

LEARNING FROM TESTS: TESTING AS A STUDY TOOL

by

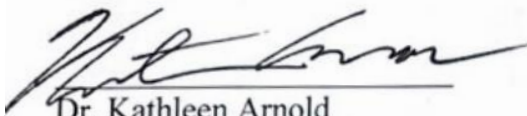
Lesli A. Taylor

A thesis submitted to the faculty of Radford University in partial fulfillment of the requirements  
for the degree of Master of Arts in the Department of Psychology


Thesis advisor: Dr. Kathleen Arnold

May 2019

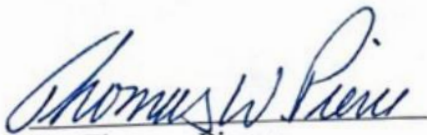
Copyright 2019, Lesli A. Taylor

  
Dr. Kathleen Arnold  
Thesis Advisor

5/10/19  
Date

  
Dr. Catherine Middlebrooks  
Committee Member

5/10/19  
Date

  
Dr. Thomas Pierce  
Committee Member

5/10/19  
Date

## TESTING AS A STUDY TOOL

### **Abstract**

The present study investigated the role of test taking as a study tool, particularly, if test taking is a means of making people more effective in their studying. It was also investigated whether or not receiving feedback on interpolated tests encouraged participants to study more effectively. It was predicted that participants who took interpolated tests would study items of higher value longer than those of lower value, have higher final value scores, and would recall more items on a final test than any other participants, all of which are indicative of more strategic studying. After studying a list of words paired with values ranging from 1-10, some participants took interpolated free recall tests and either received feedback or not, while others completed a distractor task, and the remaining participants took only one test. All participants studied a final list and took a final test, which is the test of interest. Taking tests with feedback prompted participants to study items of higher value longer. In addition, participants who took tests without feedback had higher value scores and were more selective in their studies. Overall, the results of this study indicate that taking tests is superior to not taking tests and that feedback may increase the effectiveness of testing as a study tool.

*Keywords:* testing, studying, interpolated tests, feedback

Lesli A. Taylor, M.A.

Department of Psychology, 2019

Radford University

**Table of Contents**

Abstract.....	ii
Table of Contents.....	iii
List of Tables and Figures .....	iv
Chapter 1. Introduction.....	1
Chapter 2. Methods.....	15
Chapter 3. Results.....	20
Chapter 4. Discussion.....	43
References.....	48-51

**List of Tables and Figures**

Table 1.....	23
Table 2.....	24
Figure 1.....	25
Figure 2.....	29
Figure 3.....	32
Figure 4.....	36
Figure 5.....	40

## Chapter 1 – Introduction

### Overview of Metacognition

Students have to make a lot of decisions while they study. For example, students must decide how long to study, what to study, and what study strategies to use. Nelson and Narens (1990) developed a model that describes how students make these decisions. This model suggests that students make these decisions, at least in part, by using their metacognition. Metacognition is the knowledge someone possesses about how his/her own cognitive processes operate. Through metacognition, students observe their own knowledge, make judgments about their knowledge based on those observations, and then make study decisions based on those judgments.

According to Nelson and Narens (1990), cognitive processes are split into two levels: the meta-level and the object-level. Meta-level information is more abstract knowledge, such as what we know about what we know. Object-level is more concrete, such as studying. This model states that there are two dominant relations in metacognition that deal with the flow of information between these two levels. The first relationship is called metacognitive control. Metacognitive control is the flow of information from the meta-level to the object-level, which changes the object-level process. That change can be to either initiate an action, to continue an action, or to terminate an action. An example of this would be what occurs during studying. A student is studying and then makes judgements about his/her learning of the material, a meta-level process. The meta-level influences the student to either continue studying or terminate studying (object-level processes) based on the individual's meta-level judgments, such as judging that he/she has or has not learned material.

## TESTING AS A STUDY TOOL

The second relationship that was discussed by Nelson and Narens (1990) is one of monitoring. Object level observations can influence meta-level changes, through the use of metacognitive monitoring. This monitoring would occur when someone is studying. If a person is reading material over and over again until he/she feels as though the material is easier to understand, also known as the material becoming fluent, that is an object level observation. Then the individual uses that object level observation of fluency to make a judgment on his/her studying, meaning the individual is monitoring his/her thoughts. Lastly, the person then uses that thought to terminate studying upon the material becoming fluent, demonstrating control. That could mean he/she either begins studying, continues studying, or terminates studying based on his/her monitoring of object-level activities, which influences metacognitive level judgments, which in turn directs metacognitive control processes (Nelson & Narens, 1990).

One important decision students make during studying is how long to study any one concept. People have to decide how much time to allocate to studying individual items, decisions which are limited by the time available between the induction of studying and the taking of a future test (Nelson & Narens, 1990). Students must figure out how to divide that limited time in order to maximize their exam scores and their knowledge. People continue to study an item until they judge that they have achieved a level of knowledge of the item that is consistent with their study goals (Nelson & Narens, 1990). Study goals are set by learners and they are used as goalposts for the learner to reach in order to terminate study.

Son and Metcalfe (2000) examined the relationship between metacognitive judgments and decisions learners make while studying. Specifically, they examined the allocation of study time when items varied in difficulty. Using undergraduate students, the researchers divided them into two groups: high time pressure and low time pressure. In each condition, the participants

## TESTING AS A STUDY TOOL

had to study eight different sonnets and then were tested on them. Ease of learning judgements were made by the participants and were a rating of how easy the given material was. They found that people used their time strategically and chose to study items that were easier if time constraints and pressure were more prevalent. This finding from Son and Metcalfe (2000) suggests that metacognitive judgements are a driving force in learning because they help to direct a person's study behavior to be strategic for the time that the individual is allowed. This experiment is a demonstration that metacognitive judgements are used to make study time decisions.

### **Studying Overview**

Studying encompasses a variety of behaviors. According to Winnie and Hadwin (1998), any behavior that leads to a relatively permanent change in cognitive structure can be considered studying. They suggest that studying has six key features that distinguish it from other learning activities. First, studying rarely includes intervention that is frequent or direct from instructors. This means that teachers often do not tell the students what is or is not important to study, nor do they tell them how to study. Instead, they outline lecture material and the student is left to decide what to study. Second, while group studying sometimes occurs, studying is an activity that is often done alone. Third, studying usually starts with a goal set forth by the instructor that is therefore interpreted by the student. Fourth, studying usually involves incorporating information from more than one source. This means that students must take information from sources other than lecture, such as a textbook, class notes, and/or journals and other materials that can be accessed via a database. Fifth, studying occurs in the environment of the learner's choosing. This means that the person engaged in studying can set the environment up the way that he/she prefers, quiet versus with music or television, with or without food or drink, and with as many or

## TESTING AS A STUDY TOOL

as little distractors as possible. Lastly, studying almost always has a trace of cognitive processing that is observable. This can be through rewritten class notes, highlighted text, diagrams, or notecards, among other study materials.

Winne and Hadwin (1998) suggested that studying involves complex cognitive and motivational processes, which are goal-directed. Those processes are what drive people to get studying done. The processes suggested by Winnie and Hadwin (1998) involve what is known as metacognitive monitoring, which influences the studiers' goals. This goal-directed studying involves the learner regulating his/her own learning, a process known as self-regulated learning.

### **Self-Regulated Learning**

Self-regulated learning is a type of learning during which learners are able to regulate their study time, what study strategies to use, when to terminate or begin study, and many more behaviors. It is not uncommon for any learner to be presented with more information than he/she is able to remember in a limited amount of time (Castel et al., 2013). Learners need to be able to selectively attend to information that is goal-relevant, or information that relates directly to the test, paper, or other work that the student is attempting to learn about in order to score well (Castel et al., 2013). Decisions of how to allocate study time are therefore controlled by people's agendas, which they develop to achieve their goals (Ariel, Dunlosky, & Bailey, 2009).

Agendas, which are created by learners, are used to make an allotment of study time to various items. Agendas also contain criteria for what items the learner will decide to study. Goals, a person's objective for learning, and agendas, a plan created by the person to achieve his/her goal, are based on various items that influence learners, such as the desired grade, the difficulty of the material, as well as how much study time has been allotted for the material.



## TESTING AS A STUDY TOOL

Dunlosky and Ariel (2011) suggested that learners use agendas to make decisions about their learning. This agenda-based-regulation framework suggests that learners will spend more time studying items that are more difficult unless there is not enough time available for sufficient learning or the reward for correctly recalling easier information is higher. That is, this framework suggests that people are capable of making efficacious study decisions that are tailored for their specific agenda, which is based on their study constraints and personal goals. This framework is supported by a study conducted by Ariel et al. (2009), which found that people chose to study items that had a higher likelihood of appearing on a future test more than those items that had a lower likelihood of appearing on a future test. This suggests that the participants' agendas influenced the way they studied; while studying for a future test, they chose to study items that may appear on that test.

The framework set forth by Dunlosky and Ariel (2011) is complementary to the framework set forth by Nelson and Narens (1990). Both parties proposed the idea that agendas set learners' priorities. They also proposed that metacognitive monitoring is used as a tool by which learners make study decisions within that agenda.

### **Value Directed Learning**

Given the supporting evidence that learners choose to study based on a regulated agenda, it is important to understand how that agenda can be influenced by outside factors such as time constraints or study formatting (Dunlosky & Ariel, 2011). One way to measure the influence of these outside factors is to change the value of the items being studied, as well as the way that they are presented. Middlebrooks and Castel (2018) designed an experiment to observe participants' ability to self-regulate their learning under different encoding conditions. The goal of the study was to determine whether or not people would manipulate their learning agendas.

