilities examined included EM and gF and serve as a replication attempt of part of Brewer & Unsworth (2012) and Pan et al., (2015). Additionally, a new individual difference measure, variability in strategy use, is also examined. Variability in strategy use will be measured using the retrospective strategy report explored in Study 1.

Participants

Two-hundred four students were recruited from the University of Maryland undergraduate research pool and received course credit or $10 for participation. Four participants were excluded due to not completing the testing effect task. This resulted in 200 participants included in analyses. This sample size was the number decided prior to data collection to be necessary for SEM analysis (Kline, 2011).

Materials

Paired Associates Testing Effect Task. The measure of the testing effect was modeled after the task used by Brewer & Unsworth (2012) and Pan et al (2015) with three modifications. First the word pairs used in past studies consisted of common nouns. The word pairs used in the present study consisted of 5 categories: Related – High Imageability nouns, Related – Low Imageability nouns, Unrelated – High Imageability nouns, Unrelated – Low Imageability nouns, and Nonsense words. A complete list of word pairs is included in Appendix A. For more details on how the word pairs were generated, see Study 1.

During the encoding phase of the task, participants were presented with all 40 word pairs, one at a time, for 6 seconds each. During the study phase, participants restudied half of the

already encoded pairs for 6 seconds each and completed retrieval practice on the remaining words pairs. During retrieval practice, participants were presented with the first word of each pair (the cue) and tasked with retrieving the second word (the target). Participants were given 5 seconds to retrieve the target word and then presented with the correct word for 1 second in order to keep total time studying equivalent across the two conditions.

During Brewer & Unsworth's (2012) study phase, the restudied pairs were always presented before the retrieval pairs in order to reduce the possibility of a testing strategy being carried over onto the restudied items. In contrast, Pan and colleagues (2015) randomly mixed restudy and retrieval trials, which allowed for the possibility of carry over effects. Abel and Roediger (2017) however, have shown that the testing effect is equivalent across blocked and intermixed restudy. In the present study, participants were randomly assigned to complete either restudy first (n = 101) or practice retrieval first (n = 99). This manipulation allows to test for carry-over effect by examining if the magnitude of the testing effect differs based on which condition occurs first.

The final deviation from Brewer and Unsworth (2012) & Pan et al (2015) is the delay between study and final test. The prior two studies used 24 hour delays, however due to the feasibility and time frame of the present study, final test took place approximately 30 minutes after the study portion. Although this delay differs from the previous studies, empirical results have shown that delays longer than 5 mins show a testing effect whereas some studies with delays under 5 minutes show an advantage for restudying (Roediger & Karpicke, 2006; Roediger & Butler 2011). Additionally, Carpenter et al., (2008, study 1) compared delays of 5 mins, 1 day, 2 days, 1 week, 2 weeks, or 42 days, and found no interaction between the testing effect and retrieval interval and testing effects were present at all delays (also see Able & Roediger, 2017;

Avci et al., 2017). At final test participants completed self-paced cued recall of all 40 word pairs. Participants took, on average, 2.9 minutes to complete the retrieval portion of the test, or, 4.4 seconds per word pair. The dependent measures are the proportion of target words retrieved correctly at final test for both the restudied and retrieval practice pairs.

Episodic Memory Measures

Delayed Free Recall of Unrelated Words. Participants were presented with six, ten-word lists consisting of common nouns. All words were presented for 1 second each. The words were predetermined to be part of a specific list, but within each list the presentation order of the words was random. Immediately following the presentation of the 10^{th} word in a particular list, participants completed a distractor task where two sets of colored squares flashed on the screen and participants were asked to make a judgment whether the squares were the same or different. Participants repeated this colored squares matching task for 24 seconds. Following the distractor task participants had 45 seconds to retrieve as many of the words as they could in any order. A participant's score is the total number of words they were able to recall from all 6 lists.

Cued Recall. The cued recall task followed the same format as the testing effect task. Participants were presented, one at a time, with 40 word pairs from 5 different categories. Word pairs were presented for 6 seconds each. The pairs were grouped such that blocks of 4 pairs from the same word pair category were presented together. After the participants were presented with all 40 pairs, they immediately completed cued recall by being presented with first word and asked to retrieve the second. Retrieval was self-paced. Participants took on average 3.5 minutes to complete the retrieval portion of the task, or, 5.2 seconds per word pair. A participant's score is the proportion of words they were able to correctly recall.

Picture-Source Recognition. Participants were presented with 30 pictures in one of four corners of the computer screen. All pictures were presented for 1 second. Participants were specifically instructed to remember not only the pictures, but the locations in which they appeared. During retrieval, participants viewed the 30 old pictures and 30 new pictures and made judgements as to whether the picture was new, or, if old, which corner of the screen they had viewed it in. Participants were given 5 seconds to give their responses. Participants have two scores for this task, the proportion of old items they correctly identify in the correct quadrant and the proportion of new items correctly identified as new.

Gender-Source Recognition. Participants listened to 30 words which randomly varied as to whether they were read aloud by a male or female speaker. Participants were specifically instructed to remember not only the words, but the gender of the speaker. At retrieval 30 old words and 30 new words were presented on the computer screen and participants were asked to make judgment whether the word was new, said by a male speaker, or said by a female speaker. Participants were given 5 seconds to give their responses. Participants have two scores for this task, the proportion of old words they identified with the correct speaker and the proportion of new words they identified as new.

General Fluid Intelligence Tasks

Ravens Advanced Progressive Matrices. Participants were presented with up to 18 logic problems with increasing difficulty. Each problem presented a 3x3 matrix of complex geometric shapes displaying a pattern. The bottom right shape was always missing and participants had to select from 8 choices the shape they thought would complete the pattern. Participants were given 10 minutes to complete as many problems as they could. A participant's score is the total number of correct solutions.

Number Series (Thurstone & Thurstone, 1962). Participants were presented with a series of numbers that followed a specific sequence. Participants were tasked with determining the pattern of the sequence by selecting the appropriate next number from a set of 5 options. Participants completed 5 practice trials and were then given 4.5 minutes to complete up to 15 problems. A participant's score is the total number of correct solutions.

Letter Sets. Participants were presented with up to 20 problems, containing 5 sets of 4 letters each. In each problem 4 of the 5 sets followed a rule and one did not. Participants were tasked with identifying the letter set that did not follow the same rule as the other sets. Participants completed 4 practice problem and were then given 5 minutes to complete up to 20 problems. A participant's score is the total number of correct solutions.

Retrospective Strategy Report. Participants were presented with all of the word pairs from the Paired Associated Testing Effect Task an additional time to collect self-reported strategy use data. Presentation of all word pairs from all categories was randomized. Participants made a forced choice decision regarding which of 5 memory strategies they felt they used most *when originally learning the word pair*. Available strategies options were: Visualization, Sentence Generation, Rote Rehearsal, Other, and None (Dunlosky & Hertzog, 2001). A definition of each strategy was provided to the participants and can be found in the methods section of Study 1b. If participants selected "Other", they were prompted to type out an explanation of what strategy they felt they used. Responses on this task were used to calculate strategy composite scores and CWS ratios for each participant.

All "Other" responses were independently coded by three coders (myself and two trained research assistants) to determine if they really were "Other" strategies or if they belonged in one of the four provided categories. A total of 53 "Other" responses were given in this study. The

coding of "Other" responses in this study followed the same guidelines as those described in Study 1. Pair-wise Kappas were as follows: Kappa (95%CI), .68 (.53 to .83), .60 (.44 to .76), .68 (.51 to .84). The magnitude guidelines of Landis and Koch (1997) categorize all agreements as substantial.

Strategy composite scores were calculated for each participant by taking the average score of all strategies they reported using. Visualization and Sentence Generation responses were scored as a positive one (+1) and Rote Rehearsal and None responses were scored as a negative one (-1). After initial coding, 3 responses remained coded as "Other". These responses were again independently coded and all coders were in agreement as to their scores.

Additionally, strategy variability scores were determined by calculating CWS ratios for each participant (see Study 1). Measures of "unalikeability" in strategy use (i.e., variability of a categorical variable, Kader & Perry, 2007) were calculated within word pair categories for each participant using the equation below,

$$u_2 = 1 - \sum_i p_i^2 \quad (2)$$

where p_i represents the proportion of a specific option within category i. Variation in the proportion of strategies used between categories was calculated using a sample measure of standard deviation. Using the variability between as the measure of discriminability and the variability within as the measure of consistency, CWS ratios were calculated for each participant using the equation below.

$$CWS = \frac{Discriminability}{Consistency} \quad (1)$$

Procedure

The study took place in one session lasting approximately an hour and a half. Participants first completed 3 of the EM measures to make sure that strategies learned during the testing

effect task would not carry over. Next participants completed the encoding and study phases of the paired associated testing effect task. Participants were randomly assigned to complete restudy or retrieval practice first. The delayed free recall episodic memory task occurred after the first portion of the testing effect task to ensure the delay between the study and final tests phases would be at least thirty minutes. Participants then completed the gF measures followed by the remainder of the testing effect task. This allowed for an approximately 30 minute delay between the study and final test portions of the testing effect task. Finally, participants completed the retrospective strategy report. This order was set for all participants.

General Analysis Plan.

All analyses were run using Bayesian model comparison, with proportion data transformed using the logit transformation. This approach provides an index of the degree of support for any model of interest (the Bayes Factor, BF), including the null model. BFs provide the evidence of support for alternative or more complex models relative to a null or reduced model, such that BFs greater than 1 represent support for the more complex model and BFs less than one represent support for the reduced model. Additionally, BFs represent a continuum with values further from 1 representing greater support. For more explanation on BF interpretation see Study 1.

The exact models run will be discussed in more detail in the results section, however a few common elements will be discussed here. All individual differences will be tested by comparing a model with an interaction with Study Type to a model with only main effects. This approach is different than the analysis used by Brewer & Unsworth (2012) and Pan et al (2015),

however I believe it allows for a greater understanding of potential individual differences[2]. Previous research has tested for individual differences by predicting differences in memory performance from a single continuous variable. Although that approach would reveal individual differences, it has two shortcomings. First, difference scores are known to be unreliable. Second, it does not provide insight into whether differences are due to difference in performance on restudy trials, retrieval trials, or both. Instead, by testing for an interaction and plotting this effect, conclusions can be made as to which factor is driving in the interaction. All analyses provided in the manuscript include interactions, however in order to compare with previous work, analyses using differences scores are provided in Appendix D.

Results

Descriptive statistics and a correlation matrix for all measured variables can be found in Table 2. Additionally, boxplots displaying the distributions for all variables can be found in Appendix C.

Question 1: Is final test performance higher for items that received retrieval practice compared to items that were restudied (i.e., is there a testing effect)?

Memory performance for items that received practice retrieval and items that were restudied for both participants that competed restudy first and participants who completed retrieval first are presented in Figure 7. Analyses revealed decisive support for a testing effect ($BF_{10} = 2.28 \times 10^{11}$, $\omega^2 = 0.25$), with participants remembering on average 7.5% more of the items that received retrieval practice compared to restudy. Breaking down the results further, in

[2] The data from Brewer & Unsworth(2012), were re-run using the Bayesian model comparison with interaction approach used in the present paper. There was still support that both EM and gF served as individual differences $BFs_{10} = 4.79, 6.91$.

this particular sample, 66% of participants showed a testing effect, 15% showed no effect, and 19% of participants performed better on items that were restudied.

Question 2: Does the order participants complete the portions of the study phase impact the magnitude of the testing effect?

Bayesian model comparison was run to determine whether the order participants complete the study phases (i.e., restudy first or practice retrieval first), influenced the magnitude of the testing effect.

Table 2. Correlation matrix and descriptives of all measured variables

	RES	RET	TE	Scomp	CWS	CR	DFR	PS-SC	PS-CR	GS-SC	GS-CD	Ravens	Letter	Num	Mean	St.Dev.	n	Rel
RES	-														0.45	0.18	200	0.81
RET	0.74	-													0.53	0.17	200	0.81
TE	-0.56	0.15	-												0.08	0.12	200	0.68
Scomp	0.38	0.33	-0.15	-											-0.82	11.76	200	0.76
CWS	-0.15	0.03	-0.19	0.17	-										0.61	0.27	200	0.32
CR	0.5	0.52	-0.09	0.29	0.01	-	-								0.42	0.15	200	0.77
DFR	0.4	0.42	-0.07	0.26	0.08	0.51	-								29.58	7.8	200	0.73
PS-SC	0.48	0.47	-0.13	0.31	0.08	0.42	0.27	-							0.75	0.18	200	0.93
PS-CR	0.39	0.38	-0.11	0.21	0.04	0.29	0.31	0.56	-						0.86	0.21	200	0.73
GS-SC	0.19	0.22	0	0.23	0.02	0.32	0.32	0.23	0.17	-					0.56	0.15	199	0.86

			-														
	0.37	0.37	0.08	0.27	0.04	0.42	0.45	0.46	0.46	-				0.66	0.22	199	0.63
			-														
Ravens	0.35	0.26	0.19	0.17	0.11	0.29	0.32	0.24	0.31	0.18	0.26	-		10.45	4.19	200	0.94
			-														
Letter	0.14	0.15	0.02	0.2	0.05	0.16	0.29	0.14	0.14	0.07	0.22	0.24	-	10.23	9.82	200	0.84
	0.14	0.17	0.01			0.28	0.26	0.12	0.15	0.06	0.15					200	0.85
Num				0.13	0.02							0.24	0.48	- 3.12	2.73		

Note. Variable names are as follows: CR = Cued Recall, DFR = Delayed Free Recall, PS-SC = Picture Source Source Correct, PS-CR = Picture Source Correct Rejection, GS-SC = Gender Source Source Correct, GS-CR = Gender Source Correct Rejection, Restudy = Proportion of items from the Restudy condition that were correctly retrieved at final recall, Retrieval = Proportion of items from the Retrieval Practice condition that were correctly retrieved at final recall, TE = Testing Effect measured as the difference between the proportion of Retrieval Practice and Restudy items correctly retrieved at final recall. Reliabilities were calculated using split-half reliability.

*Final Test Accuracy ~ Study Type + Study Order + Study Type*Study Order*

There was strong support for no interaction between Study Type and Study Order (BF_{10} = 0.18, ω^2 = 0.007), suggesting that the magnitude of the testing effect did not differ between participants who completed the restudy trials first and the participants who completed the retrieval practice trials first. Results were inconclusive regarding whether there was a main effect of Study Order (BF_{10} = 0.51, ω^2 = 0.01)[3]. Additionally, when controlling for order there was still a main effect of study type (BF_{10} = 2.50 x 10^{11}, ω^2 = 0.25).

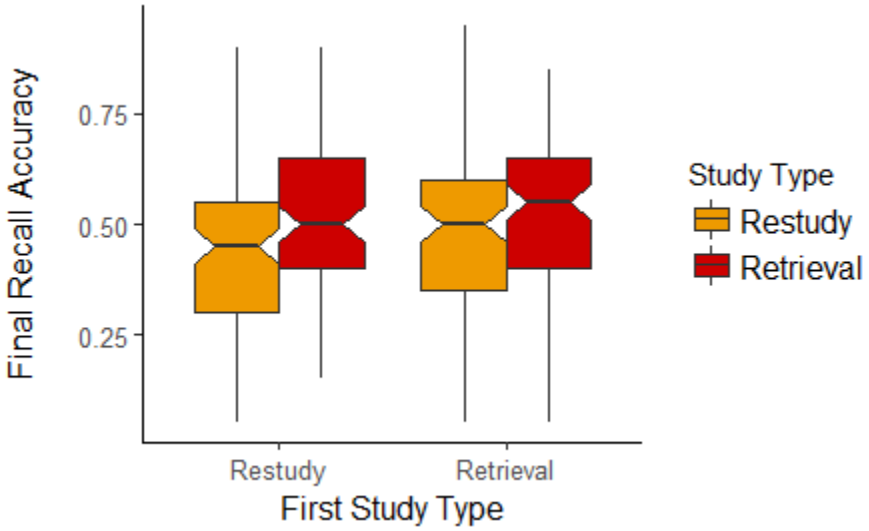


Figure 7. Boxplots showing final recall accuracy for items that were restudied and items that received practice retrieval.

Question 3: Does each student factor individually relate to the magnitude of the testing effect?

[3] Because results regarding the effect of order were inconclusive all subsequent analyses were run both with and without Study Order as an additional factor. All results involving Study Order however were either support for the null or inconclusive therefore they are not included in the manuscript. Additionally, including Study Order as a covariate did not change the support for any other effect. All results report are collapsed across Study Order.

Bayesian model comparisons were run for each student factor separately to determine its effect on the magnitude of the testing effect. The four student factors tested included composite episodic memory, composite fluid intelligence, strategy composite scores, and CWS ratios. Composite measures for episodic memory and fluid intelligence were calculated by z-scoring each measure and taking the average z-score for each ability.

Episodic Memory. The following model was tested.

*Final Test Accuracy ~ Study Type + Episodic Memory + Study Type*Episodic Memory*

There was no support for an interaction between Study Type and EM ($BF_{10} = 0.23$, $\omega^2 = 0.01$) suggesting that EM does not serve as an individual difference measure in explaining the magnitude of the testing effect (See Figure 8). There was however support for a main effect of EM in predicting final test accuracy ($BF_{10} = 3.98 \times 10^{18}$, $\omega^2 = 0.37$) and the testing effect was still present after controlling for the effects of EM ($BF_{10} = 2.03 \times 10^{11}$, $\omega^2 = 0.26$).

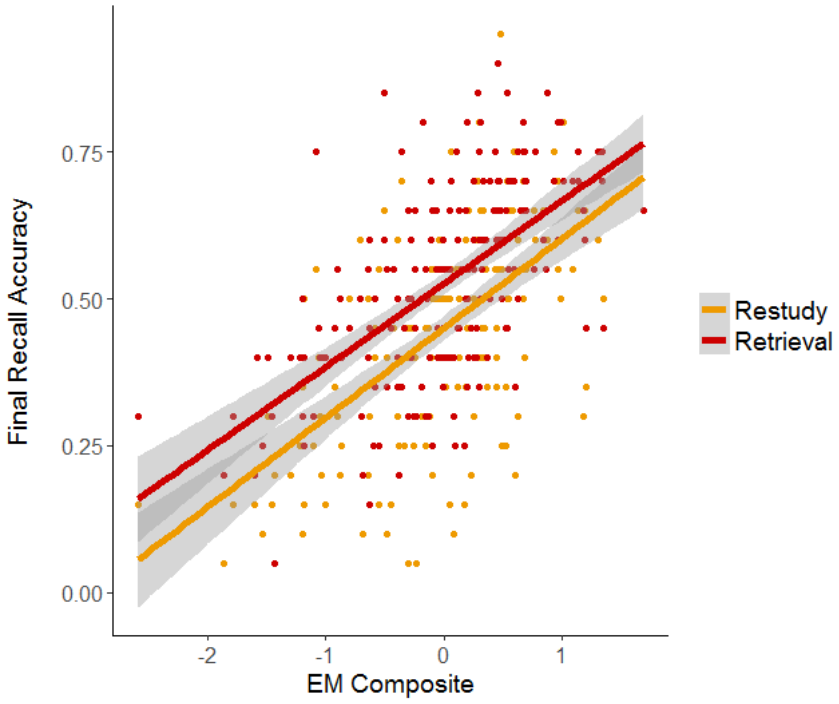


Figure 8. The relationship between episodic memory and final recall accuracy for items receiving restudy and retrieval practice.

Fluid Intelligence. The following model was tested.

*Final Test Accuracy ~ Study Type + Intelligence + Study Type*Intelligence*

There was no support for an interaction between Intelligence and Study Type (BF_{10} = 0.21, ω^2 = 0.01) suggesting that IQ does not serve as an individual difference in predicting the magnitude of the testing effect (see Figure 9). There was however a main effect of IQ (BF_{10} =810.13, ω^2 = 0.09) and the testing effect remained after controlling for the effects of IQ (BF_{10} = 2.29 x 10^{11}, ω^2 = 0.25).