## TESTING AS A STUDY TOOL

In their first experiment, participants were assigned to either a simultaneous study condition or a sequential study condition. A simultaneous study condition is a condition in which learners are presented with stimuli on the same page at the same time. A sequential study condition is a condition in which learners are presented with stimuli one at a time. The study materials were lists of 20 words, with each word assigned a value from one to 10 (two words for each point value). After one study period, participants were then given a test to see how many of the words they were able to remember. Participants who were in the sequential study condition recalled fewer items. Additionally, when items were studied in a sequential fashion, value of the items had a smaller effect, meaning that participants in the sequential list condition paid less attention to value when remembering items than in the simultaneous condition. This experiment provides evidence that people's agendas do change when the way they are able to attend to the material has been shifted. This is due to the item value not mattering as much when study material presentation differed (Middlebrooks & Castel, 2018).

Additionally, Middlebrooks and Castel (2018) found evidence of value directed learning. High value information was recalled more often than low-value information. This finding is consistent with prior research. For example, Ariel and Castel (2014) found participants prioritized high value information when they were forced to make a choice of what to remember because there was too much information to remember in one sitting. Similarly, Middlebrooks et al. (2016) also found evidence that participants prioritize high value items when they are under constrained study time.

This research provides evidence for how students make study decisions. Students use agendas in order to guide their learning, and these agendas are influenced by study time, difficulty of material to be learned, and what is deemed to be important by the student. Those

## TESTING AS A STUDY TOOL

study decisions are guided by the students' metacognition and their ability to monitor their own learning.

### **Testing Overview**

While traditional studying is important to student learning, so is testing. Testing can have many effects on learning. One of the most well-studied effects of testing on learning, traditionally referred to as “the testing effect,” is the impact on long-term retention of information; learners remember more information after being tested on it than after using more traditional methods of studying (Karpicke & Roediger, 2008). This testing effect occurs even when more time is afforded to studying than testing (Karpicke & Roediger, 2008). This effect of testing has been found across populations, including elementary school children (McDermott, Agarwal, D’Antonio, Roediger, & McDaniel, 2013) and college students (Roediger & Karpicke, 2006).

What about testing is so powerful? There are many theories as to why testing may help students. One theory suggests that taking tests enhances learning by strengthening the connection between the retrieval cue and the information to be retrieved. When people take tests, they have to engage in the cognitive process known as retrieval. When engaging in retrieval, people must try to remember, or “retrieve,” the information to which they have at some point been exposed. Retrieving information may lead to elaboration of the memory, which makes it more likely that the memory can be retrieved again in the future (Roediger & Butler, 2011).

While testing is an incredibly effective learning tool, many learners still do not engage in retrieval practice when they are attempting to learn new material (Tullis, Finley, & Benjamin, 2013). This outside of the classroom learning, or studying, is often self-directed with little educator input. Self-regulated learning, as discussed earlier, is a process by which people have to

## TESTING AS A STUDY TOOL

rely on their own metacognitive knowledge, including their knowledge of what study strategies are most effective. Being an effective learner implies that learners must make appropriate study decisions (e.g., Metcalfe, 2009), guided by accurate metacognitive monitoring. Studies on the testing effect suggest that using retrieval would be an appropriate study decision in many situations considering its wide range of benefits. One testing effect that students should consider when making decisions on how to study is the interpolated testing effect.

### **Interpolated Testing**

In a now seminal paper, Szpunar, McDermott, and Roediger (2008) found evidence for what is now known as the interpolated testing effect. Over a series of four experiments, they found that testing, when administered between studying different lists of items, benefits learners. Specifically, persons who are tested on the first set of material prior to studying the next set of material are not only better able to recall information from that second set of material, but they also experience fewer intrusions from the prior information. Intrusions are thought to be reduced through the suppression of irrelevant cues from prior lists, which help to protect the attempted to-be-studied information from proactive interference (Szpunar et al., 2008).

More recently, this effect was replicated by Szpunar, Khan, and Schacter (2013) using video lectures. In two different experiments, participants were instructed to watch a 21-minute video lecture that was broken up into four segments with a break between each segment. In experiment one, during these breaks, students first completed one minute of unrelated arithmetic problems. After completing the problems, the participants were told the computer would randomly decide if they would complete two more minutes of arithmetic problems, again unrelated to the lecture, or have a 2-minute knowledge test on the most recently watched portion of the lecture. In reality, the participants were divided into two groups: the interpolated tests

## TESTING AS A STUDY TOOL

condition in which they took a test after watching each video segment and the no interpolated tests condition in which they did additional math problems after watching the first three segments. Both conditions took a test on the fourth segment as well as a final cumulative test. In the second experiment, a third condition was added: a restudy condition in which after the first three segments, participants studied test questions and their answers, which were presented together. Just as in Experiment 1, the participants were told that they would be randomly assigned a task (math, test, or restudy) after each segment. In all three conditions, participants took a test on the fourth segment and took a final cumulative test (Szpunar et al., 2008).

The results of both experiments revealed that, in conditions where participants had to complete interpolated tests, there were fewer instances of off-task behaviors and mind-wandering. In addition, participants in the interpolated testing condition learned more on the test following the fourth segment in comparison to the distractor task and restudy conditions. These results suggest that the uses of interpolated tests, or the insertion of tests in the middle of learning, are effective learning tools. They aid in helping learners sustain their attention and also serve to discourage off-task behaviors and reduce proactive interference, while encouraging appropriate study-related behaviors (Szpunar et al., 2013).

### **Testing and Item Value**

Testing may also help self-regulated learners study more efficiently. Learners often show better memory for things they have deemed valuable to remember. In a series of experiments, Cohen, Rissman, Hovhannisyan, Castel, and Knowlton (2017) investigated the role that item value plays in encoding strategy selection. The experiments consisted of participants being presented with sets of words that were associated with point values, both high and low. Some participants were not given a test after each list of word-value pairs, while some participants

## TESTING AS A STUDY TOOL

were given a free recall test after each list of word-value pairs, meaning that they were asked to recall as many of the words as they could from the prior list, after each list. All participants were tested after being presented with the final list of words and value pairs.

The results of the experiment suggested that when paired with free recall tests with feedback value increased the participants' likelihood that they would recall the word. If the word was paired with a high value number, numbers 10, 11, or 12, then the participant was significantly more likely to recall the word than if it was paired with a low value number, 1, 2, or 3. Thus, a word being given more value does play an important role in increasing the likelihood of remembering that word. When participants were given tests after their study lists, they were more likely to recall and be familiar with the words they were being tested on than the participants who did not receive multiple tests. This study did not include a one test condition, however, so there was no way to exclude the possibility that one test is enough to produce these same results.

### **Testing with Feedback**

While testing can enhance learning, another factor that has been suggested to contribute to learning is feedback. Testing is sometimes followed by feedback, which often enhances the effect of testing on learning (Kang, McDermott, & Roediger, 2007). Feedback can either be immediate, usually found in online formatting, or delayed, like in a traditional classroom setting. Feedback can also come in many forms, including more traditional formats such as being given the correct answer, being only told if you are right or wrong, or just being told an overall score. It has been argued that feedback is important for learners because it has the potential to enhance long-term retention (Butler, Karpicke, & Roediger, 2007).

## TESTING AS A STUDY TOOL

In a 2007 experiment, Butler et al. investigated the role of feedback on learning. Participants experienced either immediate or delayed feedback in either the traditional or answer-until-correct fashion. They found that regardless of feedback type, participants who received feedback had a higher proportion correct on a final cued recall test than those who did not have feedback. This suggests that feedback as a whole is important to learners when it comes to their ability to learn and retain material; however, there is a lack of literature in both the interpolated testing effect as well as the value-directed learning literature on what role, if any, feedback plays in learning through tests.

### **Purpose of Current Experiment**

Studying and testing play an intricate role in the life of students. The purpose of the current study was to investigate the relationship between testing and studying, specifically how testing may affect the efficacy of studying. Prior studies have shown that interpolated testing, which is taking tests after you finish learning one chunk of information before studying the next chunk, increases how much you can learn on the next chunk of information. One reason that interpolated testing benefits learners may be because they make better study decisions. Prior studies have shown that learners make study decisions based on their goals, for example, getting an A. To achieve these goals, they create agendas, which serve as guides to help them achieve their goals, based on an assortment of things such as the difficulty of the material, as well as how much study time has been allotted for the given material. Agendas can be affected by manipulations such as time constraints and presentation of study material. Interpolated tests may also affect agendas by teaching students what worked and what did not, information that is part of one's metacognitive knowledge. Therefore, interpolated testing may make students more efficacious students. Feedback may also be beneficial to learners and provide them with

## TESTING AS A STUDY TOOL

information about what they do and do not know. The presence of feedback may also make students more efficacious studiers.

In this study, the effects of interpolated testing and feedback on study effectiveness will be tested by examining value-item directed learning. Specifically, items will be given different values and learners will be instructed that their goal is to get as high a score as possible. Participants will study a series of lists. One fourth of participants will take tests without feedback after each list, a fourth will take a test with feedback after each list, a fourth of participants will not be given interpolated tests, and the final condition will take a test after they study the third list. After studying the final list, all participants will take a test. If testing alone increases the effectiveness of study strategies, on the final list learners in both testing conditions should spend more time studying high value items and this strategy should result in a higher test score. If feedback enhances the effect of interpolated testing, learners in the feedback condition should show more effective study strategies and increased learning on the last list than learners who took tests without feedback.

### **Primary Hypotheses**

Based on past literature regarding study behaviors and testing, the following hypotheses were developed:

H1: Participants who take interpolated tests will study items of higher value (items with a value of 8 or above) longer, indicating more strategic study, when compared to participants who do not take tests.

Rationale: Retrieval practice via interpolated testing is an incredibly effective form of studying, such that learners attend better to the material and practice better study habits (Szpunar



## TESTING AS A STUDY TOOL

et al., 2013). If participants take more tests, they may make the decision to study higher value items to increase their score instead of studying lower value items.

H2: Participants who receive feedback on tests will be more strategic in their study material selection than participants who do not receive feedback.

Rationale: Prior research (e.g., Middlebrooks & Castel, 2018) has only explored tests with feedback regarding value-directed learning. Additionally, most interpolated testing experiments have not explored the effect of feedback. It is known that feedback enhances the testing effect, so it may help learners be more strategic in their studies. This hypothesis is an exploration of if feedback has an influence on study strategies.