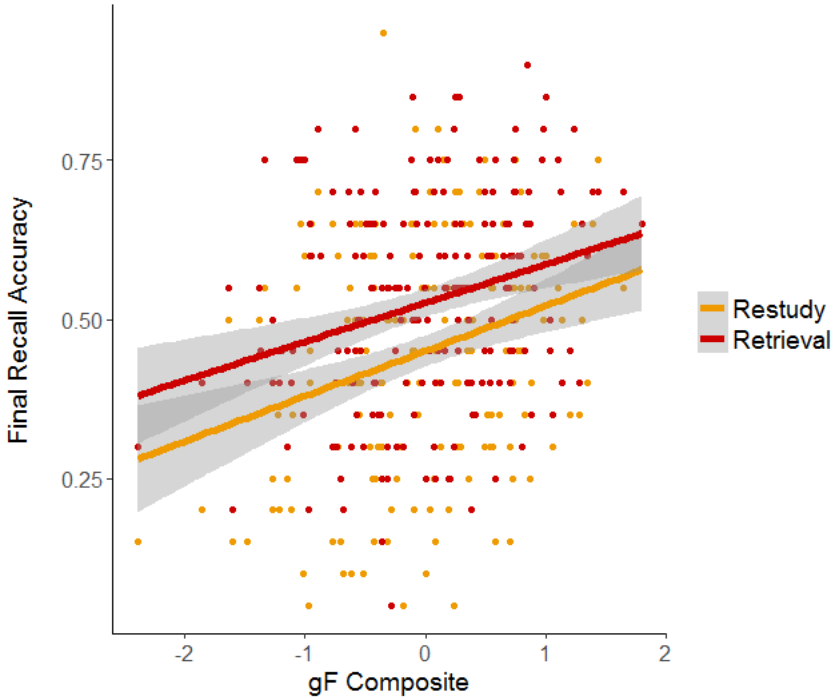


Figure 9. The relationship between general fluid intelligence and final recall accuracy for items receiving restudy and retrieval practice.

Strategy Variability. To test the effects of *what* strategies are used the following model was tested:

*Final Test Accuracy ~ Study Type + Strategy Composite + Study Type*Strategy Composite*

Results were inconclusive regarding whether an interaction existed between Strategy Composite scores and Study Type ($BF_{10} = 0.89$, $\omega^2 = 0.02$, see Figure 10). To ensure these results were not being driven by outliers in Strategy Composite Scores, these points were removed (n = 2) and analyses were re-run. Results shifted in favor of the null, but were still inconclusive ($BF_{10} = 0.44$, $\omega^2 = 0.02$). With both analyses there was a main effect of Strategy Composite ($BFs_{10} = 401769, 42018.42$, $\omega^2 = 0.14, 0.12$), and the testing effect remained when controlling for strategy composite ($BFs_{10} = 2.13 \times 10^{11}, 1.16 \times 10^{11}$, $\omega^2 = 0.26, 0.25$).

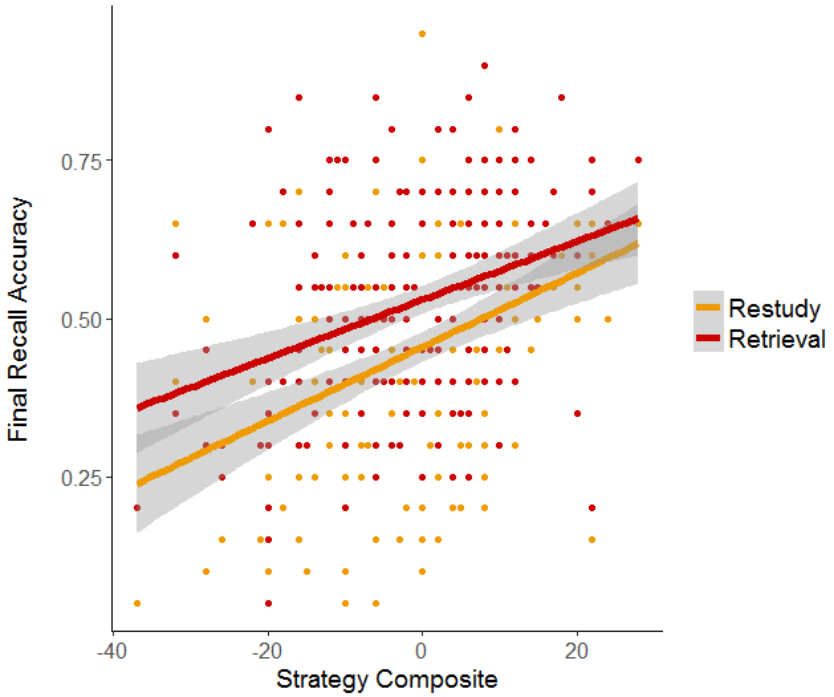


Figure 10. The relationship between strategy composite scores and final recall accuracy for items receiving restudy and retrieval practice

To test the effects of *how* strategies are used the following model was tested:

*Final Test Accuracy ~ Study Type + CWS + Study Type*CWS*

Results showed support for an interaction between CWS ratio and Study Type (BF_{10} = 5.5, ω^2 = 0.04) suggesting that variability in how learners use strategies does serve as an individual difference that can predict the magnitude of the testing effect. Additionally, as can be seen in Figure 11, performance on items that experienced retrieval practice is equal across all levels of strategy variability whereas performance in the restudied items increases as strategy variability scores increased. Based on this pattern of results, retrieval practice serves as a mechanism to equate performance across learners with differences in strategy use, whereas when learners are free to restudy on their own, more skilled strategy users perform better because they likely make better use of the restudy opportunity. To confirm that these effects were not driven by outliers in

CWS ratios, these data points were removed (n = 8) and the analysis was rerun. With the outliers removed the support for an interaction was even greater (BF_{10} = 28.57, ω^2 = 0.06). Now memory performance appears to increase as strategy variability increases for both restudied and retrieved items, however the increase is greater for restudied items (See Figure 11). With outliers included, it was inconclusive whether there was a main effect of CWS (BF_{10} = 0.58, ω^2 = 0.01), however with outliers removed there was strong support (BF_{10} = 66.29, ω^2 = 0.06). In both cases the testing effect still showed strong support after controlling for the effects of CWS (BFs_{10} = 2.4 x 10^{11}, 1.39 x 10^{11}, ω^2 = 0.26, 0.27).

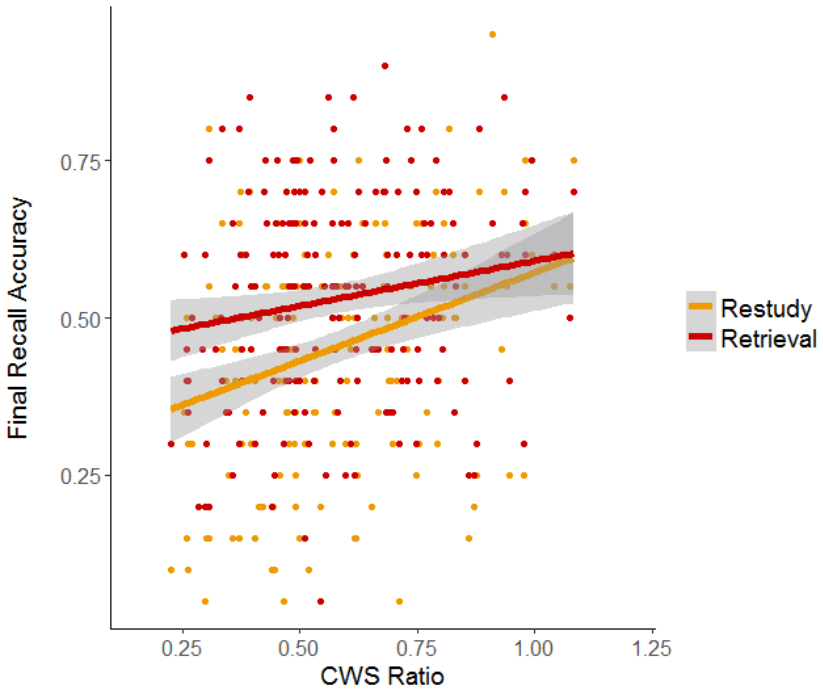


Figure 11. The relationship between CWS ratios and final recall accuracy for items receiving restudy and retrieval practice without outliers.

Question 4. Which combination of student factors best explains the magnitude of the testing effect?

Bayesian all subsets regressions were run to determine which set of cognitive abilities best predicts the magnitude of the testing effect. For this analysis, the testing effect was represented by the difference in final recall accuracy between items that were restudied and items that received retrieval practice. The following independent variables were included: EM composite, gF composite, Strategy Composite, CWS ratio, and Order of Restudy/Retrieval Practice. The best model included only the CWS ratio (BF_{10} = 5.02), however this model was not substantially different than a model including both the CWS ratio and Strategy Composite (BF_{10} = 0.87). To ensure these results were not driven by outliers, all subsets were rerun with outliers removed. Outliers were defined as data points greater than 1.5 times the interquartile range of any variable (n = 9). Again the best model included only the CWS ratio (BF_{10} = 19.27), and this model was substantially more supported than the second best model which included the CWS ratio and Strategy Composite (BF_{10} = 0.32).

To provide converging evidence for the model found above, two additional analyses were run, structural equation modeling and dominance analysis. One benefit of structural equation modeling is it eliminates the measurement error associated with composite scores allowing for a cleaner observation of the relationship. Again, all 4 cognitive abilities were used to predict the magnitude of the testing effect (difference in accuracy between restudy and practice retrieval items), but with EM and gF represented as latent factors rather than composite scores. The full model is presented in Figure 12.