H3: Participants who take interpolated tests will have higher total value scores on the final test, which will be indicative of a more effective study strategy.

Rationale: Retrieval practice is an incredibly effective form of studying, such that learners attend better to the material and practice better study habits (Szpunar et al., 2013). If participants take more tests, they may study higher value items, which should lead to higher scores, indicating more strategic study choices.

H4: Participants who receive feedback on interpolated tests will produce higher value scores on the final test than participants who do not receive feedback after interpolated tests.

Rationale: Prior research (e.g., Middlebrooks & Castel, 2018) has only explored tests with feedback regarding value-directed learning. Additionally, most interpolated testing experiments have not explored the effect of feedback. This hypothesis is an exploration of if value-directed learning only happens in the presence of feedback, as well as if interpolated testing can in part be attributed to study effectiveness if the tests do not include feedback.

## TESTING AS A STUDY TOOL

H5: Participants who take tests will recall more total items, regardless of item value, on the final test when compared to participants who do not take tests.

Rationale: Interpolated testing increases recall of subsequent material via reduction of proactive interference as well as increasing test-taker attention. This hypothesis is a replication of the interpolated testing effect found in Szpunar et al. (2013).

## **Chapter 2 – Method**

### **Participants**

Participants were 32 Radford University undergraduate students, randomly divided into four between-subject conditions. Of the 32 participants, 69% were female and 31% were male. Most (81%) had completed some college, while 12% had an associate degree and 6% had a bachelor's degree. All participants indicated that they were fluent in the English language. Most participants indicated English was their first language (87%), and the remaining indicated learning English after age 2 (13%). Analyses were performed twice: first, with all participants included, and again excluding 8 participants who indicated that they experienced an issue during the study, creating an *N* of 24. The most common issue these eight participants reported was a skipped screen, meaning they either skipped a word that they wanted to study or skipped a distractor task, which would affect the spacing of the tests and therefore could skew results. Participants received partial fulfillment of class requirements by participating in the present study.

### **Design**

The experiment included two independent variables each with three levels, all manipulated between-subjects: tests versus no tests and feedback versus no feedback. Participants were randomly assigned to one of four conditions: test after every list without feedback condition, test after every list with feedback condition, no test condition, and halfway test (one test given after the middle list) condition. In both test after every list conditions, participants were asked to recall words from the previously studied list before studying the next list. The test with feedback condition received a score for each test, which was calculated by adding together the values for each word the participant accurately remembered. In contrast, the

## TESTING AS A STUDY TOOL

test with no feedback condition as well as the halfway test condition received no score on each test. In the no test condition as well as the halfway test condition, participants took part in a distractor task between studying lists.

This study has four dependent variables: study time of items on the final list, value score on the final test, selectivity index, and proportion of items recalled on the final test. Study time was used to measure participant study behavior. Strategic study is defined as spending more time studying high value items than low value items and was measured using study time. In order to measure if the participant studied strategically, the selectivity index was used. The selectivity index is a calculation that compares a participant's value score to an idealized value score, which is the maximum score he/she could have gotten given the number of items he/she recalled (Castel et al., 2002). A selectivity index of positive one is indicative of perfectly selective study habits, meaning a person studied and remembered the highest value items. If a person's selectivity index is negative, he/she was perfectly unselective; meaning he/she only remembered the lowest value items. A score of zero would indicate the absence of selectivity (i.e., item value had no bearing on what the person remembered).

A final value score was also calculated for each participant. This was determined by the sum of the values assigned to the words that were accurately recalled by the participant. Higher value scores are indicative of a more effective study strategy. Percentage of words correctly recalled (regardless of value) was also calculated for each participant. This was used to determine if tests and/or feedback have an effect on how much information is remembered.

### **Materials**

The study was designed and presented to participants via the Collector program in the lab (Garcia, Kerr, Blake, & Haffey, 2015). Stimuli consisted of six lists with 20 words each, which

## TESTING AS A STUDY TOOL

were selected from a pool of words previously used for this type of study (Middlebrooks & Castel, 2018). Each of the words was randomly assigned a value ranging from 1 to 10 points, with two words assigned to each value per list. From a bank of 665 nouns, verbs, and adjectives, the words in each list were randomly selected without replacement. The words in this bank all ranged in length from four to six characters with an average of 7.52 ( $SD = 1.02$ ) characters. The 120 words given to each participant to study were selected at random from the word bank to avoid any potential item effects (Murayama, Sakaki, Yan, & Smith, 2014). Distractor items consisted of basic arithmetic problems. These consisted of multiplication problems with numbers that have two or more digits (e.g.,  $123 \times 456 = \underline{\hspace{1cm}}$ ).

### **Procedure**

All parts of the experiment were done inside the lab and all task-related materials were administered via laptop. Participants were recruited via the Psychology Department's Research Participation Scheduling System (SONA) website. After signing up for the experiment on SONA, participants were presented with the informed consent by the experimenter at the start of their appointment. All participants were given instructions stating that they would be studying lists that each contain 20 words that range in value from 1 to 10 points. The participants were instructed to try and remember as many words as they could in 60 seconds from each list while trying to achieve the highest possible score if they were to be tested on the words.

After the instructions, the participants were presented with each item value individually (Middlebrooks & Castel, 2018). The participants were able to click on a box that covers the word to study the word associated with that value or press the space bar to move on to the next value. Participants were given the option to select an item for restudy while it was on the screen. If the participants did not select the item for restudy, they would not see the value item pair again in

## TESTING AS A STUDY TOOL

that study period. If the participants selected the item for restudy, it was presented again after they cycled through the complete list one time. Participants were allowed to restudy items as many times as they liked during the 60-second study period.

To illustrate, during the initial presentation of the list, a participant may have selected to restudy *break*, *hound*, and *cookie*. The selected items would be presented for a second time after the participant was given the chance to study all 20 items. During the first restudy phase, the participant could elect to choose any of those words to restudy again, or none of them. If the participant selected *hound* for a second restudy, the word was presented after the initial restudy of the initially selected three words. Words could only be selected for restudy while the item was on the screen (Middlebrooks & Castel, 2018).

In the test with feedback condition, after studying the word-value pairs, participants were asked to recall as many of the items as they could. After being given the opportunity to recall the material, participants were shown their score (the sum of the value of the items they recalled) out of a possible 110 points. After being given their score, participants repeated the same procedure for lists two through six.

In the test with no feedback condition, after studying the word-value pairs, participants were asked to recall as many of the items as they could. After being given the opportunity to recall the material, participants then repeated the same procedure for lists two through six.

In the no test condition, participants were given a distractor task of math problems to answer. Participants answered as many of the problems as they could within one minute. After answering each math problem, participants repeated the same procedure for lists two through five. On the sixth list, participants were asked to recall as many words as they could from the sixth list.

## TESTING AS A STUDY TOOL

In the halfway test condition, after studying the word-value pairs, participants were asked to answer math problems for one minute. Participants repeated the same procedure for lists two through three. On list three, after studying the word-value pairs, participants were asked to recall as many of the items as they could. The participant was then asked to, after studying the word-value pairs, answer math problems for one minute. On the sixth list, participants were asked to recall as many words as they could from the sixth list.

After being presented with the final test, the participants were given a demographic questionnaire consisting of questions about their age, education level, English language acquisition, and sex. Additionally, questions regarding their use of outside resources were asked to ensure that participants were not writing down the words on the lists or using any outside resource to remember the words on the lists.

### Chapter 3 – Results

#### Analytic Approach

To preview, I first examined participants' study behavior to see if it was affected by prior testing and/or feedback by comparing study time as a function of item value across conditions. I did this by looking for differences in the time they spent studying high versus low valued items when they had taken a test (collapsing across test every list with feedback, test every list without feedback, and the halfway test conditions) versus when they had not taken a test (the no test condition) and when they had received feedback (test every list with feedback) or had not received feedback (collapsing across the test every list with no feedback, the halfway test, and the no test conditions). To further test the effects of testing, the test every list with feedback condition as well as the test every list without feedback condition and the halfway test condition were compared to the no test condition. In addition, test every list with feedback and test every list without feedback were compared to determine if there were differences in the feedback versus no feedback conditions.

Differences in time spent studying were examined in two ways: calculating a correlation for each participant between item value and time spent studying and calculating a difference score for each participant. The correlation, which was calculated using time spent studying extreme high values (8, 9, and 10) to time spent studying the extreme low values (1, 2, and 3), was used because it can capture the time spent studying on an individual item based on its value, with a strong positive correlation meaning that more time was spent studying higher value items. This analyses and most following analyses are conditional, meaning that they were only calculated based off of the items the participants had the opportunity to study. A difference score was also calculated for each participant. The difference score was used because it is an



## TESTING AS A STUDY TOOL

illustration of how long participants spent studying items as a whole, with a large difference score indicating more time spent studying high value items. The difference score was calculated using extreme z-scores. The study time was z-scored and then the lowest three value scores (1, 2, and 3) were averaged together and subtracted from the average of the three highest value scores (8, 9, and 10). This score was used to capture the most effective study strategy of studying the highest value items more than the lowest value items.

Next, I examined if participants' study behavior was more effective when they had taken prior tests with or without feedback. This was done by examining a proportion of the value of the recalled words to the possible value score of all words the participant studied as a function of prior testing and/or feedback. I did this by looking at the total value score for the participant on the final test and comparing that score to the possible score that could have been obtained if the participant recalled each word that he/she had the opportunity to study.

Third, I looked at the study efficacy of the participants. The selectivity index is a calculation that compares a participant's value score to an idealized value score, which is the maximum score the participant could have gotten given the number of items he/she recalled (Castel et al., 2002). While this is not a conditional proportion, all participants were told that the number of words they had to study of each value and therefore could choose to be selective. A selectivity index of positive one is indicative of perfectly selective study habits, meaning a person studied and remembered the highest value items. Selectivity index was compared across all conditions. Furthermore, the average selectivity index for those who took tests was compared to those who did not take tests, and the selectivity index of those who received feedback was compared to those who did not receive feedback.

## TESTING AS A STUDY TOOL

Lastly, I investigated if the basic interpolated test effect was replicated. This was done by examining the proportion of the words recalled in proportion to the words they studied. This was done in a different way than prior research because this experiment limited the amount of time the participants had to study the material and therefore changes the way this could be observed. This was compared across testing conditions as well as for a comparison to see if feedback was advantageous to participants.

Participants were given the option to indicate if there was an issue while taking the experiment. Most commonly, the issue was a skipped screen which may have affected the spacing of the experiment. Due to this the analyses were run both with and without these participants. The results presented are the analyses with those participants left in as excluding them did not change the conclusions drawn from these results. See Table 1 and Table 2 for the reported means.

### Study Behavior

**Correlation between item value and study time.** In order to analyze participants' study behavior, first a correlation was calculated for each participant between the time he/she spent studying each item and that item's value. This correlation was then used to look for differences across the test every list with no feedback ( $M = 0.22$ ,  $SD = 0.45$ ), the halfway test ( $M = 0.16$ ,  $SD = 0.65$ ), and the no test conditions ( $M = 0.23$ ,  $SD = 0.22$ ) using a one-way between-subjects ANOVA. Overall, participants had a positive average correlation  $M = 0.20$ ,  $SD = 0.42$ . There was not a significant difference across conditions,  $F < 1$ ,  $\eta_p^2 = 0.05$ , which indicates that participant did not differ on the amount of time dedicated to studying an item based upon the items value regardless of condition (see Figure 1).

## TESTING AS A STUDY TOOL

Table 1

*Means and standard deviations for all dependent variables across conditions*

	Time by Value	Extreme	Value	Selectivity	Proportion of
	Correlation	Z-Score	Proportion	Index	Words Recalled
Test Every List	0.19	0.85	0.48	-0.05	0.48
with Feedback	(0.41)	(2.76)	(0.27)	(0.32)	(0.25)
Test Every List	0.22	2.01	0.64	0.18	0.61
with no	(0.45)	(5.02)	(0.26)	(0.23)	(0.25)
Feedback					
No Test	0.23	2.98	0.20	0.05	0.19
	(0.22)	(3.28)	(0.16)	(0.16)	(0.14)
Halfway Test	0.16	1.39	0.54	-0.03	0.52
	(0.65)	(3.55)	(0.36)	(0.56)	(0.30)

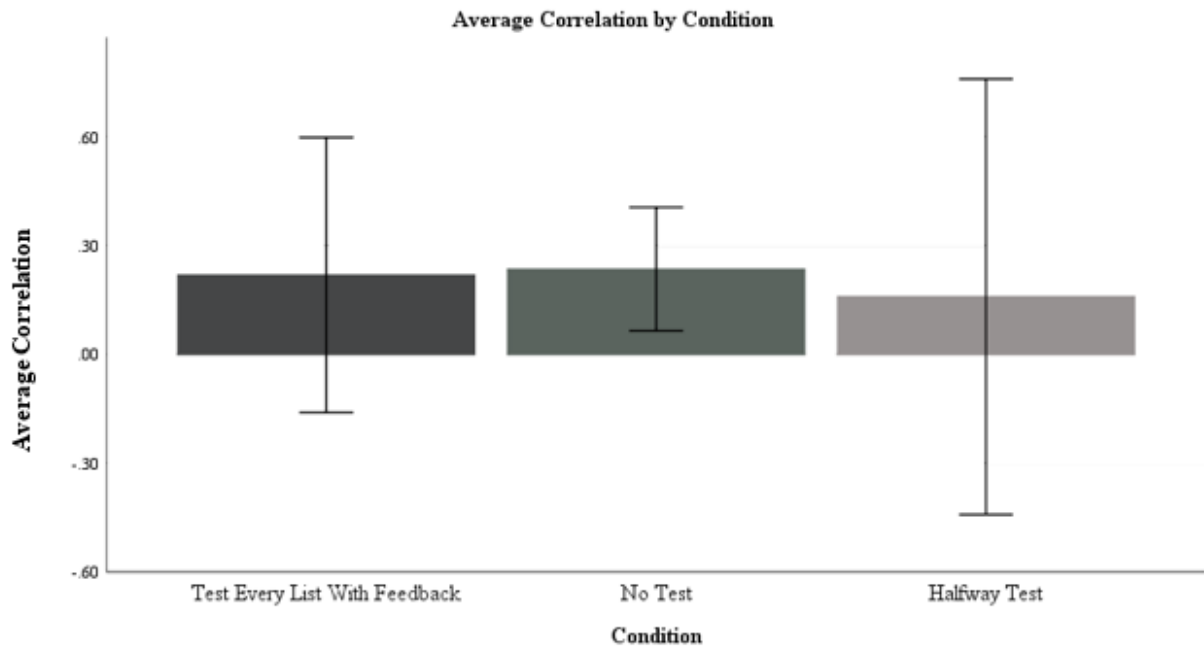
## TESTING AS A STUDY TOOL

Table 2

*Means and standard deviations for all dependent variables across conditions excluding participants who indicated an issue*

	Time by Value Correlation	Extreme Z-Score	Value Proportion	Selectivity Index	Proportion of Words Recalled
Test Every List	0.18	0.80	0.42	-0.06	0.44
with Feedback	(0.44)	(2.98)	(0.24)	(0.34)	(0.24)
Test Every List	0.14	1.12	0.63	0.16	0.60
with no Feedback	(0.43)	(4.70)	(0.28)	(0.23)	(0.27)
No Test	0.14	1.52	0.22	0.12	0.21
	(0.20)	(3.06)	(0.12)	(0.41)	(0.11)
Halfway Test	0.13	1.15	0.57	-0.04	0.55
	(0.71)	(3.82)	(0.38)	(0.61)	(0.31)

## TESTING AS A STUDY TOOL



*Figure 1.* Average conditional correlation compared across test every list with feedback, no test, and halfway test conditions.  
Error bars 95% CI.

## TESTING AS A STUDY TOOL

To compare participants' study behavior for the test and no test conditions, an independent samples t-test was used to determine if there were differences between the test and no test conditions. This test answers the question of whether or not interpolated tests had an effect on value of the word's participants chose to study. There was not a significant difference between the test every list without feedback condition ( $M = 0.22$ ,  $SD = 0.45$ ) and the no test condition ( $M = 0.23$ ,  $SD = 0.22$ ),  $t < 1$ ,  $d = 0.03$ . This means that taking interpolated tests does not affect how participants choose words to study when basing that choice on item value.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of testing, a 2 (list three, list six) x 2 (test every list/no feedback, halfway test) repeated measures ANOVA was used. There was not a significant interaction between list and taking interpolated tests when comparing the test every list without feedback condition ( $M = 0.40$ ,  $SD = 0.35$  for list three,  $M = 0.22$ ,  $SD = 0.45$  for list six) to the halfway test condition ( $M = 0.61$ ,  $SD = 0.35$  for list three,  $M = 0.16$ ,  $SD = 0.65$  for list six),  $F < 1$ ,  $\eta_p^2 = 0.06$ . There was not a significant main effect for interpolated tests,  $F < 1$ ,  $\eta_p^2 = 0.11$ . This means that taking interpolated tests does not change the study behavior of participants. There was a significant main effect for list,  $F(1, 13) = 4.60$ ,  $p = 0.05$ ,  $\eta_p^2 = 0.26$ . This means that the more lists that participants studied the more time that he/she studied higher value items.

To answer the question as to whether or not one interpolated test is enough to change study behavior, the halfway test condition was compared to the no test condition using an independent samples t-test. There was no significant difference between the halfway test ( $M = 0.16$ ,  $SD = 0.65$ ) and the no test ( $M = 0.23$ ,  $SD = 0.22$ ) conditions,  $t < 1$ ,  $d = 0.14$ . This means that taking one test is not enough to change participants study behavior by changing the way the study words in accordance with their item value.

## TESTING AS A STUDY TOOL

In order to answer the question as to whether or not taking multiple interpolated tests is better than taking only one test, the test every list without feedback condition ( $M = 0.22$ ,  $SD = 0.45$ ) was compared to the halfway test condition ( $M = 0.16$ ,  $SD = 0.65$ ) using an independent samples t-test. There was no significant difference between the two conditions,  $t(13) = -1.13$ ,  $p = 0.28$ ,  $d = 0.11$ . This means that the act of taking one test has no less effect than taking multiple tests when it comes to participants study choices in relation to the items value.

In order to answer the question as to if feedback was a helpful study tool the test every list with feedback condition was compared to the test every list without feedback condition using an independent samples t-test. The test with feedback condition ( $M = 0.19$ ,  $SD = 0.41$ ) was compared to the test without feedback condition ( $M = 0.22$ ,  $SD = 0.45$ ). There was not a significant difference between the test and no test conditions,  $t < 1$ ,  $d = 0.07$ . This means that feedback was not a helpful study tool.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of feedback a 2 (list three, list six) x 2 (test every list/feedback, test every list/no feedback) repeated measures ANOVA was used. There was not a significant interaction between list and feedback when comparing the test every list without feedback condition ( $M = 0.60$ ,  $SD = 0.23$  for list three,  $M = 0.61$ ,  $SD = 0.25$  for list six) to the test every list with feedback condition ( $M = 0.44$ ,  $SD = 0.28$  for list three,  $M = 0.48$ ,  $SD = 0.25$  for list six),  $F < 1$ ,  $\eta_p^2 = 0.01$ . There was not a significant main effect for feedback,  $F = 1.71$ ,  $\eta_p^2 = 0.11$ . There was not a significant main effect for list,  $F(1, 14) = 1.71$ ,  $\eta_p^2 = 0.12$ . This means that the presence of feedback does not change participants study behavior.

**Difference score using extreme z-scores.** For this difference score only z-scores calculated using extreme values (values 8, 9, and 10 as well as values 1, 2, and 3) were used.