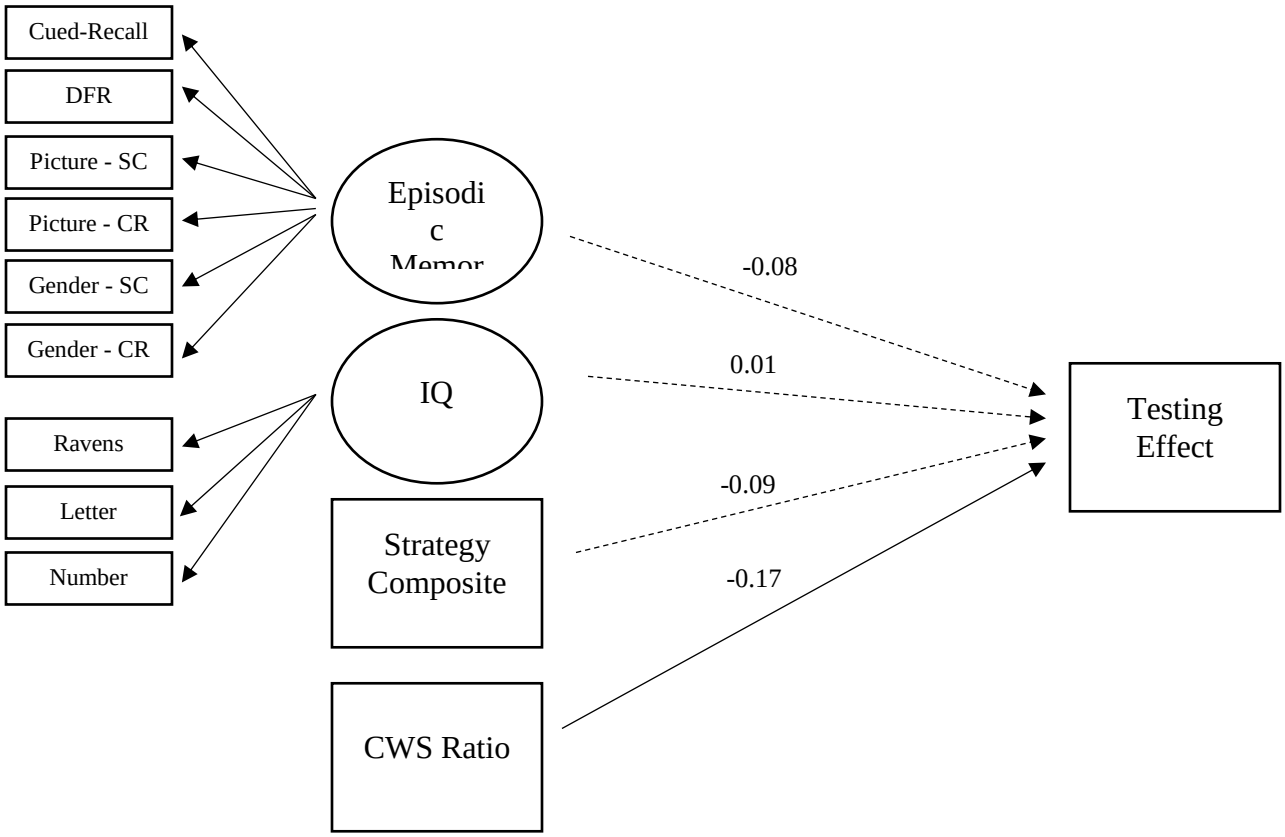


Figure 12. Structural SEM Model standardized results with outliers included. Solid lines represent significant paths.

The data were analyzed using a two-step SEM approach with the Satorra-Bentler Scaling method to adjust for non-normality (Satorra & Bentler, 2010). All variables were standardized. In Step 1, the initial measurement model was tested to confirm that the measured variables load appropriately onto the cognitive abilities (i.e., EM and gF). In this step all factors were permitted to co-vary. Additionally, the error terms from the two picture source recognition and gender source recognition tasks were allowed to co-vary. Overall the measurement model had good fit based on the guidelines proposed by Hu & Bentler (1995), $\chi^2(44) = 69.07$, $p = .01$, AIC = 8839.10, RMSEA = .05 (95% CI = 0.03 to 0.08), CFI = 0.94, SRMR = 0.05. No modifications were made to this model.

In Step 2, the structural portions of the model were added to the existing measurement model. Because all variables without directional paths were allowed to co-vary, the structural model fit was equal to the measurement model. The only significant path to the testing effect was the CWS ratio, b = -0.17, $t(198) = 2.61$, p = 0.01. The paths from the other cognitive abilities to the testing effect were not significant, EM: b = -0.08, $t(198) = 0.68$, p = 0.50, IQ: b = 0.01, $t(198) = 0.07$, p = 0.95, Strategy Composite: b = -0.09, $t(198) = 1.14$, p = 0.26.

To ensure these findings were not driven by outliers, both steps of the SEM approach were again run with outliers removed. Outliers were defined using the same metric as during the all subsets regressions. Again, all fit indices were acceptable, $\chi^2(44) = 65.24$, p = .02, AIC = 8334.98, RMSEA = .05 (95% CI = 0.02 to 0.08), CFI = 0.95, SRMR = 0.06. Again, the only significant path to the testing effect was from the CWS ratio b = -0.20, $t(190) = 3.10$, p = 0.002. None of the other paths were significant, EM: b = -0.07, $t(190) = 0.60$, p = 0.55, IQ: b = 0.03, $t(190) = 0.25$, p = 0.80, Strategy Composite: b = -0.07, $t(190) = 0.87$, p = 0.39.

The goal of the all subsets regression was to determine which set of cognitive abilities best explain the magnitude of the testing effect. Another approach to answering this question is to use dominance analysis (Budescu, 1993). Dominance analysis provides a qualitative assessment of the relative importance of all predictors in a multiple regression. Each predictor is given a dominance score indicating its relative importance compared to the other predictors. The dominance scores sum to the R^2 of the model with all predictors included. One of the biggest strengths of this methods is it avoids issues

relating to multicollinearity. Dominance analysis was run predicting the magnitude of the testing effect from the four cognitive abilities and study order.

Like with other analyses, the dominance analysis was run once with outliers and once without. Outliers were defined the same way as previous analyses. The results of the dominance analysis can be seen in Table 3. Consistent with findings from above, both with and without outliers, the CWS ratio is clearly the dominating predictor. Additionally, with both models, Strategy Composite is the second most important predictor, but to a much lesser extent than the CWS ratio. EM, IQ, and Order play almost no role in predicting the magnitude of the testing effect.

Table 3. Results of dominance analysis both with and without outliers.

Data	EM	IQ	Strategy Composite	CWS	Order	R^2
All Data	0.007	0.004	0.014	0.032	0.001	0.057
No Outliers	0.005	0.002	0.007	0.044	0.002	0.061

Additional exploratory analyses can be found in Appendix E.

Study 2 Discussion

The testing effect is a robust finding, but across samples, some learners benefit and some do not. The goal of Study 2 was to explore if any cognitive abilities of the learner could predict the magnitude of the testing effect or if differences in the benefit were due to random chance. Across multiple methods of analysis, CWS ratios, a measure of *how* participants use strategies, were able to explain differences in the benefit or retrieval practice. More specifically, learners with low CWS ratios (i.e., learners who were not flexible/adaptable in their strategy use) benefited more from retrieval practice than learners with high CWS ratios. In fact, learners with high CWS ratios showed

essentially no difference in final recall accuracy between items that were restudied and items that received practice retrieval.

Findings regarding strategy composite scores, a measure of *what* strategies learners use, were largely inconclusive. Results of the dominance analysis suggest that strategy composite scores were the second most important factor, however the Bayes factors were inconclusive regarding whether an interaction existed between study type and strategy composite scores. Additionally, there was not substantial support that strategy composite scores improved the predictability of the testing effect beyond a model with just CWS ratios. The support for strategy composite scores became even weaker when outliers were removed.

Across all methods of analysis there was no support for either EM or IQ predicting the magnitude of the testing effect. Additionally, in methods using Bayesian model comparison there was consistently support for the null hypothesis. Past research has been inconsistent as to whether EM serves as an individual difference in predicting the testing effect. Specifically, Brewer & Unsworth (2012) found a significant negative relationship where learners with lower EM showed a greater testing effect, whereas Pan et al (2015) found no relation for either of their studies. One problem with attempting to draw conclusions from a single study, however, is that small effect sizes can be hard to detect, especially without large sample sizes. The current study was intentionally designed so that data could be pooled across the four studies, as all of the same measures were used, with only minor methodological differences. With all four studies combined (N = 547), there is modest support for an interaction between EM and study type in predicting memory accuracy ($BF_{10} = 5.19$), however this effect is quite small ($\omega^2 = 0.02$).

When using EM scores to predict the difference between restudy accuracy and retrieval accuracy, EM only explains 2% of the variance in the testing effect. Although the data do provide support for the hypothesis that EM can predict the magnitude of the testing effect, the relation is extremely small and may not be relevant in applied settings.

In addition to exploring individual differences in the testing effect, this study also examined how the order of restudy and retrieval practice trials influences the magnitude of the testing effect. Previous work has shown that the testing effect is equivalent across blocked (restudy first) and intermixed study trials (Able & Roediger, 2017), however, the design of the present study was to specifically test the order blocked trials with half of the participants completing restudy first and half completing retrieval practice first. It was expected that completing retrieval practice first may lead to a carry-over of the effect into the restudy trials reducing the testing effect, however this interaction was not found. These results combined with previous studies suggest that the benefits of retrieval practice are resistant to changes in when the retrieval practice trials occur relative to restudy.

Chapter 9: General Discussion

In Study 1 a new measure was tested to examine the how variation in strategy use, both *what* strategies are used and *how* strategies are used relate to overall memory performance. Across a series of three studies, strong relations were found between Strategy Composite scores (i.e., *what* strategies are used) and memory recall accuracy, but no such relations were found for CWS ratios (i.e., *how* strategies are used). In Study 2, variation in strategy use along with EM and IQ were explored as potential individual differences in the testing effect. Across a variety of analyses, CWS ratios were the only cognitive ability to consistently relate to the magnitude of the testing effect.