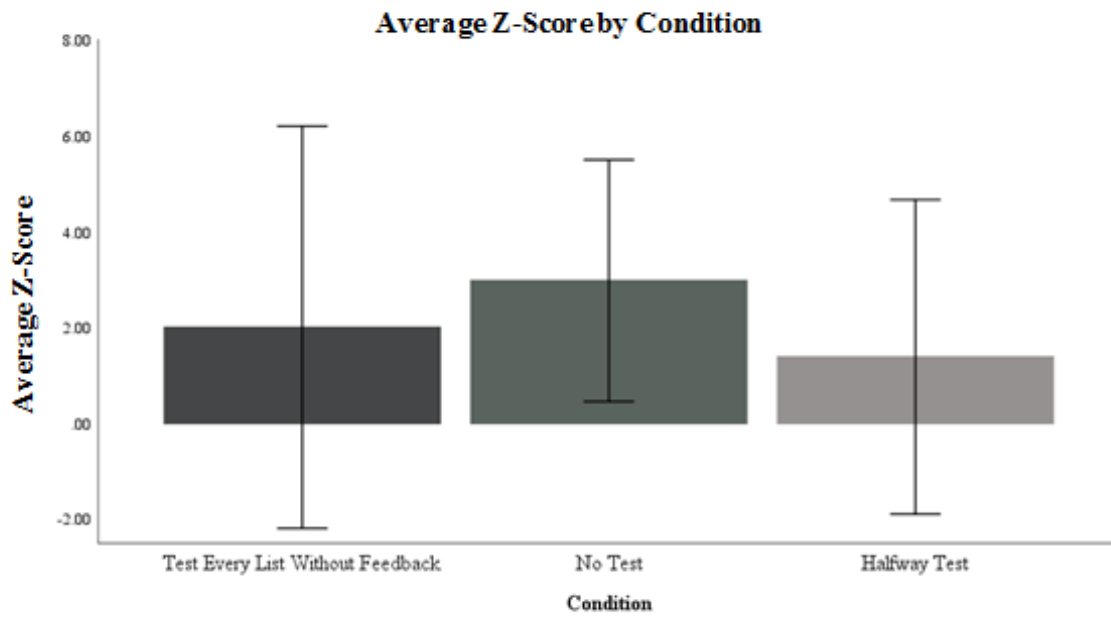
## TESTING AS A STUDY TOOL

These extreme values were used because of the small number of participants as a way to more accurately capture study time on high and low values. This score was then used to look for differences across the test every list without feedback ( $M = 2.01$ ,  $SD = 5.02$ ), halfway test ( $M = 1.39$ ,  $SD = 3.55$ ), and no test ( $M = 2.98$ ,  $SD = 3.28$ ) conditions using a one way between-subjects ANOVA. Overall, participants had a z-score of  $M = 1.86$ ,  $SD = 3.65$ . There was not a significant difference across conditions,  $F < 1$ ,  $n_p^2 = 0.03$ , which means that difference scores were not different between conditions, indicating that no condition spent more time studying higher value items than another (see Figure 2).

To compare participants' study behavior for the test and no test conditions, an independent samples t-test was used to determine if there were differences between the test and no test conditions. This test answers the question of whether or not interpolated tests had an effect on participants study behavior. There was not a significant difference between the test every list without feedback condition ( $M = 2.01$ ,  $SD = 5.02$ ) and the no test ( $M = 2.98$ ,  $SD = 3.28$ ),  $t < 1$ ,  $d = 0.23$ . This means that taking interpolated tests does not affect how participants choose words to study when basing that choice on item value.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of testing, a 2 (list three, list six) x 2 (test every list/no feedback, halfway test) repeated measures ANOVA was used. There was not a significant interaction between list and taking interpolated tests when comparing the test every list without feedback condition ( $M = 2.48$ ,  $SD = 1.77$  for list three,  $M = 2.01$ ,  $SD = 5.02$  for list six) to the halfway test condition ( $M = 2.12$ ,  $SD = 1.40$  for list three,  $M = 1.39$ ,  $SD = 3.55$  for list six),  $F < 1$ ,  $n_p^2 < 0.01$ . There was not a significant main effect for interpolated tests,  $F < 1$ ,  $n_p^2 < 0.01$ . There was also not a significant main effect for list,  $F < 1$ ,  $n_p^2 = 0.04$ . This means that taking





*Figure 2.* Average extreme z-score compared across test every list with feedback, no test, and halfway test conditions. Error bars 95% CI.

## TESTING AS A STUDY TOOL

interpolated tests does not change the study behavior of participants.

To answer the question as to whether or not one interpolated test is enough to increase the time participants spent studying higher value words, the halfway test condition was compared to the no test condition using an independent samples t-test. There was not a significant difference between the halfway test ( $M = 1.39$ ,  $SD = 3.55$ ) and the no test ( $M = 2.98$ ,  $SD = 3.28$ ) conditions,  $t < 1$ ,  $d = 0.47$ . This means that taking only one test is not enough to increase the time spent by participants on items of higher value.

In order to answer the question as to whether or not taking multiple interpolated tests is better than taking only one test, the test every list without feedback condition ( $M = 2.01$ ,  $SD = 5.02$ ) was compared to the halfway test condition ( $M = 1.39$ ,  $SD = 3.55$ ) using an independent samples t-test. There was no significant difference between the two conditions,  $t < 1$ ,  $d = 0.14$ . This means that taking only one test has no less effect than taking multiple tests.

In order to answer the question as to if feedback was a helpful study tool the test every list with feedback condition was compared to the test every list without feedback condition using an independent samples t-test. The test with feedback condition ( $M = 0.85$ ,  $SD = 2.76$ ) was compared to the test without feedback condition ( $M = 2.01$ ,  $SD = 5.02$ ). There was not a significant difference between the test and no test conditions,  $t < 1$ ,  $d = 0.29$ . This indicates that feedback did not help.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of feedback a 2 (list three, list six) x 2 (test every list/feedback, test every list/no feedback) repeated measures ANOVA was used. There was not a significant interaction between list and feedback when comparing the test every list without feedback condition ( $M = 2.48$ ,  $SD = 1.77$  for list three,  $M = 2.01$ ,  $SD = 5.02$  for list six) to the test

## TESTING AS A STUDY TOOL

every list with feedback condition ( $M = 2.10$ ,  $SD = 2.86$  for list three,  $M = 0.85$ ,  $SD = 2.76$  for list six),  $F < 1$ ,  $n_p^2 = 0.01$ . There was not a significant main effect for feedback,  $F < 1$ ,  $n_p^2 = 0.02$ .

This means that taking interpolated tests does not change the study behavior of participants.

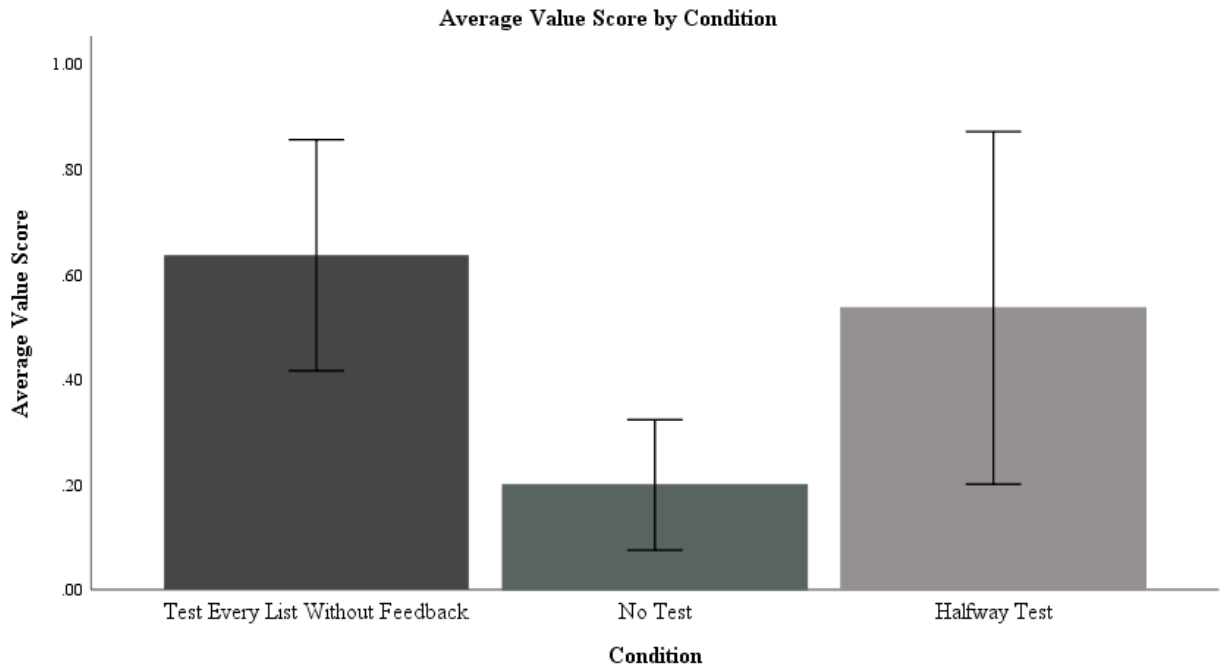
There was not a significant main effect for list,  $F < 1$ ,  $n_p^2 = 0.06$ . This means that the presence of feedback did not change the amount of time participants spent studying high value items.

### Study Efficacy

**Proportion of total value score out of total possible value score.** In order to examine if participants' study behavior resulted in better test performance, a proportion of total value score out of total possible value score was calculated for each participant. The value score is calculated using the sum of the value of the words correctly recalled by the participant, and the possible value score was calculated by summing all of the values attached to the words the participant was able to study, meaning that items that he/she did not get to study due to running out of time or electing not to study them were excluded. Overall, participants had an average proportional value of  $M = 0.45$ ,  $SD = 0.30$ . To examine the effect of number of interpolated tests, this value was compared across conditions using a one way between-subjects ANOVA comparing the test with no feedback condition ( $M = 0.64$ ,  $SD = 0.26$ ), the halfway test condition ( $M = 0.54$ ,  $SD = 0.36$ ), and the no test condition ( $M = 0.20$ ,  $SD = 0.16$ ). There was a significant difference between the conditions,  $F(2,21) = 6.34$ ,  $p < 0.01$ ,  $n_p^2 = 0.38$ . This means that number of interpolated tests had an effect on value score (see Figure 3). Both conditions who took interpolated tests scored higher than the no test condition.

Independent samples t-test was used to determine if there were differences between the test and no test conditions. This test answers the question of whether or not interpolated test had an effect on the proportion of words a participant recalled in comparison to the amount of words

## TESTING AS A STUDY TOOL



*Figure 3.* Average conditional value score compared across test every list with feedback, no test, and halfway test conditions. Error bars 95% CI.

## TESTING AS A STUDY TOOL

he/she studied. There was a significant difference between the test every list without feedback condition ( $M = 0.64$ ,  $SD = 0.26$ ) and the no test ( $M = 0.20$ ,  $SD = 0.16$ ),  $t(15) = 4.19$ ,  $p < 0.01$ ,  $d = 2.04$ . This means that taking interpolated tests allows participants to remember more words that he/she studied and therefore obtain a higher value score as compared to participants who do not take tests.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of testing, a 2 (list three, list six) x 2 (test every list/no feedback, halfway test) repeated measures ANOVA was used. There was not a significant interaction when comparing the test every list without feedback condition ( $M = 0.63$ ,  $SD = 0.26$  for list three,  $M = 0.64$ ,  $SD = 0.26$  for list six) to the halfway test condition ( $M = 0.27$ ,  $SD = 0.27$  for list three,  $M = 0.54$ ,  $SD = 0.36$  for list six),  $F(1, 13) = 1.47$ ,  $p = 0.25$ ,  $\eta_p^2 = 0.10$ . There was not a significant main effect for list,  $F(1, 13) = 1.67$ ,  $p = 0.22$ ,  $\eta_p^2 = 0.11$ . There was a trending significant main effect for test,  $F(1, 13) = 4.58$ ,  $p = 0.05$ ,  $\eta_p^2 = 0.26$ . This means that value score increased when participants took interpolated tests.

To answer the question as to whether or not one interpolated test is enough to increase the proportion of words recalled in comparison to the words the participant studied, the halfway test condition was compared to the no test condition using an independent samples t-test. There was a significant difference between the halfway test ( $M = 0.54$ ,  $SD = 0.36$ ) and the no test ( $M = 0.20$ ,  $SD = 0.16$ ) conditions,  $t(14) = 2.50$ ,  $p = 0.02$ ,  $d = 1.22$ . This means that taking one test is enough to aid participants in recalling words that he/she studied beyond those participants who never take a test, and therefore will lead to an increase in value score.

In order to answer the question as to whether or not taking multiple interpolated tests is better than taking only one test, the test every list without feedback condition ( $M = 0.64$ ,  $SD =$

## TESTING AS A STUDY TOOL

0.26) was compared to the halfway test condition ( $M = 0.54$ ,  $SD = 0.36$ ) using an independent samples t-test. There was no significant difference between the two conditions,  $t < 1$ ,  $d = 0.32$ . This means that the act of taking one test has no less effect than taking multiple tests when it comes to the ability to recall words that were studied.

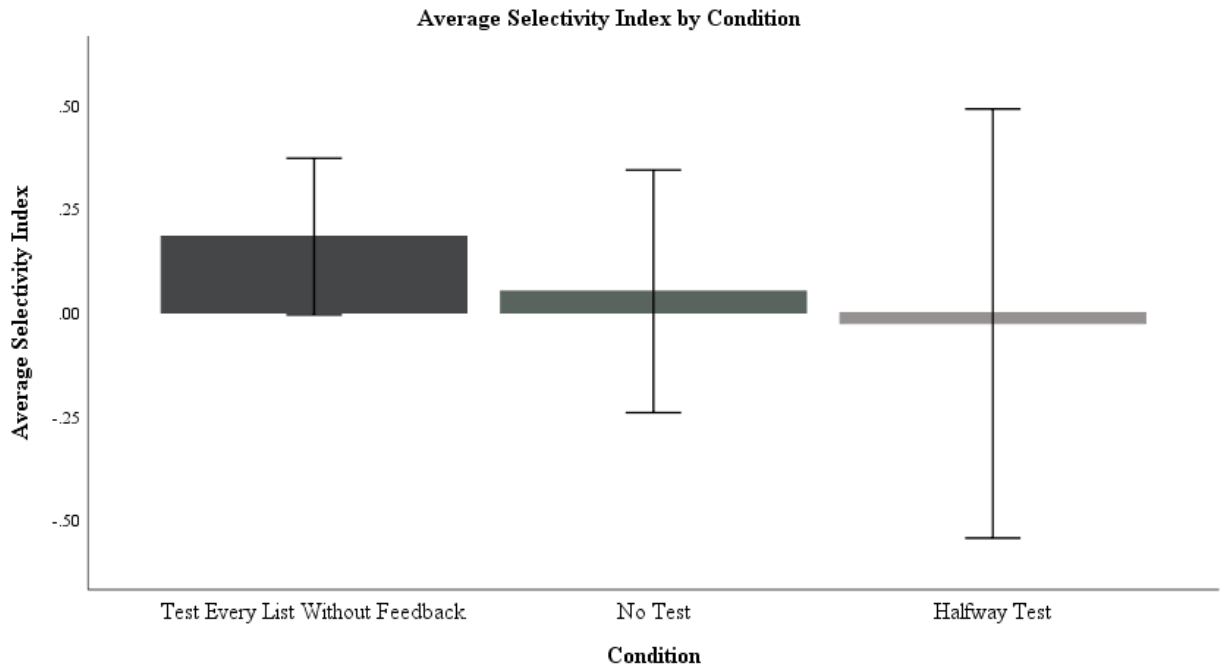
In order to answer the question as to if feedback was a helpful study tool the test every list with feedback condition was compared to the test every list without feedback condition using an independent samples t-test. The test with feedback condition ( $M = 0.48$ ,  $SD = 0.27$ ) was compared to the test without feedback condition ( $M = 0.64$ ,  $SD = 0.26$ ). There was a numeric trend that the test with feedback condition recalled fewer items than he/she studied when compared to the test without feedback condition. There was not a significant difference between the test and no test conditions; however, it was trending that participants in the test/no feedback condition had a larger proportion of words recalled,  $t(14) = -1.21$ ,  $p = 0.25$ ,  $d = 0.60$ . This indicates that feedback may not have helped.

To answer the question as to whether or not there was a change participants value score from list three to list six due to the effect of feedback a 2 (list three, list six) x 2 (test every list/feedback, test every list/no feedback) repeated measures ANOVA was used. There was not a significant interaction between list and feedback when comparing the test every list without feedback condition ( $M = 0.60$ ,  $SD = 0.23$  for list three,  $M = 0.61$ ,  $SD = 0.25$  for list six) to the test every list with feedback condition ( $M = 0.44$ ,  $SD = 0.28$  for list three,  $M = 0.48$ ,  $SD = 0.25$  for list six),  $F < 1$ ,  $\eta_p^2 = 0.01$ . There was not a significant main effect for feedback,  $F < 1$ ,  $\eta_p^2 = 0.02$ . There was not a significant main effect for list,  $F < 1$ ,  $\eta_p^2 = 0.02$ . This means that the presence of feedback does not change participant's value score.

**Selectivity Index.** Although value score is indicative of strategic study, the selectivity index is a superior way to capture how selective participants were because it considers the amount of words studied and instead of using only the total value of all words. However, unlike the proportion score calculated previously, this value takes into consideration all words, regardless if participants studied them or not. While this may be limiting, participants were all made aware that they had 20 items to study and had the opportunity to skip low value items if they chose not to study them. To determine whether or not participants were selective in their study behaviors the selectivity index was calculated for each participant. This value was then used in one way between-subjects ANOVA to determine if there were any differences across the conditions. Overall, participants had an average selectivity index of  $M = 0.04$ ,  $SD = 0.38$ . There was not a significant difference across the test every list with no feedback ( $M = 0.18$ ,  $SD = 0.23$ ), no test ( $M = 0.05$ ,  $SD = 0.38$ ), and the halfway test ( $M = -0.03$ ,  $SD = 0.56$ ) conditions,  $F < 1$ ,  $\eta_p^2 = 0.05$ , which indicates that participants were not more selective in their studies in one condition as compared to others (see Figure 4).

To compare participants' study behavior for the test and no test conditions, an independent samples t-test was used to determine if there were differences between the test and no test conditions. This test answers the question of whether or not interpolated tests had an effect on how selective were in their studying. There was not a significant difference between the test every list without feedback condition ( $M = 0.18$ ,  $SD = 0.23$ ) and the no test ( $M = 0.05$ ,  $SD = 0.38$ ),  $t < 1$ ,  $d = 0.41$ . This means that taking interpolated tests does not affect how selective participants are when studying.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of testing, a 2 (list three, list six) x 2 (test every list/



*Figure 4.* Average selectivity index compared across test every list with feedback, no test, and halfway test conditions. Error bars 95% CI.



## TESTING AS A STUDY TOOL

/no feedback, halfway test) repeated measures ANOVA was used. There was not a significant interaction between list and taking interpolated tests when comparing the test every list without feedback condition ( $M = 0.16$ ,  $SD = 0.28$  for list three,  $M = 0.18$ ,  $SD = 0.23$  for list six) to the halfway test condition ( $M = -0.05$ ,  $SD = 0.53$  for list three,  $M = -0.03$ ,  $SD = 0.56$  for list six),  $F < 1$ ,  $n_p^2 < 0.01$ . There was not a significant main effect for interpolated tests,  $F = 1.99$ ,  $p = 0.18$ ,  $n_p^2 = 0.13$ . There was not a significant main effect for list,  $F < 1$ ,  $n_p^2 < 0.01$ . This means that participants did not change how selective they were in their studies if they took more tests.

To answer the question as to whether or not one interpolated test is enough to increase the proportion of words recalled in comparison to the words the participant studied, the halfway test condition was compared to the no test condition using an independent samples t-test. There was not a significant difference between the halfway test ( $M = -0.03$ ,  $SD = 0.56$ ) and the no test ( $M = 0.05$ ,  $SD = 0.38$ ) conditions,  $t < 1$ ,  $d = 0.17$ . This means that taking one test is not enough to aid participants in being more selective.