It may seem surprising that in Study 1 the CWS ratio was not related to memory retrieval, whereas in Study 2 the CWS ratio was able to predict the magnitude of the testing effect. More specifically, for items that received retrieval practice there was minimal to no relation between CWS ratios and recall accuracy, whereas for items that were restudied there was a strong positive relation between CWS ratios and recall accuracy. This may be contradictory as the restudy trials are, on the surface, the trials more similar to the design used in Study 1. Upon closer examination, however, there are 2 key differences between the cued recall task and retrospective strategy judgements collected in Study 1 and Study 2.

The first difference in the two designs is the delay between initial encoding and retrospective strategy judgments. For all three studies in Study 1, encoding was immediately followed by retrieval and strategies judgements, whereas in Study 2 there was a 30 minute delay between initial encoding and final retrieval. Retrospective strategy judgments are themselves somewhat of a memory task as the initial strategy used must be

retrievable at the time of judgment. One of the weaknesses of retrospective strategy reports as compared to concurrent strategy reports is strategies are forgotten over time, and therefore the retrospective reports are less accurate (Dunlosky & Hertzog, 2001). Conversely, retrospective reports may be less accurate regarding what strategies participants used when originally learning the pairs, but they may in fact be more accurate at capturing strategies that were beneficial for improving long-term retention, particularly when measured after a delay. For example, in Study 1, because the delay was so short, strategies that were initiated, but not fully successful (i.e., would not have benefited long term retention) may have still been available at the point of strategy judgment. In contrast, with the delay present in Study 2, it is unlikely that strategies there were unhelpful at retrieval were remembered when providing strategy judgments. This would result in the CWS ratios found after a delay (i.e., Study 2) to be a better representation of variation in beneficial strategy use.

The second major difference between Study 1 and Study 2, particularly for the items that were restudied, is the number of study opportunities with each item. For all studies under Study 1 and the retrieval practice items in Study 2, participants were given a single 6 second learning opportunity with each word pair. In contrast, for the items that were restudied, participants received two 6 second study opportunities. For all items that received just 1 study opportunity no relation was found with the CWS ratio, whereas for the items that received two study opportunities a strong relationship was found. In Study 1c, an additional categorized cued-recall task was included with the hope that participants would learn from the first task that (1) there were different word pair categories included, and (2) some strategies were more or less beneficial for certain categories compared to

others. It was expected that this knowledge from the first task would carry-over to the second task and be reflected in the CWS ratio, however pattern was not found. The results of Study 2 suggest that an additional categorized recall task is not enough to cause learners to use strategies flexibly between categories, and instead, familiarization with the words of that specific task may be needed. In this scenario it is likely that the retrospective strategy report was capturing a combination of both the strategies used during the restudy opportunity and those used during initial learning. Additionally, the CWS ratio did not correlate with any other cognitive measures collected during Study 2, suggesting again that the measure is extremely task specific and not generalizable to strategy flexibility beyond the measured stimuli.

Implications for theories of the testing effect. While several theories of the underlying mechanisms of the testing effect have been proposed, there is still much work to be done, as no existing theories account for potential individual difference. The findings of the present study support multiple existing theories, however they do not lend support for one existing theory over another.

First, the testing effect in the present study was smaller than in many past studies. This is likely due to the specific stimuli that were used. In most past studies involving word pairs, unrelated pairs were used, however in the present study both related and unrelated pairs were used. The related word pairs were more easily remembered minimizing the benefits of testing, whereas the unrelated word pairs were hard to retrieve and showed a typical larger testing effect. This finding is consistent with the retrieval effort hypothesis, suggesting that harder to retrieve information experiences greater

elaboration during retrieval practice and therefore shows a greater benefit at final retrieval (Carpenter & DeLosh, 2006; Pyc & Rawson, 2009).

Additionally, the finding that learners who are better able to flexibly use strategies during learning show a smaller testing effect is consistent with both the Elaborative Hypothesis (Carpenter & DeLosh, 2006) and the Episodic Context Account (Karpicke, Lehman, & Aue, 2014). Under the Elaborative Retrieval Hypothesis, retrieval practice benefits memory performance by strengthening or creating additional memory traces that can be used at later retrieval. Learners who are able to apply strategies flexibly to the to-be-remembered material are able to strengthen the memory traces more during restudy than the learners with poor strategy use, making the elaboration that occurs with retrieval practice less beneficial. It should be noted however, that under this hypothesis, it is surprising that greater support for individual differences based on strategy composite scores were not found.

Similarly, the same pattern of results would have been expected under the Episodic Context Account which proposes that retrieval practice improves memory performance through the reinstatement of contextual information associated with the original learning episode. One context that could be reinstated during retrieval is the semantic context, which would have similar effects to using a strategy that focused on semantic information such as “Verbal Association”. Under this hypothesis, learners who make use of this contextual information during restudy (i.e., using semantic-related strategies when semantic information is available) would show less benefit of semantic reinstatement during retrieval practice. One limitation of the present study however is that strategies involving the use of other pieces of contextual information (such as

temporal information) were not probed and including such strategies may lead to larger individual differences. In summary, although the results of the present study to not help identify the specific mechanisms responsible for the testing effect, they do support the existing theories.

Real World Applications. In the last decade, the testing effect has begun to be incorporated into the classroom at almost all levels of education. Retrieval practice has been shown to improve student outcomes in college level statistics courses and psychology courses (Batsell, Perry, Hanley, Hostetter, 2017; Lyle & Crawford, 2011), middle school history and science classes (Carpenter, et al., 2009; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011), elementary school classes (Karpicke, Blunt, Smith, & Karpicke, 2014), and even medical school courses (Larsen, Butler, & Roediger, 2009). While all of these studies found group level difference, none examined the benefits of retrieval practice at the individual student level. An important aspect for educators to know is which students in their classroom are expected to benefit from retrieval practice and just as important, if any students would be at a disadvantage. The findings of the present study would support the use of retrieval practice in educational setting. Specifically, the testing effect appears to benefit students equally regardless of their episodic memory skills, general fluid intelligence, or *what* strategies they use. Additionally there is evidence that students who are less adaptable in *how* they use strategies will show a greater testing effects suggesting that lower preforming students will show a greater benefit. Furthermore, students who are good at adaptively using strategies do not appear to be disadvantaged by retrieval practice.

The present findings suggest that the testing effect is larger for learners who are less adaptive in their strategy use, however the benefits appear to be equal across other cognitive factors. It is important to note however, that cognitive factors are just one domain of student factors that could influence the benefits of retrieval practice. Some work has begun to look at academic factors such as prior knowledge and general class performance, however personality related factors such as belief in the benefits of testing still remain unexplored. Research exploring the benefits of retrieval practice continually shows support for its use in applied settings, however much work is still needed to understand the nuances of its effects.

Appendix A

Cue Word	Target Word	Category
Attempt	Try	Related - Low Imageability
Cause	Effect	Related - Low Imageability
Matter	Fact	Related - Low Imageability
Motive	Reason	Related - Low Imageability
Stay	Put	Related - Low Imageability
Promise	Keep	Related - Low Imageability
Theory	Idea	Related - Low Imageability
System	Method	Related - Low Imageability
Extent	Culture	Unrelated - Low Imageability
Common	Former	Unrelated - Low Imageability
Way	Moral	Unrelated - Low Imageability
Know	Ego	Unrelated - Low Imageability
Manner	Issue	Unrelated - Low Imageability
Unit	Real	Unrelated - Low Imageability
Factor	Ease	Unrelated - Low Imageability
Result	Event	Unrelated - Low Imageability
Ape	Man	Related - High Imageability
Table	Chair	Related - High Imageability
Bat	Ball	Related - High Imageability
Teeth	Dentist	Related - High Imageability
Cash	Money	Related - High Imageability
Tractor	Trailer	Related - High Imageability
Honey	Bee	Related - High Imageability
School	Boy	Related - High Imageability
Piano	Rubber	Unrelated - High Imageability
Hotel	Car	Unrelated - High Imageability
House	Island	Unrelated - High Imageability
Feet	Paper	Unrelated - High Imageability
Road	Heart	Unrelated - High Imageability
Farm	Bone	Unrelated - High Imageability
Army	Queen	Unrelated - High Imageability
Fire	Hall	Unrelated - High Imageability
Gluce	Sprogs	Nonsense
Yipe	Tren	Nonsense
Ock	Chalds	Nonsense
Plince	Tranch	Nonsense
Snurf	Shroon	Nonsense
Flince	Stilch	Nonsense
Bliff	Yold	Nonsense
Vab	Gwerp	Nonsense

Appendix B.

The below examples represent manually simulated data that demonstrate the CWS ratio (Shanteau et al. 2002). The first example (A) represents a participant with poor strategy use. All decisions were random resulting in an approximately even distribution of strategies across categories. The second example (B) represents a participant with good strategy use. Only two types of the strategies were used for each category and different strategies were used for different categories. The third example (C) represents good strategy use plus noise. Ranking these strategy users from best to worst would results in Participant B, then C, then A. As can be seen below, the CWS ratios follow the same order.

For all tables, the word pair categories are as follow: RH – Related High Imageability, RL – Related Low Imageability, UH – Unrelated High Imageability, UL – Unrelated Low Imageability, N – Nonsense Words.

Variability within was calculated using the following formula:

$$u_2 = 1 - \sum_i p_i^2$$

Where p_i represents the proportion of a specific option within category i.

CWS ratios were then calculated using the following formula:

$$CWS\ Ratio = \frac{Discrimination}{Consistency}$$

Where *Consistency* is represented by within category variability and *Discrimination* is represented by between category variability.

Participant A.

Strategy responses for a participant with poor strategy use

	Word Pair Category				
Strategy	RH	RL	UH	UL	N
Visualization	3	2	2	1	3
Sentence Generation	2	3	1	2	1
Rote rehearsal	2	1	2	3	2
None	1	2	3	2	2

Strategy	Word Pair Category					Between Variability
	RH	RL	UH	UL	N	
Visualization	0.375	0.25	0.25	0.125	0.375	0.094
Sentence Generation	0.25	0.375	0.125	0.25	0.125	0.094
Rote Rehearsal	0.25	0.125	0.25	0.375	0.25	0.079
None	0.125	0.25	0.375	0.25	0.25	0.079
Variability =	0.719	0.719	0.719	0.719	0.719	

Consistency = .719 *Discrimination* = .086

CWS ratio = $\frac{0.086}{0.719} = 0.12$

Participant B

Strategy responses for a participant with good strategy use.

	Word Pair Category				
Strategy	RH	RL	UH	UL	N
Visualization	4	0	6	0	0
Sentence Generation	0	4	0	6	0
Rote Rehearsal	4	4	0	0	4
None	0	0	2	2	4

Strategy	Word Pair Category					Between Variability
	RH	RL	UH	UL	N	
Visualization	0.5	0	0.75	0	0	0.316
Sentence Generation	0	0.5	0	0.75	0	0.316
Rote Rehearsal	0.5	0.5	0	0	0.5	0.245
None	0	0	0.25	0.25	0.5	0.187
Variability =	0.5	0.5	0.375	0.375	0.5	

Consistency = .45 *Discrimination* = .266

CWS ratio = $\frac{0.266}{0.45} = .59$

Participant C.

Strategy responses for a participant with good strategy use plus noise.

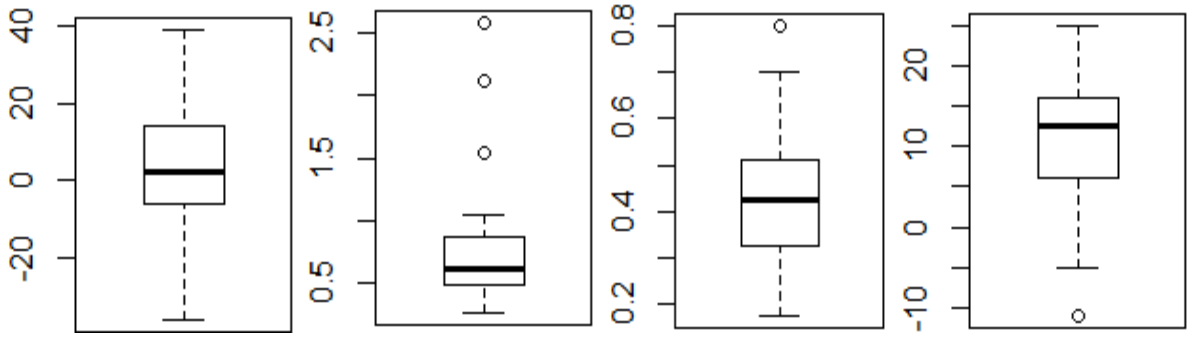
	Word Pair Category				
Strategy	RH	RL	UH	UL	N
Visualization	3	1	5	1	1
Sentence Generation	1	3	1	5	1
Rote Rehearsal	3	3	1	1	3
None	1	1	1	1	3

Strategy	Word Pair Category					Between Variability
	RH	RL	UH	UL	N	
Visualization	0.375	0.125	0.625	0.125	0.125	0.275
Sentence Generation	0.125	0.375	0.125	0.625	0.125	0.275
Rote Rehearsal	0.375	0.375	0.125	0.125	0.375	0.275
None	0.125	0.125	0.125	0.125	0.375	0.175
Variability =	0.688	0.688	0.563	0.563	0.688	

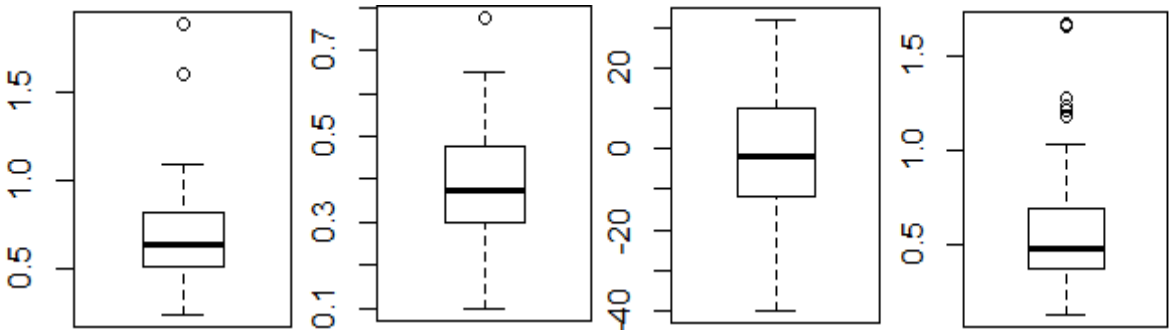
Consistency = .638 *Discrimination* = .25

CWS ratio = $\frac{0.25}{0.638} = .392$

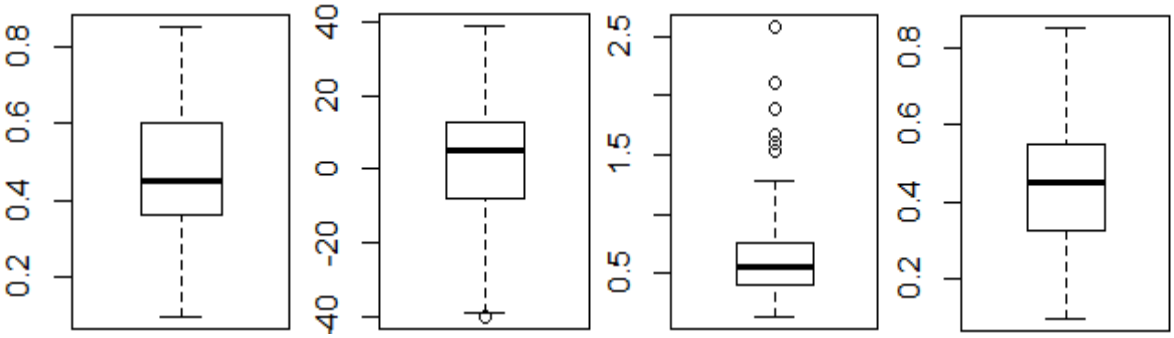
Appendix C.



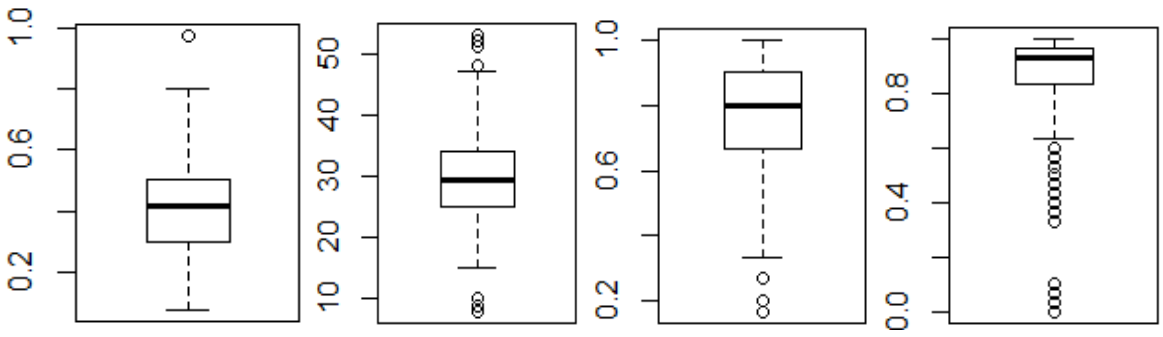
S1a – Strategy Comp. S1a - CWS S1a – Accuracy S1b – Strategy Comp.



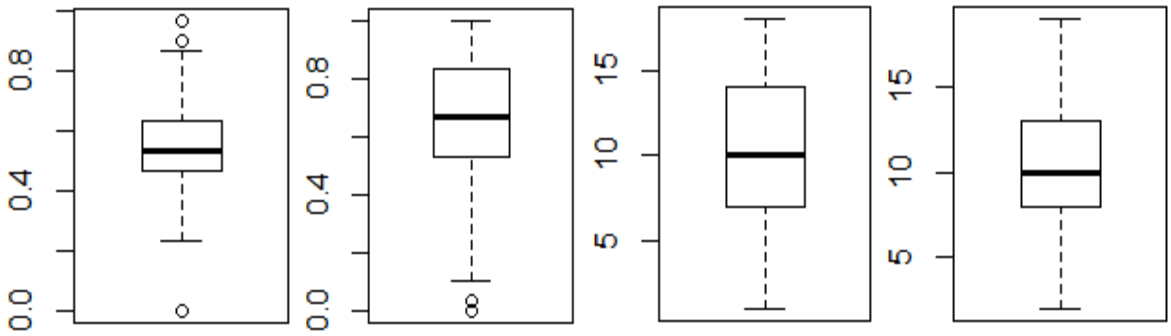
S1b – CWS S1b – Accuracy S1c – Strategy Comp. S1c – CWS



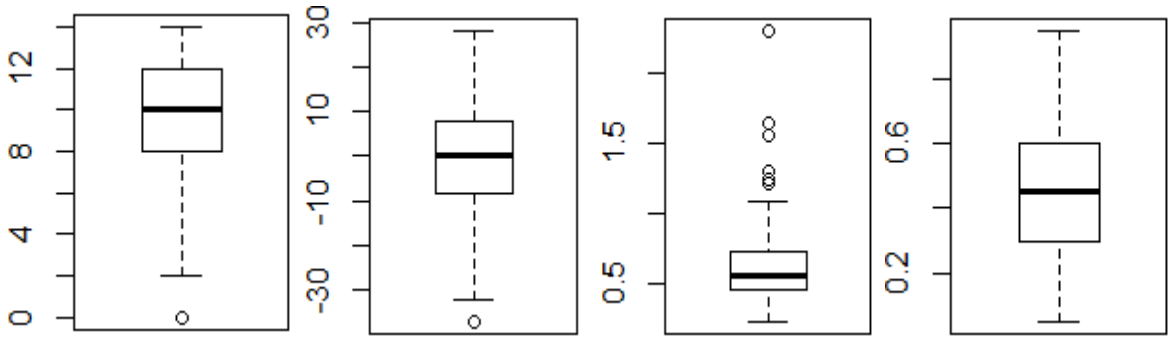
S1c – Accuracy S1all – Strategy Comp. S1all – CWS S1all – Accuracy



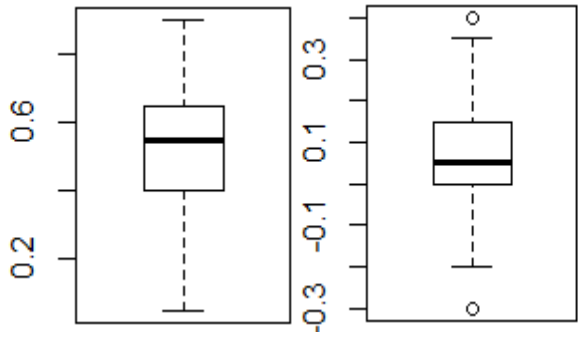
S2 – Cued Recall S2 – DFR S2 – PS–SC S2 – PS–CR



S2 – GS–SC S2 – GS–CR S2 – Ravens S2 – Letter



S2 – Number S2 – Strategy Comp. S2 – CWS S2 – Restudy



S2 – Retrieval S2 – Testing Effect

Appendix D.