In order to answer the question as to whether or not taking multiple interpolated tests is better than taking only one test, the test every list without feedback condition ( $M = 0.18$ ,  $SD = 0.23$ ) was compared to the halfway test condition ( $M = -0.03$ ,  $SD = 0.56$ ) using an independent samples t-test. There was no significant difference between the two conditions,  $t < 1$ ,  $d = 0.49$ . This means that the act of taking one test has no less effect than taking multiple tests when it comes to participant's selectivity.

In order to answer the question as to if feedback was a helpful study tool the test every list with feedback condition was compared to the test every list without feedback condition using an independent samples t-test. The test with feedback condition ( $M = -0.05$ ,  $SD = 0.32$ ) was compared to the test without feedback condition ( $M = 0.18$ ,  $SD = 0.23$ ). There was not a

## TESTING AS A STUDY TOOL

significant difference between the two conditions,  $t(14) = -1.70$ ,  $p = 0.11$ ,  $d = 0.83$ . This indicates that feedback did not help.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of feedback a 2 (list three, list six) x 2 (test every list/feedback, test every list/no feedback) repeated measures ANOVA was used. There was not a significant interaction between list and feedback when comparing the test every list without feedback condition ( $M = 0.16$ ,  $SD = 0.28$  for list three,  $M = 0.18$ ,  $SD = 0.23$  for list six) to the test every list with feedback condition ( $M = 0.16$ ,  $SD = 0.28$  for list three,  $M = -0.05$ ,  $SD = 0.32$  for list six),  $F < 1$ ,  $\eta_p^2 = 0.27$ . There was not a significant main effect for feedback,  $F = 2.60$ ,  $p = 0.13$ ,  $\eta_p^2 = 0.16$ . There was also not a significant main effect for list,  $F < 1$ ,  $\eta_p^2 = 0.01$ . This means that feedback did not aid participants in their selectivity.

### Overall Test Performance

**Proportion of total proportion of words recalled to total words studied.** In order to examine if participants' study behavior resulted in better test performance, a proportion of total words recalled out of total possible recallable words was calculated for each participant. This proportion was created by summing all words accurately recalled and comparing it to the number of words the participant studied. Overall, participants had an average proportion value of  $M = 0.44$ ,  $SD = 0.28$ . This value was compared across conditions using a one way between-subjects ANOVA comparing the test with no feedback condition ( $M = 0.61$ ,  $SD = 0.25$ ), the halfway test condition ( $M = 0.52$ ,  $SD = 0.30$ ), and the no test condition ( $M = 0.19$ ,  $SD = 0.14$ ). There was a significant difference between the conditions,  $F(2,21) = 7.78$ ,  $p < 0.01$ ,  $\eta_p^2 = 0.43$ . This means that the number of interpolated tests students had affected how many words he/she could recall

## TESTING AS A STUDY TOOL

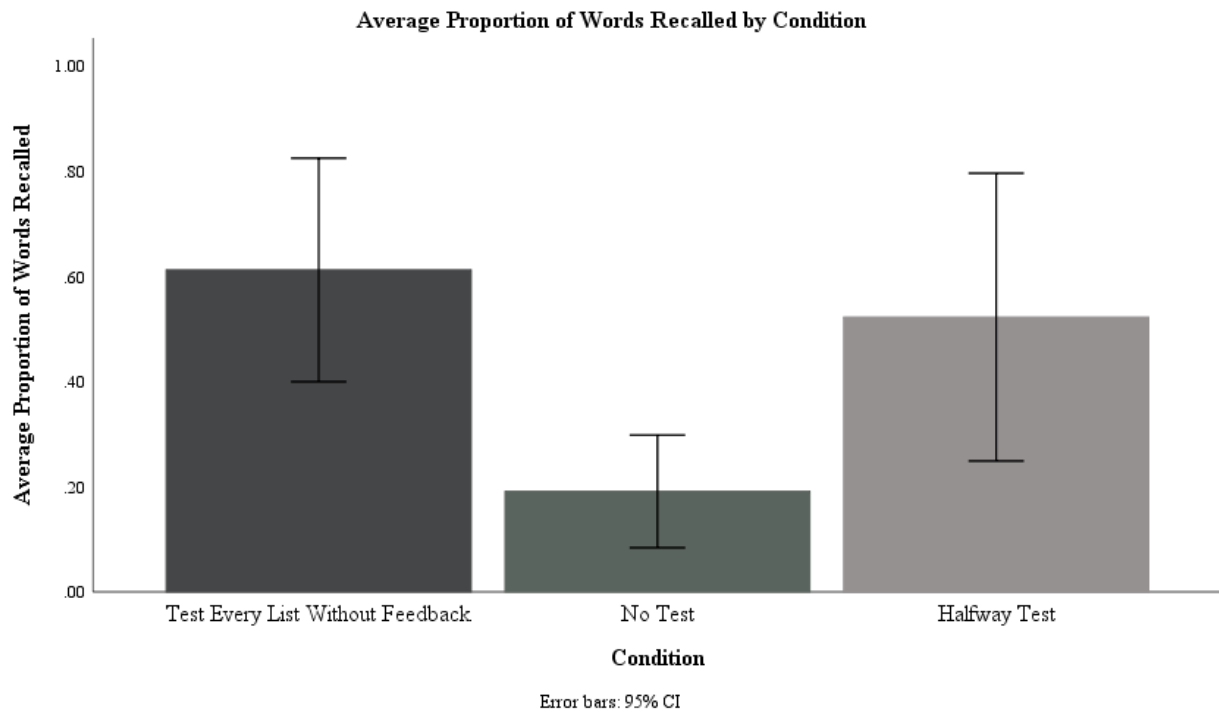
on the last list (see Figure 5). The test with no feedback condition could recall more words that they studied than any other condition.

In order to analyze participants' study behavior for the test and no test conditions, independent samples t-tests were used to determine if there were differences between the test and no test conditions. This comparison was made in three ways. The first answers the question of whether or not interpolated tests had an effect on the proportion of words recalled in comparison to the amount of words he/she studied. There was a significant difference between the test every list without feedback condition ( $M = 0.61$ ,  $SD = 0.25$ ) and the no test ( $M = 0.19$ ,  $SD = 0.14$ ),  $t(15) = 4.31$ ,  $p < 0.01$ ,  $d = 2.07$ . This means that taking interpolated tests allows participants to remember more words that he/she studied as compared to participants who do not take tests.

To answer the question as to whether or not one interpolated test is enough to increase the proportion of words recalled in comparison to the words the participant studied, the halfway test condition was compared to the no test condition using an independent samples t-test. There was a significant difference between the halfway test ( $M = 0.52$ ,  $SD = 0.30$ ) and the no test ( $M = 0.19$ ,  $SD = 0.14$ ) conditions,  $t(14) = 2.98$ ,  $p = 0.01$ ,  $d = 1.41$ . This means that taking one test is enough to aid participants in recalling words that he/she studied beyond those participants who never take a test.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of testing, a 2 (list three, list six) x 2 (test every list/no feedback, halfway test) repeated measures ANOVA was used. There was not a significant interaction when comparing the test every list without feedback condition ( $M = 0.60$ ,  $SD = 0.23$  for list three,  $M = 0.61$ ,  $SD = 0.25$  for list six),  $F(1, 13) = 1.43$ ,  $p = 0.25$ ,  $\eta_p^2 = 0.10$ . There was not a significant main effect for list,  $F(1, 13) = 1.78$ ,  $p = 0.21$ ,  $\eta_p^2 = 0.12$ . There was a significant

## TESTING AS A STUDY TOOL



*Figure 5.* Average proportion of words recalled compared across test every list with feedback, no test, and halfway test conditions. Error bars 95% CI.

## TESTING AS A STUDY TOOL

main effect for test,  $F(1, 13) = 4.64, p < 0.01, n_p^2 = 0.26$ . This means that taking interpolated tests led to participants recalling more words that were studied when compared to not taking interpolated tests.

In order to answer the question as to whether or not taking multiple interpolated tests is better than taking only one test, the test every list without feedback condition ( $M = 0.61, SD = 0.25$ ) was compared to the halfway test condition ( $M = 0.52, SD = 0.30$ ) using an independent samples t-test. There was no significant difference between the two conditions,  $t < 1, d = 0.33$ . This means that the act of taking one test has no less effect than taking multiple tests when it comes to the ability to recall words that were studied.

In order to answer the question as to if feedback was a helpful study tool the test every list with feedback condition was compared to the test every list without feedback condition using an independent samples t-test. The test with feedback condition ( $M = 0.48, SD = 0.25$ ) was compared to the test without feedback condition ( $M = 0.61, SD = 0.25$ ). There was a numeric trend that the participants in the test with feedback condition recalled less items that they studied when compared to the participants in the test without feedback condition. There was not a significant difference between the test and no test conditions,  $t(14) = -1.04, p = 0.32, d = 0.52$ . This indicates that feedback did not help.

To answer the question as to whether or not there was a change in study behavior from list three to list six due to the effect of feedback a 2 (list three, list six) x 2 (test/ with feedback, test/ no feedback) repeated measures ANOVA was used. There was a not a significant interaction between list and feedback from list three to list six when comparing the test every list with feedback condition ( $M = 0.44, SD = 0.28$  for list three,  $M = 0.48, SD = 0.25$  for list six) to the test every list without feedback condition ( $M = 0.60, SD = 0.23$  for list three,  $M = 0.61, SD =$

## TESTING AS A STUDY TOOL

0.25 for list six),  $F < 1$ ,  $n_p^2 = 0.01$ . There was no main effect for list,  $F < 1$ ,  $n_p^2 = 0.02$ . There was also no significant main effect for feedback,  $F = 1.71$ ,  $p = 0.21$ ,  $n_p^2 = 0.12$ . This means that feedback did not aid participants.