The analysis in the present paper differs from the analysis run by Brewer & Unsworth (2012) and Pan et al (2015) in two ways. First two of the EM tasks were scored differently. Past studies scored the picture-source and gender-source tasks as overall correct responses, whereas the present study separated correct responses into source correct and correct rejections. Second, past studies tested for individual differences by predicting testing effects as a difference score from each of the cognitive abilities, however the present study looked for interaction between study type and cognitive abilities. In order to compare the results of the present study to past finding, an additional set of analysis will be run mimicking those used in past studies.

New composites were created for the EM measures, with picture-source and gender-source represented by overall correct proportion of responses. Then each cognitive was used to predicted the magnitude of the testing effect represented by the difference in final recall accuracy for items that received retrieval practice and items that received restudy. Only the CWS ratio had support for being related to the testing effect (BF_{10} = 5.02, r = -0.19). Finding for both EM composite and strategy composites were inconclusive (BFs_{10} = 0.47, 1.41, rs = -0.10, -0.15). There was no relation between the testing effect and IQ composite scores (BF_{10} = 0.32, r = -0.09). All relationships can be seen in Figure 13.

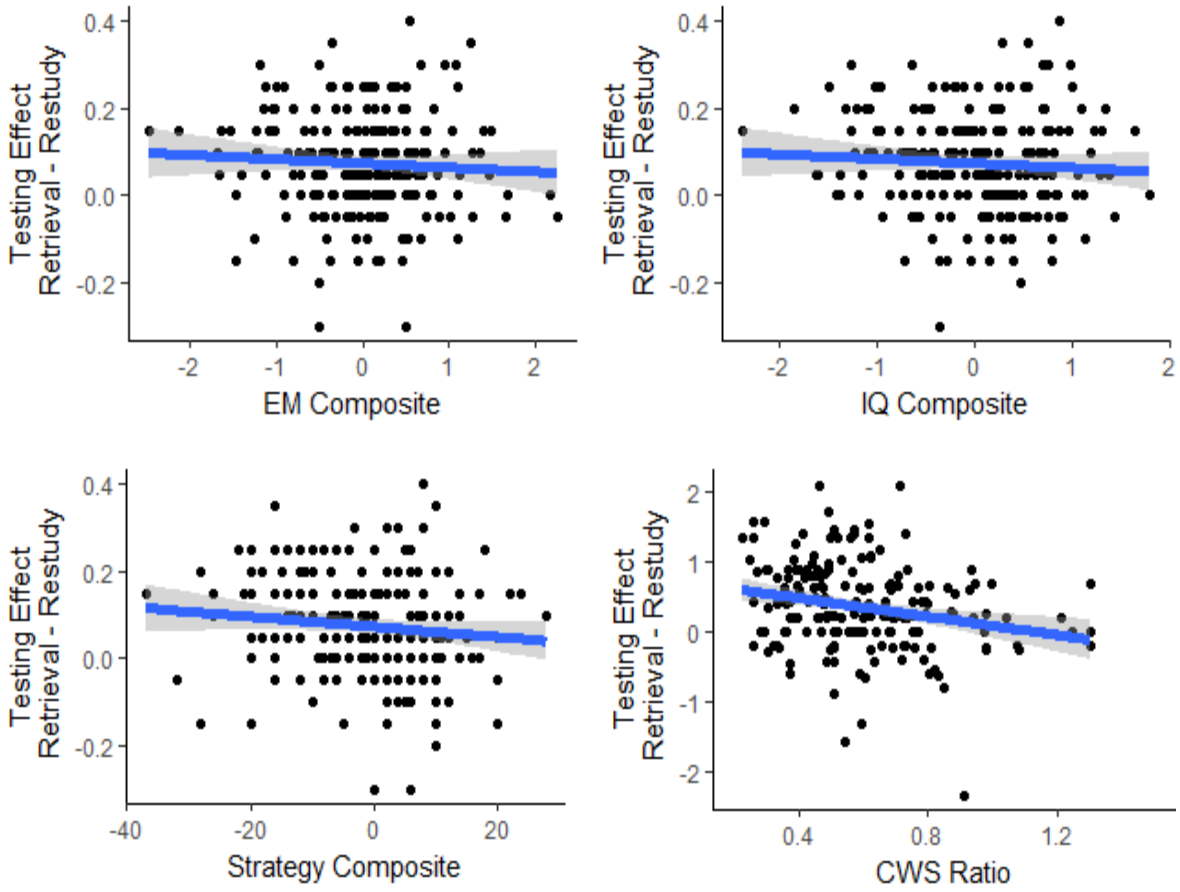


Figure 13. The relationship between all four cognitive abilities and the testing effect represented by the difference in final retrieval accuracy for items that received retrieval practice and the items that received restudy.

Appendix E.

This appendix contains the results of two additional exploratory analyses not included in the original analysis plan.

Exploratory Analysis Questions 1: Did the testing effect vary between the different word pair categories?

As reported in the discussion section, the overall testing effect found in the present study (7.5% increase) was smaller than many other testing effect experiments were a difference of around 10% has been observed. One potential explanation of the decrease in the testing effect is that five specific word-pair categories were used in the present study, whereas as most studies use just one type, and it is possible that the magnitude of the effect varied between categories. To explore differences in the testing effect across the word pair categories, an interaction between word pair type and study type was tested, and moderate support for an interaction was found ($BF_{10} = 5.3$). Next the testing effect was tested within each categories separately. Support for testing effect was found in all categories expect the nonwords, where numerically, performance was better on the restudied word pairs, Related-High Imageability: $BF_{10} = 9230905$, Related-Low Imageability: $BF_{10} = 6.98$, Unrelated-High Imageability: $BF_{10} = 6.98$, Unrelated-Low Imageability: $BF_{10} = 19061.5$, Nonwords: $BF_{10} = 0.50$, See Figure 14.

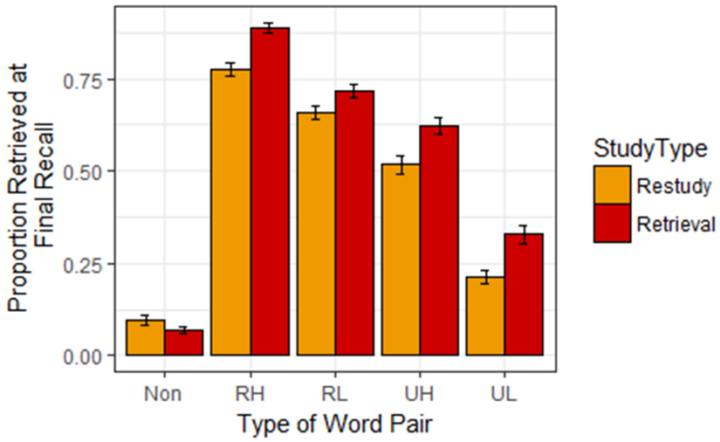


Figure 14. The testing effect for each of the five word pair categories.

Exploratory Analysis Question 2: In addition to retrieval practice having an impact on final test accuracy, does it also impact final test reaction time?

This question can be informative in two ways. First, it might be expected that some of the differences between restudies and retrieval pairs may be driven by differences in motivation. For example, after retrieval practice, a student may be more motivated to and retrieve the correct answer, leading to more time spent on retrieval, specifically for items that were not retrieved. Second, existing beliefs about the testing effect state that retrieval practice improves the accessibility of information and not the availability (Kang, McDermott, & Roediger, 2007). Under this idea, it could be expected that better accessibility of information would lead to faster retrieval of information, leading to short reaction times for items that are correctly retrieved.

Exploratory analyses was run comparing the final test RTs for correct and incorrectly retrieved items following both restudy and practice retrieval. It was expected that if motivation played a role in improve memory retrieval after retrieval practice that, retrieval practice RTs would be longer for incorrectly retrieved words. Additionally, since retrieval practice is thought to improve the availability of information, it was expect

that for correctly retrieved items, RTs would be faster for items that received retrieval practice.

Support was found for an interaction between study type and final accuracy in predicting retrieval RT, $BF_{10} = 174.12$. Follow-up analyses found no differences in RT for incorrectly retrieved words, $BF_{10} = 0.12$, suggesting that motivation did not appear to be differing for the two type of study. Different than expected, decisive support was found for differences in RT for correctly retrieved words, $BF_{10} = 4.0 \times 10^{13}$, but with RTs being greater retrieval practice pairs than restudied pairs, see Figure 15. Although this finding is in the opposite direction than what was found, it can still be consistent with existing theories of the testing effect. Specifically, the ERH posits that retrieval practice enhances later memory performance by elaborating on the existing memory trace and creating more potential routes for future success (Carpenter & DeLosh, 2006). Relating this to the observed RT data, if learners have more "memory routes" to explore after retrieval practice, this could leading to more overall time spent processing, and longer retrieval RTs. In contrast, additional routes have not been created for the restudied items, and there for less processing and short RTs would be needed. To ensure that the benefits of retrieval practice were not driven solely by increases in processing during final retrieval analyses were run to test for a benefit of retrieval practice after statistically controlling for difference in retrieval RT. Decisive evidence for a testing effect was still found, $BF_{10} = 189676574863$.

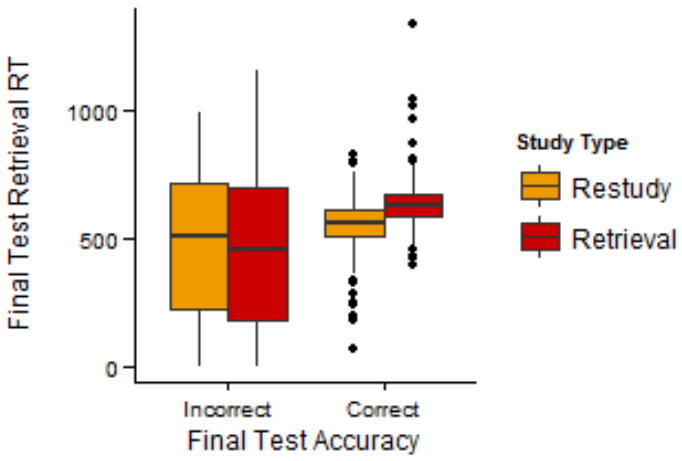


Figure 15. RT distributions for correct and incorrectly recalled items after restudy or retrieval practice.

References

Abel, M. & Roediger, H. L. (2017). Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affect by practice format. *Memory & Cognition, 45*(1), 81-92.

Avci, G., Woods, S. P., Verduzco, M., Sheppard, D. P., Sumowski, J. F., Chiaravalloti, N. D., & Deluca, J. (2017). Effect of retrieval practice on short-term and long-term retention in hiv+ individuals. *Journal of the International Neuropsychological Society, 23*(3), 214-222.

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging, 4*, 3-9.

Balota, D. A. Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. K. III. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging, 21*, 19-31.

Batsell, W. R. Jr., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology, 44*(1), 18-23.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranges associates. *The Journal of General Psychology, 49*(2), 229-240. doi: 10.1080/00221309.1953.9710088

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

Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of verbal learning and verbal behavior, 9,* 529-533.

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

Budescu, D. (1993). Dominance Analysis: A new approach to the problem of relative importance of predictor in multiple regression. *Psychological Bulletin, 114*(3), 542-551.

Camp, C. J., Markley, R. P., & Kramer, J. J. (1983). Naïve mnemonics: What the "do-nothing" control group does. *The American Journal of Psychology, 96*(4), 503-511. doi: 10.2307/1422571

Canal, L., Bonini, N., Micciolo, R., & Tentori, K. (2012). Consistency in teachers' judgments. *European Journal of Psychology of Education, 27*(3), 319-327.

Cantor, A. B. (1996). Sample Size calculations for Cohen's Kappa. *Psychological Methods, 1*(2), 150-153.

Carpenter, S. K. & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34* (2), 268-276.

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2015). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review,* 1-23. doi:10.1007/s10648-015-9311-9.

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, 6,* 760-771.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cogniton, 36*(2), 438-448. doi: 10.3758/MC.36.2.438.

Carrol, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect, T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, test integration, and test format. *European Journal of Cognitive Psychology, 19,* 580-606.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61,* 153-170.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 37-46.

Dunlosky, J. & Hertzog, C. (2001). Measuring strategy production during associative learning: The relative utility of concurrent versus retrospective reports. *Memory & Cognition, 29*(2), 247-253.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising direction from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7,* 86-106.

Einhorn, J. J. (1974). Expert judgment: Some necessary condition and an example. *Journal of Applied Psychology, 59,* 562-571.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392-399.

Hammond, K. R. (1996). Human judgment and social policy. Oxford University Press, New York.

Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues, and applications (pp. 76-99). Thousand Oaks, CA: Sage.

Kader, G. D., & Perry, M., (2007). Variability for categorical variables. *Journal of Statistical Education, 15*(2), 1 – 16.

Kass, R. E. & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795.

Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology, 7*, 350: 1-8.

Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for juided retrieval in elementary school children. *Journal of Applied Research in Memory and Cognition, 3*, 198-206. doi: 10.1016/j.jarmac.2014.07.008.

Karpick, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation, 61*, 237-284. doi:10.1016/B978-0-12-800283-4.0007-1

Karpicke, J. D. & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaboartive encoding. *Journal of Memory and Language, 67,* 17-29. doi:10.1016/j.jml012.02.004.

Kirchhoff, B. A. (2009). Individual differences in episodic memory: The role of self-initiated encoding strategies. *The Neuroscientist, 15*(2), 166-179.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In Aitken, A.J., Bailey, R. W., and Hamilton-Smith, N. (Eds.), *The Computer and Literary Studies.* Eidenburgh: University Press.

Kline, R. B. (2011). *Principles and practice of structural equation modeling, 3rd ed.* New York: The Guildford Press.

Landauer, T. K. & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In *Practical Aspect of Memory* (Gruneberg, M. M. et al., eds), pp. 625-632, Academic Press.

Landis, J.R.; Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1): 159–174. *doi:10.2307/2529310.*

Larsen, D. P., Butler, A. C., Roediger, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomized controlled trial. *Medical Education, 43,* 1174-1181.

Lehman, M., Smith, M. A. & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-bases learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40,* 1787-1794.

Loveday, T., Wiggins, M. W., Searle, B. J., Festa, M., & Schell, D. The capability of static and dynamic features to distinguish competent from genuinely expert practitioners in pediatric diagnosis. *Human Factors, 55*(1), 125-137.

Lyle, B. K. & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*(2), 94-97.

Ma, X., Yang, X., Yanru, L., & Zhao, A. (2016). Prior Knowledge level dissociates effects of retrieval practice and elaboration. *Learning and Individual Differences, 51*, 210-214.

Marsh, E. J., Roediger, H, L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple choice questions. *Psychonomic Bulletin and Review, 14*(2), 194-199.

Martin, C. J., Boersma, F. J., & Cox, D. L. (1965). A classificiation of associative strategies in paired-associate learning. *Psychological Science, 3*, 455-456.

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 effect of quiz feedback and placement. *Journal of Educational Psychology*, *103*(2), 399-414. doi: 10.1037/a0021782.

McDaniel, M. A., Kowitz, M. D., & Dunay, P. K. (1989). Altering memory representation through recall: The effects of cued-guided retrieval processing. *Memory & Cognition, 17*, 423-434.

McDermott, K. B., Kang, S. & Roediger, H. L., III. (2005). *Test format and its modulation of the testing effect.* Paper presented at the biennial meeting of the

Society for Applied Research in Memory and Cognition, Wellington, New Zealand.

McGee, R. (1980). Imagery and recognition memory: The effects of relational organization. *Memory and Cognition, 8*(5), 394-399. doi: 10.3758/BF03211135

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*(5), 519-533. doi: 10.1016/S0022-5371(77)80016-9

Naveh-Benjamin, M., Brav, T. K., & Levy, O. (2007). The associative memory deficit of older adults: The role of strategy utilization. *Psychology of Aging, 22*(1), 202-208.

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.

Paulet, K., O'Hare, D., & Wiggins, M. Measuring expertise in weather-related aeronautical risk perception: The validity of the Cochran-Weiss-Shanteau (CWS) Index. *The International Journal of Aviation Psychology, 19*(3), 201-216.

Pashler, H., Cepeda, N. J., Wixted,J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 3-8.

Paivio, A. & Yuille, J. C. (1969). Changes in associative strategies and paired-associate learning over trials as a function of word imagery and type of learning set. *Journal of Experimental Psychology, 79*, 458-463.

Paivio, A., Yuille, J. C., & Smythe, P. C. (1966). Stimulus and response abstractness, imagery, and meaningfulness, and reported mediators in paired-associate learning. *Canadian Journal of Psychology, 20,* 362-377.

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.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358, 534 nonwords: The ARC nonword database. Quarterly Journal of Experimental Psychology, 55A, 1339-1362.

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*(3), 283-302. doi: 10.1037/a0023956.

Richardson, J. T. E. (1978). Reported mediators and individual differences in mental imagery. *Memory and Cognition, 6,* 376-378.

Richardson, J. T. E. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. *Psychonomic Bulletin and Review, 5,* 597-614.

Roberts, W. A. (1968). Alphabetic coding and individual differences in modes of organization in free-recall learning. *American Journal of Psycholgy, 81,* 433-438.

Roediger, H. L., III. & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20-27.

Roediger, H. L. III. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improved long-term retention. *Psychological Science, 17*(3), 249-255.

Roth, W. & Mavin, T. (2015). Peer assessment of aviation performance: Inconsistent for good reasons. *Cognitive Science, 39*(2), 405-433.

Satorra, A., & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. Psychometrika, 75, 243-248.

Shanteau, J., Weiss, D. J., Thomas, Rickey. P., & Pounds, J. C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research, 136*, 253-263.

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

Stoff, D. M. & Eagle, M. N. (1971). The relationship among reported strategies, presentation rate, and verbal ability and their effects on free recall learning. *Journal of Experimental Psychology*, *87*, 423-428.

Thurstone, L. L. & Thurstone, J. (1962). *Tests of primary mental abilities* (revised ed.). Chicago: Chicago Science Research Association.

Tse, C., Balota, D. A., & Roediger, H. L. III. (2010). The benefits and costs of repeated testing on the learning of face-name pairs on healthy older adults. *Psychology of Aging*, *25*(4), 833-845. doi: 10.1037/a0019933.

Whiffen, J. W. & Karpicke, J. D. (in press) The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning Memory and Cognition.*

Wiggins, M. W. (2014). Differences in situation assessments and prospective diagnoses of simulated weather radar returns amongst experience pilots. *International Journal of Industrial Ergonomics 44*(1), 18-23.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, Version 2, *Behavioural Research Methods, Instruments, and Computers, 20*, 6-11.

Witteman, C., Weiss, D. J., & Metzmacher, M. (2012). Assessing diagnostic expertise of counselors using the Cochran-Weiss-Shanteau (CWS) index. *Journal of Counseling & Development*, *90*(1). 30-34.

Witteman, C. L. M. & Tollenaar, M. S. (2012). Remembering and diagnosing clients: Does experience matter? *Memory*, *20*(3), 266-276.