### Chapter 4 – Discussion

It was predicted that participants who take interpolated tests will study items of higher value (items with a value of eight or above) longer, indicating more strategic study, when compared to participants who do not take tests. The reason this was predicted is that retrieval practice via interpolated testing is an incredibly effective form of studying, such that learners attend better to the material and practice better study habits (Szpunar, Khan, & Schacter, 2013). It was also predicted that participants who receive feedback on tests will be more strategic in their study material selection than participants who do not receive feedback. This was a predication based on the exploration of feedback as it has not been studied thoroughly in the interpolated testing literature.

I also predicted that participants who take interpolated tests will have higher total value scores on the final test, which will be indicative of a more effective study strategy. This is because retrieval practice is an incredibly effective form of studying, such that learners attend better to the material and practice better study habits (Szpunar, et al., 2013). If participants take more tests, they may study higher value items which should lead to higher scores, indicating more strategic study choices. It was also predicted that participants who receive feedback on interpolated tests will produce higher value scores on the final test than participants who do not receive feedback after interpolated tests. This hypothesis was an exploration of if value-directed learning only happens in the presence of feedback, as well as if interpolated testing can in part be attributed to study effectiveness if the tests don't include feedback. Lastly, as a replication of the interpolated testing effect (Szpunar, et al., 2013), it was predicted that participants who take tests will recall more total items, regardless of item value, on the final test when compared to

## TESTING AS A STUDY TOOL

participants who do not take tests. Overall, evidence was found to support the hypotheses that testing encourages learners to be more strategic in their studies.

### **Study Behavior**

In order to analyze study behavior several different methods were used. In previous research it was demonstrated that participants would study words given a higher value for longer than he/she would study words given a smaller value (Castel, 2008). There was some evidence for the test every list with no feedback condition being superior to the test every list with feedback condition. This may be due to the type of feedback being given. While other studies have found this type of feedback to be effective, there may be some fundamental difference in the populations used in these studies as compared to the population used in the current research. In the present study when using a correlation between item value and study time, limited evidence was found for people being more likely to engage in this kind of behavior, as there was only a significant difference of list causing a change in study behavior.

In addition to the correlation, difference scores were also calculated using z-scores that were calculated using the extreme ends of the study possibilities (using only eight, nine, and ten value items to compare to items valued one, two, and three). When the extreme z-scores were compared across conditions there was no statistically significant difference between the conditions. Lastly, a proportional value score was used to determine if there was a difference in study behavior between conditions. There was a significant difference in people's study behavior. The hypothesis was based on previous research that suggested that when feedback is given with tests, learners are better able to direct their study habits (Castel, et al., 2002). While these results must be interpreted with caution due to the small number of participants, the results support the hypothesis that taking more tests will change study behavior. When the feedback and



## TESTING AS A STUDY TOOL

no feedback conditions were analyzed it was found that the no feedback condition had higher means than the feedback condition. This did not support the hypothesis that feedback would encourage learners to study higher value items longer than he/she studies lower value items. This lack of support for feedback may be due to the type of feedback as well as the population used in the current experiment. Given that this effect has been found previously, it is important to continue this research in order to make definitive conclusions about the role of feedback.

### **Study Efficacy**

It has been previously shown that learners who take more tests have higher test scores (Karpik & Blunt, 2008). It was found that there was a significant difference across the conditions, with the test without feedback condition having a higher average value score than all other conditions. In addition, it was found that participants who take tests have a higher average test score than those who do not take tests when comparing all of those who took tests as well as those who took only one test. Both of these findings support the hypothesis that those who take tests will have higher test scores than those who do not take tests, in addition to supporting the idea that people who take tests study more efficiently. This also provides evidence for the strength of testing as this effect can be found even when participants only take one test. However, the test with no feedback condition being better than the test with feedback condition was not consistent with the hypothesis. This could be due to fundamental differences within the populations.

It was predicted, based on prior research done by McDaniel and Fisher (1991), that participants who take tests with feedback will be more effective students than those who do not take tests. This was analyzed using a selectivity index ranging from negative one to one, with

## TESTING AS A STUDY TOOL

one being the most effective study. There were no significant differences found between the conditions on their selectivity index.

### **Total Item Recall**

The proportion of items recalled on a final test compared to the number of words studied was used to determine if interpolated testing had an effect on the number of items a participant was able to recall. The interpolated testing effect was replicated in this study, in line with prior research (Szpunar, et al., 2013). This effect being found in spite of having limited time supports the idea that interpolated testing is a valuable study tool. While this effect was replicated, it was not found that feedback had any effect on participant's ability to recall items. It is possible that feedback wasn't effective due to a potential population difference between the population who took the present study as compared to the population used for prior research.

### **Overall Discussion**

The current study sought to investigate the role that test taking plays in student study decisions, specifically what makes people more strategic in their studies. A strength of the study is that a population was used that is very applicable to the study. College students are exposed to learning on a daily basis. The results of this study could have a direct application on their lives, as well as the lives of their professors. If professors are able to identify ways to aid students in being better learners, they are able to help their students retain more class material and help them succeed. The results of the study suggest that professor should be testing their students more. Adding interpolated questions within a lecture can help their students retain the information that would be important for future exams.

A small number of participants made interpretation of the results difficult. Beyond the small  $n$ , there are several other possible limitations. The first limitation being that the population

## TESTING AS A STUDY TOOL

may have been unmotivated to complete the experiment correctly. While it was difficult to get participants for this study that may be due to the fact that data collection was started towards the end of the semester. By collecting data at this time, it is likely that many of the participants were doing studies last minute and may not have cared as much as participants who could have been collected at the beginning of the semester. While some participants demonstrated selective study, others simply clicked through all of the words or only spent time on one word, which leads me to believe that they were not giving the experiment their best effort.

In addition, this study should be taken into the classroom. Testing in a controlled environment is an important step for educational research; however, there are outside factors in students' lives, such as the distraction of others within the classroom and the temptation of technology use, which may distract them from the test at hand, which may influence the results.

Prior research has shown that testing, with and without feedback, is an effective study tool (Kang, et al., 2007; Butler, et al., 2007). Additionally, research has also found that learners use agendas to guide their study decisions (Ariel, et al., 2009). The present study sought to investigate the relationship between testing and studying. Specifically, what role do tests play in making a person a more effective studier? By using measures of study time and selectivity index as well as value score and proportion of words recalled, this question was analyzed. The research supported some of the hypothesis, but further research, especially on the relationship between feedback and study efficiency, is needed to be able to draw definitive conclusions.

### References

- Ariel, R. & Castel, A. D. (2014). Eyes wide open: Enhanced pupil dilation when selectively studying important information. *Experimental Brain Research*, 232(1), 337-344.
- Ariel, R., Dunlosky, J., & Bailey, H., (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, 138(3), 432-447.
- Ariel, R. & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Psychology: Applied*, 24, 43-56.
- Butler, A.C., Karpicke, J.D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *The Journal of Experimental Psychology: Applied*, 13(4), 273-281.
- Castel, A. D. (2008). The adaptive and strategic use of memory by older adults: Evaluative processing and value-directed remembering. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation* (Vol. 48, pp. 225–270). San Diego, CA: Academic Press.
- Castel, A.D., Benjamin, A.S., Craik, F.I.M., & Watkins, M.J. (2002). The effects of aging on selectivity and control in short-term recall. *Memory and Cognition*, 30(7), 1078-1085.
- Castel, A. D., McGillivray, S. & Worden, K. M. (2013). Back to the future: Past and future era-based schematic support and associative memory for prices in younger and older adults. *Psychology and Aging*, 28, 996-1003.

## TESTING AS A STUDY TOOL

- Cohen, M. S., Rissman, J., Hovhannisyan, M., Castel, A.D., & Knowlton, B. J. (2017). Recall test experience potentiates strategy-driven effects of value on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(10), 1581-1601.
- Dunlosky, J. & Ariel, R. (2011). Self-regulated learning and the allocation of study time. *Psychology of Learning and Motivation*, 54, 103-140.
- Dunlosky, J & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction*, 22, 271-280.
- Dunlosky, J., Rawson, Marsh, Nathan, & Willingham (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4-58.
- Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, 138(4) 469-486.
- Karpicke, J. D. & Blunt J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 11(331), 772-775.
- Karpicke, J. D., Roediger, (2008). The critical importance of retrieval for learning. *Science*, 319 (5864). 966-968.
- McDaniel, M.A., & Fisher, R. P. (1991) Tests and test feedback as learning sources. *Contemporary Educational Psychology*, 16(2), 192-201.
- McDermott, K. B., Arnold, K. M., & Nelson, S. M. (2014). The Sage Handbook of Applied Memory, Sage Publications, 183-197.

## TESTING AS A STUDY TOOL

McDermott, K. B., Agarwal, P.K., D'Antonio, L., Roediger, H.L., & McDaniel, M.A. (2014).

Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology*, 20(1), 3-21.

Metcalfe, J. (2009). Metacognitive judgements and control of study. *Current Directions in Psychological Science*, (18)3, 159-163.

Metcalfe, J. & Finn, B. (2008). Evidence that judgements of learning are causally related to study choice. *Psychonomic Bulletin and Review*, (15)1, 174-179.

Middlebrooks, C. D. & Castel, A. D. (2018). Self-Regulated Learning of Important Information Under Sequential and Simultaneous Encoding Conditions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, (44) 5, 779-792.

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G.M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: a generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5), 1287-1306.

Nelson, T. O. & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect" *Psychological Science* (2) 4, 267-270.

Nelson, T.O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science* (3) 4, 207-213.

## TESTING AS A STUDY TOOL

- Nelson, T.O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, 26, 125-173.
- Roediger, H. L. & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20-27.
- Roediger, H. L. & Karpicke, J.D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249-255.
- Son, L. K. & Metcalfe, J., (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (26) 1, 204-223.
- Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16), 6313-6317.
- Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6), 1392-1399.
- Tullis, J.G., Finley, J. R., & Benjamin A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory and Cognition*, 41(3), 429-442.
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated leaning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Mahwah, NJ: Erlbaum.